Open Lab In'Pact / université de BORDEAUX

21 octobre 2021

Projet EQAE / Évaluation qualitative des acquis des étudiants

Livrable 1	**Comparaison des modèles d'évaluation des acquis – exemples significatifs de mise en œuvre d'évaluation qualitative**

« L'évaluation n'est pas une mesure, du simple fait que l'évaluateur n'est pas un instrument et que ce qui est évalué n'est pas un objet au sens immédiat du terme. » [Hadji, 1997].

1 Introduction

Les deux fonctions principales de l'évaluation sont de pouvoir faire aux étudiants un retour sur la qualité de leurs acquisitions (évaluation formative) et de les noter afin de valider la réussite à un enseignement ou à un programme d'étude (évaluation sommative). La notation permet également de classer des étudiants, ou encore de construire des statistiques de réussite. Dans tous les cas, l'évaluation devrait reposer sur une appréciation la plus rigoureuse possible de l'assimilation par chaque étudiant, à l'issue d'un enseignement ou programme de formation, des **acquis d'apprentissage visés** (AAV).

Pour effectuer cette évaluation, on peut s'appuyer sur deux modèles distincts : l'évaluation reposant sur une **mesure** de la performance, faisant l'hypothèse que l'on peut s'appuyer sur une norme permettant de quantifier la connaissance, et l'évaluation par **référence à des critères**, destinée à apprécier et objectiver les changements dans les performances résultant de l'apprentissage (en d'autres termes essayer d'apprécier ce qui a été appris et de quelle manière).

La première approche suppose de pouvoir ramener l'évaluation des performances des étudiants à un **processus de mesure**. Or mesurer, c'est une opération de description quantitative de la réalité, consistant à affecter, à la suite d'une observation, un nombre à un événement ou un objet selon une règle acceptable. Une mesure est objective en ce sens que, une fois définie l'unité, on doit toujours reproduire la même mesure du même phénomène. Et cela suppose que l'instrument de mesure soit fiable (modulo une incertitude de mesure connue).

Dans le cadre de l'évaluation des AAV, l'instrument de mesure est l'enseignant lui-même. Or cet instrument n'est pas fiable, comme en témoignent les biais étudiés par la docimologie depuis bientôt cent ans. Et même le recours à un barème ne garantit pas d'améliorer la précision de cette tentative de mesure (voir fiche « la docimologie pour étudier les biais en évaluation »). Pierre Merle en 1996 écrit : « l'utilisation d'un barème de notation au point près, voire au demi-point près, ne constitue pas une garantie de précision de la correction. » (Merle, 1996, 225). La plupart des recherches en docimologie font le même constat et des études ont même montré qu'en mathématiques par exemple, c'est sans barème imposé que les écarts de notation sont les moins importants. Il y a plusieurs raisons à cela :

— le barème est toujours l'objet d'interprétation ;
— les contraintes, en particulier de temps, entraînent les enseignants à prendre des libertés à l'égard du barème ;

— la finesse du barème n'augmente la précision de la notation que si ce qui est exactement attendu de l'étudiant est entièrement clarifié, ce qui est peu souvent le cas.

L'alternative que constitue une **évaluation qualitative des AAV** par référence à des objectifs caractérisés au moyen de critères permet de lever pour une large part ces difficultés. En revanche elle se heurte, du moins en apparence, à un nouvel obstacle : le besoin de pouvoir recourir à une agrégation de plusieurs évaluations pour produire une évaluation finale, notamment pour valider la réussite à un programme de formation. Ce besoin d'agrégation est une des raisons mises en avant pour maintenir une évaluation quantitative, que ce soit par notes sur 10, sur 20 ou par pourcentages (c.à.d. notes sur 100), l'agrégation s'effectuant alors simplement par une moyenne arithmétique, le cas échéant pondérée par des coefficients. Or pour que le recours à la moyenne soit légitime, il faudrait que la notation soit effectivement une mesure, et notamment que les écarts ou rapports entre les notes soit significatifs, ce qui n'est pas le cas (voir fiche précédemment citée).

L'enjeu de cette étude de faisabilité est donc d'identifier un modèle reposant sur une approche qualitative de l'évaluation, tout en permettant d'agréger des résultats de manière satisfaisante, et le cas échéant au sein d'un système de notation quantitatif. Cette condition est importante si l'on veut pouvoir effectuer une migration progressive d'un système quantitatif vers un système qualitatif au sein d'un même programme de formation, et donc permettre la coexistence temporaire des deux systèmes.

La section suivante dresse un rapide état des lieux des principes d'évaluation des AAV et des systèmes de notation dans l'enseignement supérieur. La section 3 précise les modalités de mise en œuvre de l'évaluation qualitative au moyen de rubriques et la section 4 en donne deux exemples tirés de la littérature. Enfin la section 5 revient rapidement sur la problématique d'agrégation des résultats au sein d'un programme de formation.

2 Systèmes d'évaluation et de notation

Les sections 2.1 et 2.2 reprennent pour une large part du matériel issu de [Biggs & Tang 2011]. (Teaching for Quality Learning at University. Fourth Edition. McGraw Hill 2011)

2.1 Évaluation quantitative ou qualitative

Une évaluation maitrisée s'appuie sur une définition des objectifs à atteindre par les étudiants, les acquis d'apprentissage visés (AAV), une explicitation de la manière dont ces acquisitions seront évaluées par des **activités d'évaluation** (AE), et une objectivation des critères qui seront utilisés pour apprécier le résultat de l'évaluation et de le noter que ce soit au moyen d'une lettre ou d'un nombre (voir également fiche Évaluation — De quoi parle-t-on ?).

Nous avons indiqué en préambule qu'il existe deux modèles bien distincts d'évaluation des AAV, le premier reposant sur une **mesure de la performance** (au sens quantitatif) et le second sur une **appréciation de la qualité des acquisitions** par référence à une norme à atteindre.

Évaluation reposant sur la mesure (*measurement model assessment*). Elle nécessite que les résultats de l'apprentissage de chaque étudiant soient quantifiés sous forme de scores. Ces scores sont généralement exprimés le long d'une seule dimension afin de permettre de comparer les individus entre eux à des fins de classement. Cela repose sur une évaluation de l'apprentissage en fonction d'une quantité de matière apprise correctement, et sur l'idée que les bons étudiants savent plus de choses que les mauvais (et donc que c'est à ça qu'on les reconnait).

L'évaluation quantitative est le plus souvent traduite en pourcentages dérivant par exemple du rapport entre le nombre de réponses correctes et le maximum possible multiplié par 100. On trouve aussi des systèmes avec des notes sur 20 (France, Belgique, Portugal) ou sur 10. Les autres systèmes utilisant des nombres ou des chiffres sont souvent plus proches d'une évaluation à base de lettres et semblent intermédiaires entre approche basée mesure et approche basée norme.

Dans le cas d'un système quantitatif au sens strict, disons reposant sur des pourcentages, on suppose que ces pourcentages sont une « monnaie universelle », équivalente dans toutes les matières et pour tous les élèves, de sorte que les performances de différents élèves dans différentes matières peuvent être additionnées,

moyennées et directement comparées. C'est une vision erronée et il n'y a aucun moyen de savoir si 75 % en physique est le « même niveau » que 75 % en droit, ni si ce sont les « mêmes notes », ni encore si une résultat de 75 % obtenue dans une matière donnée atteste d'un progrès par rapport à un 70 % obtenu par le même étudiant dans la même matière durant un semestre d'étude précédent.

Ce système d'évaluation va généralement de pair avec l'idée que la distribution des résultats à une évaluation suit une **courbe en cloche**. C'est d'ailleurs le principe justifiant implicitement pour faire correspondre les différents systèmes de notation européens dans le système ECTS. Or cette hypothèse est **sans fondement**, car pour que ce soit le cas, il faudrait d'une part que les aptitudes des étudiants soient distribuées selon une loi normale et donc que l'admission des étudiants à un cursus soit complétement aléatoire, et d'autre part que ces aptitudes soient le seul déterminant de la réussite, c'est-à-dire que la qualité de l'enseignement n'ait aucun impact sur le résultat. Or l'effet d'un bon enseignement devrait être de réduire l'écart initial entre les étudiants et la distribution des résultats devrait dans ce cas être asymétrique, avec des notes élevées plus fréquentes que des notes faibles.

Les approches quantitatives véhiculent une idée que l'évaluation est scientifique, précise et objective. Pourtant l'erreur de mesure est nécessairement supérieure à un point de pourcentage (ou une fraction de point sur 20). D'ailleurs lorsque les enseignants ou les jurys sont confrontés à un cas limite, il est courant de dire que puisque l'échelle n'est pas précise à un point près, on accorde le niveau supérieur au bénéfice du doute. Cette **impression de rigueur est donc également erronée**. Cette approche ne garantit nullement une appréciation objective et précise du niveau atteint. À l'arrivée, une série de « micro-jugements » subjectifs et indépendants (une note pour ceci, une note pour cela) s'accumulent et la décision finale (par exemple réussite ou échec, très bon niveau ou seulement acceptable, etc.) résulte de l'agrégation de nombres embarquant avec elle l'agrégation des erreurs dans tous ces micro-jugements.

Enfin, la manière dont sont utilisées ces échelles de notes les empêche d'être des échelles à intervalles égaux, c'est-à-dire où la différence entre deux nombres adjacents est la même que celle entre deux autres nombres adjacents. **Cela ne permet pas d'expliciter ni même de justifier un résultat obtenu** : *« je note sur dix et cette réponse est la meilleure ; elle devrait donc obtenir dix, mais je lui donne neuf car aucune réponse ne peut être parfaite »* ou encore *« cette réponse est moyenne, donc elle vaut cinq points »*. La décision finale devrait être prise, non pas sur l'accumulation de jugements mineurs imparfaits, mais sur un jugement raisonné et publiquement soutenable sur la performance elle-même. Cela nécessite un jugement holistique fondé sur des critères publiquement énoncés.

Dans ces conditions, pourquoi ce modèle continue-t-il d'être utilisé ? Biggs et Tang proposent plusieurs raisons distinctes :

1. par tradition, habitude : pourquoi remettre en question ce qui a bien fonctionné dans le passé, surtout lorsque les structures et procédures rendent le changement difficile ;
2. par commodité de mise en œuvre : illusion de la précision (laisser les chiffres prendre les grandes décisions), illusion de normalisation cohérente, etc. ;
3. par commodité pour l'enseignant : les questions d'examen peuvent être repoussées durant une bonne partie du semestre, il est facile de calculer des moyennes et combiner les notes entre les activités et les cours, les notes peuvent être utilisées à des fins disciplinaires (minoration pour soumission tardive), il est plus simple d'argumenter avec les étudiants sur la base de chiffres en cas de litige ;
4. en raison d'une croyance sincère dans le modèle de mesure (l'objectif de l'évaluation est de classer les étudiants et le rôle de l'enseignant est de faire le tri entre les bons et les mauvais).

Évaluation reposant sur une référence à des normes (*standard model assessment*). Le modèle d'évaluation par référence à des normes est conçu pour évaluer les changements dans les performances résultant de l'apprentissage, dans le but de voir ce qui a été appris et dans quelle mesure. Une telle évaluation est **critériée** en ce sens que les résultats de l'évaluation sont déterminés en fonction de la manière dont un étudiant satisfait aux critères d'apprentissage fixés. Il ne s'agit donc pas d'identifier les étudiants en fonction d'une caractéristique quelconque mais les performances indiquant ce qui a été appris et dans quelle mesure. Contrairement à l'évaluation construite sur une mesure, le résultat d'un individu est **indépendant** de celui d'un autre individu.

Il est d'ailleurs intéressant de noter qu'en dehors des établissements d'enseignement, ce modèle est utilisé pratiquement systématiquement lorsque quelqu'un enseigne quelque chose à quelqu'un d'autre. Par exemple les instructeurs de natation ont des normes qu'ils veulent que leurs apprenants atteignent. Et les parents, lorsqu'ils apprennent à leurs enfants à lacer leurs chaussures ne leur font pas passer un QCM à la fin pour savoir s'ils y parviennent mieux que l'enfant du voisin.

L'évaluation qualitative repose sur l'idée qu'il est possible de définir des normes (des critères) caractérisant les résultats d'apprentissage attendus d'un enseignement. Cela suppose de rédiger les AAV de manière appropriée (voir par exemple fiche taxonomie de Bloom). Le rôle de l'évaluation est de permettre d'indiquer dans quelle mesure ces AAV ont été atteints, le « dans quelle mesure » n'étant pas exprimé par des notes en pourcentages mais par une hiérarchie de niveaux identifiés par des notes, idéalement sous forme de lettres car celles-ci ne font pas courir le risque d'induire une notion abusive de mesure comme le font des nombres.

L'évaluation qualitative permet de **diversifier les conditions d'évaluation** alors qu'il est nécessaire de standardiser au maximum ces conditions lorsqu'on veut comparer et classer. Les individus apprennent et réalisent des performances optimales dans différentes conditions et avec différents formats d'évaluation. Certains travaillent mieux sous pression, d'autres ont besoin de plus de temps. Comme dans le travail professionnel lui-même, il y a souvent plusieurs façons d'obtenir un résultat satisfaisant, et la diversification des conditions d'évaluation devrait pouvoir permettre à chaque étudiant de montrer ses savoir-faire de différentes manières. En ce sens l'évaluation qualitative est plus adaptée à la recherche de la performance optimale de chaque individu.

Mais l'aspect essentiel de ce modèle est de permettre aux enseignants de **juger les performances en fonction de critères**, alors que ce point est généralement éludé lorsqu'on utilise un modèle basé sur la mesure. On passe en effet de la question « *Combien de points dois-je attribuer à cette section ?* » à la question « *Dans quelle mesure cette performance dans son ensemble répond-elle aux critères d'attribution d'un A ou d'un D ?* ». Afin de porter ces jugements, les enseignants doivent savoir ce qu'est une performance de mauvaise qualité, ce qu'est une performance de bonne qualité, et pourquoi.

Comparaison des deux modèles. On peut définir d'autres critères pour apprécier un système d'évaluation, et notamment son niveau de fiabilité en termes de stabilité d'un test à l'autre, de dépendance aux conditions de test, etc. Le tableau 1 [Biggs&Tang p219] récapitule différents éléments de comparaison des deux modèles.

Modèle basé sur	une mesure	un système de référence
Théorie	Quantitatif. Théorie de test classique, supposant que les notes suivent une loi normale (distribution gaussienne).	Qualitatif. Théorie de l'apprentissage permettant des appréciations cohérentes. Pas d'hypothèse sur la distribution des notes.
Stabilité	Les résultats restent stables tout au long des évaluations.	Les résultats sont meilleurs après l'enseignement qu'avant.
Dimensionnalité	Le test est unidimensionnel. Tous les items mesurent un même concept.	Le test est multidimensionnel (sauf s'il y a un seul AAV). Les items concernent l'ensemble des AAV.
Conditions de test	Les conditions doivent être standardisées pour tous les apprenants.	Les conditions doivent être optimales pour chaque apprenant.
Validité	Externe : de quelle manière les résultats du test sont corrélés avec des représentations extérieures.	Internes : de quelle manière les résultats sont reliés aux AAV et performances du domaine ciblé.
Utilisation	Sélection des étudiants. Comparaisons d'individus, de référence à des modèles de populations.	Appréciation de l'efficacité des apprentissages, pendant et à l'issue des activités d'enseignement et d'apprentissage.

TABLE 1 – Comparaison entre les modèles d'évaluation basés sur une mesure et ceux basés sur une référence à des normes. [Biggs & Tang 2011]

2.2 Évaluation et alignement pédagogique

Relier l'évaluation et l'**alignement pédagogique** (ou *constructive alignement* selon la terminologie proposée par son promoteur John Biggs) semble incontournable si l'on veut poser les bases d'une évaluation formative reposant sur une approche qualitative. On peut résumer le principe de la mise en œuvre de l'alignement pédagogique en quatre étapes :

1. décrire le résultat d'un apprentissage sous la forme d'AAV (verbe indiquant l'activité d'apprentissage + objet indiquant le contenu) et préciser une norme que les étudiants doivent atteindre ;
2. créer un environnement d'apprentissage à l'aide d'activités d'enseignement/apprentissage qui portent sur ce verbe et sont donc susceptibles de produire le résultat escompté ;
3. utiliser des tâches d'évaluation qui contiennent également ce verbe, et permettent ainsi de juger à l'aide de rubriques si, et dans quelle mesure, les performances des étudiants répondent aux critères ;
4. transformer ces jugements en critères de notation standard.

Pour reprendre l'exemple cité précédemment consistant à apprendre à lacer des chaussures, l'alignement pédagogique est dans ce cas parfait : le résultat d'apprentissage visé par les parents, l'activité d'enseignement/apprentissage et l'évaluation sont tous les mêmes : il s'agit de lacer une chaussure.

Dans le cadre de cette étude de faisabilité nous nous concentrerons sur les deux dernières étapes. On voit bien cependant que ces quatre étapes sont intrinsèquement liées et qu'il ne sera possible de mettre correctement en œuvre une évaluation qualitative et une notation littérale que si l'on adopte l'ensemble du principe d'alignement pédagogique. En particulier, il sera important de réfléchir en amont à la conception des activités qui seront évaluées de cette manière. Sans davantage développer ici cet aspect, on peut simplement énoncer trois principes que de telles activités d'évaluation (AE) devraient satisfaire :

1. une AE appropriée doit indiquer dans quelle mesure chaque étudiant a atteint le ou les AAV concernés et/ou dans quelle mesure l'activité elle-même a été exécutée ;
2. elle ne doit pas inciter les étudiants à adopter des stratégies de bas niveau telles que la mémorisation, le repérage des questions et autres formes d'évitement ;
3. les critères des notes attribuées pour décrire la façon dont les AE ont été réalisées doivent être clairement décrits sous forme de rubriques que les étudiants comprennent parfaitement.

2.3 Diversité des systèmes de notation

Quel que soit le mode d'évaluation mis en œuvre, le résultat doit ensuite être traduit dans un système de notation. La très grande diversité des systèmes de notation et l'hétérogénéité des informations disponibles rendent difficile la production d'un benchmark systématique et exhaustif. On peut s'en faire une idée en visitant le site `www.scholaro.com/pro/Countries` qui présente les systèmes de notation de 232 pays ou provinces différents (avec par exemple une dizaine de variantes simplement pour le Canada).

La **grande majorité des systèmes de notation dans le monde** expriment le résultat final à un cours ou à un programme complet de formation au moyen d'un petit ensemble de notes, généralement au nombre de quatre ou cinq, et le plus souvent exprimées aux moyens de lettres (A, B, C, F ou A, B, C, D, F), et avec des possibilités de modulations autour de ces lettres (par exemple B−, B, B+). Parfois ces niveaux sont exprimés par des chiffres qui soit sont équivalents à des lettres, soit expriment des intervalles. C'est cette information qui est qualifiée de note (*grade*). Elle est dans certains cas remplacée par une description littérale de la note (*grade description*) : *Very good*, *Sufficient*, *Not Sufficient*, etc. Chaque système de notes est en général associé à une échelle numérique. On peut également constater que dans un nombre important de pays, il existe des variantes du système majoritairement utilisé : utilisation de lettres alternatives, augmentation ou diminution du nombre de barreaux, utilisation ou non de modulations autour de la note, etc.

Une proportion importante de pays a adopté un système à cinq niveaux, souvent exprimé au moyen des lettres A, B, C, D et F, et associé à une échelle exprimée en pourcentages. C'est par exemple le cas de l'Australie

(avec les lettres N, P, C, D, HD et une variante à quatre lettres), du Brésil (avec une variante à six lettres), du Canada (avec l'exception du Québec utilisant principalement quatre lettres), de la Chine (avec une variante à quatre lettres), de l'Égypte, de la Finlande (avec les notes ET, KT, HT, TT et VT), d'Hong Kong, de l'Iran, de Madagascar (avec E à la place de F et une échelle de 0 à 20), du Pérou, de Singapour, de la Turquie et des USA. La plupart de ces pays ont recours à des modulations autour de la note exprimées au moyen des symboles + et − (ou plus rarement des combinaisons de lettres voisines comme la Turquie : BA, BB et BC pour moduler autour de la note B, etc.). Les autres pays utilisant cinq niveaux ont soit recours à des chiffres, comme l'Allemagne, la Bulgarie, la Croatie ou la Hongrie (avec une variante à sept niveaux), soit à une description littérale de ces niveaux comme l'Espagne, l'Inde, l'Irlande, Israël, le Panama ou l'Uruguay.

On retrouve les mêmes principes pour les pays utilisant quatre niveaux de notes, par exemple avec les lettres A, B, C et F comme l'Autriche, l'Argentine, le Japon (avec l'ajout de la lettre S, signifiant *Exemplaire* mais rarement attribuée) et le Mexique, avec un système de chiffres comme Cuba, la Grèce, la Pologne, le Salvador ou la Russie, ou avec une description littérale des niveaux comme l'Afrique du Sud, la Grande Bretagne, le Népal, ou le Venezuela.

On trouve plus rarement quelques pays utilisant six niveaux : Norvège, Nouvelle Zélande ou Tanzanie par exemple. On peut enfin noter quelques cas extrêmes avec la Suède utilisant seulement trois descriptions de niveaux (*Distinction*, *Pass*, *Failed*) avec une variante à quatre niveaux, la Bolivie utilisant sept niveaux et la Hollande utilisant dix niveaux de description.

On peut à ce stade déjà dresser plusieurs constats à partir de cette revue de différents systèmes de notation de l'enseignement supérieur international. Tout d'abord il existe **une forte convergence des modèles vers un système de notation basé sur quatre ou cinq niveaux**, même s'il existe quelques variantes pouvant aller de trois à six ou sept (voire dix dans un cas extrême). De même, la plupart de ces modèles permettent des ajustements en majorant ou minorant ces niveaux, par exemple au moyen des symboles + et −.

Ensuite il existe une **grande diversité dans la manière d'associer les notes et leurs ajustements à une échelle de points**, qu'il s'agisse de scores ou pourcentages. Cela résulte probablement pour une part de la méthode suivie pour attribuer ces points, pouvant aller d'une approche quantitative reposant sur des barèmes ou de type QCM par exemple, jusqu'à une approche purement qualitative par référence à une grille de critères. Dans le premier cas on peut supposer que les pourcentages jouent un rôle assez similaire à celui de nos notes sur vingt dans le processus d'appréciation du résultat d'une activité d'évaluation, et dans le second qu'ils sont plutôt utilisés pour quantifier et combiner des notes à la suite d'une évaluation qualitative.

Mais cette diversité est aussi une conséquence probable de ce que ces systèmes, qu'ils reposent sur une approche plutôt qualitative ou plutôt quantitative, **font tous la différence entre la manière d'apprécier le résultat d'une activité d'évaluation et sa traduction en une note rendant compte du niveau atteint par l'étudiant** relativement à un ou plusieurs AAV. On peut alors considérer que cette diversité d'échelles traduit également une diversité de propositions pour structurer l'étape d'appréciation du résultat avec tout ce que cela comporte de subjectivité, et d'une certaine manière la difficulté qu'il y a à relier de manière rigoureuse une échelle de points pouvant être combinés de manière arithmétique à une échelle de notes définissant un niveau d'acquisition des AAV.

Cette difficulté est naturellement éludée dans le système français, puisque la note sur vingt amalgame une tentative de « mesure » d'une performance avec la notation de cette performance, faisant ainsi l'impasse sur cet aspect pourtant essentiel de l'évaluation. Sauf naturellement lors de la prise en compte des mobilités internationales...

Pour terminer, bien que le système de notation à quatre ou cinq niveaux ne soit pas adapté à classer les étudiants en fonction de leurs résultats, il peut être converti en une échelle numérique, par exemple avec le système nord-américain GPA (*Grade Point Average*), ce qui permet alors d'effectuer ce classement en ayant recours à des moyennes, qui conservent dans ce cas une certaine légitimité (voir section 5).

3 Évaluation qualitative d'une activité

3.1 Notation qualitative

La notation d'une AE comporte deux aspects :
— l'évaluation des résultats en référence à plusieurs critères ;
— la combinaison de ces résultats pour obtenir une note globale.

Avec une approche qualitative, la note globale doit contribuer à rendre compte du degré d'acquisition du ou des AAV concernés. Dans certains cas (notamment projet tutoré, mémoire, portfolio), on peut arriver à se passer de la première première étape et **produire une note finale de manière holistique**. En voici un exemple concernant l'évaluation d'un port-folio [Biggs & Tang]. Le degré de réalisation est ici qualifié sur cinq niveaux :

A : capable de réfléchir, de s'auto-évaluer de manière réaliste, capable de formuler et d'appliquer la théorie à des situations de classe problématiques, maîtrise claire du contenu du cours ;

B : peut appliquer la théorie à la pratique, compréhension globale du cours et de ses composantes ;

C : peut expliquer les théories les plus importantes, peut décrire d'autres sujets de manière acceptable ;

D : ne peut expliquer que certaines théories ;

F : moins que D, plagiat.

La notation ne repose sur aucune utilisation quantitative de points ni de moyenne pour calculer la note finale. Chaque élément du portfolio est évalué pour savoir s'il fournit des « preuves » de niveau de qualité A, B, etc. Si par exemple l'ensemble des preuves ne reflète pas une auto-évaluation réaliste mais montre une capacité à élaborer une théorie fonctionnelle et à l'appliquer à des situations de classe, la note B est clairement attribuée.

L'appréciation du résultat peut être modulée (par exemple B+ et B− autour de B), mais **il est important d'attribuer en premier lieu la note** (la lettre) qui exprime un niveau de qualité, **puis seulement ensuite de la moduler**. Cette manière de faire diffère fondamentalement de celle consistant à attribuer une note sur une échelle supposée exprimer une mesure. Dit autrement, noter avec les treize notes A+,A, A−, B+, B, B−, C+, C, C−, D+, D, D−, F est intrinsèquement différent de noter sur une échelle numérique de 0 à 12, dont les intervalles sont supposés être régulièrement espacés.

On peut également noter séparément les différents critères et les combiner ensuite pour obtenir la note globale, soit par comptabilisation des lettres, soit en ayant recours à une correspondance numérique intermédiaire (voir exemples section 4).

Dans tous les cas, la mise en œuvre d'une évaluation qualitative repose sur la définition et l'utilisation d'**indicateurs suffisamment clairs et précis** pour être utilisés de manière opérationnelle. Dans le cas d'une évaluation holistique un indicateur sera associé à chaque note (à chaque niveau). Dans le cas d'une évaluation décomposée en plusieurs critères évalués séparément et dont les résultats sont ensuite recombinés, il est courant d'organiser ces critères sous la forme d'une **grille critériée** exprimant de quelle manière chaque activité contribue au résultat final selon une échelle d'appréciation commune. Il faut bien en effet dans ce cas que tous les critères soient appréciés selon une même échelle pour permettre ensuite d'agréger de manière cohérente les différents résultats partiels.

3.2 Utilisation de grilles critériées

Une grille critériée est un tableau qui détaille à la fois les critères utilisés pour interpréter la preuve d'apprentissage fournie par l'étudiant dans un travail à réaliser et les indicateurs ou niveaux de performance possibles pour chaque critère.

Les critères sont donc l'explicitation d'indices observables de la qualité des réponses ou des prestations des étudiants lors d'une évaluation. Ils sont directement liés aux objectifs pédagogiques, et permettent de clarifier ce qui est évalué en fournissant une description détaillée de ce qui doit être acquis, appris ou maîtrisé par les étudiants.

Les niveaux de performance ou indicateurs permettent de clarifier les différents niveaux de compétence des

étudiants et donc de faire la distinction entre les étudiants qui ont atteint les apprentissages attendus de façon « excellente », « suffisante », « insuffisante », etc. (voir exemple TABLE 2).

Il est suffisant de déterminer, pour chaque dimension, un niveau correspondant aux attentes minimales (en dessous duquel les prestations sont jugées insuffisantes), ainsi qu'un petit nombre de niveaux correspondant à des degrés significatifs de meilleure performance.

Critères		Inacceptable	Insuffisant	Correct	Excellent
Manifestation de la compréhension de la situation-problème	Respect des étapes	L'étudiant amorce certaines étapes sans les compléter	L'étudiant effectue quelques étapes	L'étudiant effectue les principales étapes	L'étudiant effectue toutes les étapes
	Prise en compte des données et des contraintes	Tient compte de certaines données sans distinguer celles qui sont pertinentes et tient compte de peu ou pas de contraintes à respecter	Tient compte de certaines données pertinentes et de certaines contraintes à respecter	Tient compte des données pertinentes et de la plupart des contraintes à respecter	Tient compte des données pertinentes et de toutes les contraintes à respecter
Mobilisation correcte des concepts et processus requis pour produire une solution appropriée	Utilisation des concepts et processus mathématiques	L'étudiant fait appel à des concepts et processus mathématiques inappropriés	L'étudiant fait appel à quelques concepts et processus mathématiques requis	L'étudiant fait appel à la plupart des concepts et des processus mathématiques requis	L'étudiant fait appel aux concepts et aux processus mathématiques requis
	Proposition de solution	Produit une démarche inappropriée ou peu appropriée comportant plusieurs erreurs conceptuelles ou procédurales majeures	Produit une démarche partielle comportant des erreurs conceptuelles ou procédurales	Produit une solution comportant quelques erreurs mineures ou peu d'erreurs conceptuelles ou procédurales	Produit une solution exacte ou comportant peu d'erreurs mineures

TABLE 2 – Exemple de grille critériée.

3.3 Structuration progressive des niveaux d'apprentissage visés : taxonomie SOLO

Une étude conduite au sein d'une variété de domaines académiques a montré de fortes similarités dans la manière dont progressent les étudiants. Il y a deux changements principaux : quantitatif, lorsque la quantité de détails dans la réponse de l'élève augmente, et qualitatif, lorsque ces détails sont intégrés dans un modèle structurel. Les étapes quantitatives de l'apprentissage se produisent d'abord, puis l'apprentissage change qualitativement, avec une complexité structurelle croissante.

Une approche à été développée, basée sur une taxonomie nommée SOLO (*Structure of the Observed Learning Outcome*), pour rendre compte de cette progression et permettre une description systématique de la complexité croissante de la performance d'un apprenant lorsqu'il maîtrise de nombreuses tâches académiques. Elle peut conjointement être utilisée pour décrire les résultats d'apprentissage visés par le cours, et pour évaluer les résultats d'apprentissage afin de savoir à quel niveau les étudiants se situent réellement.

Sans entrer dans les détails, cette progression dans les apprentissages est décrite au moyen de cinq niveaux. Les deux premiers (préstructurel et unistructurel) ne dépassent guère la compréhension de la terminologie et de concepts pris indépendamment les uns des autres, les trois suivants (multistructurel, relationnel, et

d'abstraction approfondi) permettant de passer progressivement du niveau quantitatif au qualitatif, allant pour le plus élevé jusqu'à démontrer une capacité à aller plus loin que ce qui a été enseigné et à proposer une réponse personnelle et cohérente. La taxonomie SOLO décrit cette **hiérarchie de niveaux**, dans laquelle chaque niveau d'apprentissage devient la base sur laquelle se construit le niveau suivant. Chaque niveau contient donc le niveau inférieur avec un « petit plus ».

Cette distinction entre « savoir davantage » et « savoir mieux » correspond à deux objectifs majeurs d'un programme académique : accroître les connaissances (quantitatif : unistructurel devenant de plus en plus multistructurel), et approfondir la compréhension (qualitatif : relationnel, puis abstrait étendu). La taxonomie SOLO classe les résultats d'apprentissage en fonction de leur qualité structurelle, ce qui la rend utile pour définir des niveaux de compréhension progressifs des AAV, et donc servir de base à la définition des éléments d'appréciation d'une grille critériée.

3.4 Évaluation des AE, des AAV ou des compétences ?

Une dernière question peut se poser concernant ce que l'on évalue réellement entre activités d'évaluation et acquis d'apprentissage visés. Généralement il est difficile, et pas forcément souhaitable, de faire correspondre de manière biunivoque les AE avec les AAV. Cependant une évaluation qualitative cohérente permet d'évaluer des AAV répartis sur plusieurs AE, et donc de faire un retour précis aux étudiants sur la réalité de leurs acquisitions.

Activité	AAV 1	AAV 2	AAV 3	AAV 4	AAV 5	Résultat
projet 1			A+	A+	B+	**A**
projet 2			A–	A	B	**A–**
projet 3			A	A–	A–	**A–**
test intermédiaire	C+	C–	C			**C**
examen final	C	C+	B			**C+**
participation					A	**A**
Résultats aux AAV	**C**	**C+**	**B**	**A**	**A–**	**B**

TABLE 3 – Exemple d'évaluation conjointe d'AE et AAV.

Sur l'exemple du tableau 3 (adapté de [Biggs & Tang, p342]), chacune des six AE présentées en ligne contribue à l'évaluation d'un ou plusieurs des cinq AAV présentés en colonne. Si la matrice est construite de manière cohérente, il y a deux façons différentes de parvenir à la note finale (ici B) :

— à partir des résultats obtenus à chaque AE : A, A–, A–, C, C+ et A (dernière colonne du tableau),

— à partir de ceux obtenus pour chaque AAV : C, C+, B, A– et A (dernière ligne du tableau).

Il devient ainsi possible, si on le souhaite, d'évaluer conjointement les AE et les AAV. Cela permet de mettre en œuvre une évaluation en termes d'AAV qui soit plus fine qu'un simple résultat « tout ou rien » qui découlerait de la note finale à l'enseignement. Cela peut par conséquent permettre une évaluation également plus précise en termes de compétences en reliant de manière plus précise les résultats aux UE à l'évaluation des AAV, par exemple pour une certification, et cela sans surcout d'évaluation ni pour l'enseignant ni pour l'étudiant.

4 Deux exemples de notation par évaluation qualitative

4.1 Notation qualitative et multicritère d'une AE

Ce premier exemple [Biggs & Tang pp.241] propose une manière d'évaluer et de noter de manière qualitative, mais de pouvoir rendre compte de manière quantitative. Il s'agit de l'évaluation d'un travail portant sur la présentation d'une argumentation (éléments pour, contre et conclusions), décrit par un AAV correspondant au verbe « expliquer ». Le système de notation est un système à cinq niveaux et les notes correspondantes sont exprimées au moyen des lettres A,B,C,D et F, avec des possibilités de modulation autour de chaque note.

La notation est décomposée en quatre critères : introduction, argumentation, conclusions et références. Les

		D	C– C C+	B– B B+	A– A A+
Points de pourcentage		1-3	5-7	9-11	13-15
Introduction	15%	Suffisant pour dire quoi traite le sujet mais faible prise en compte des priorités	Décrit le sujet, fait référence aux travaux antérieurs et à la proposition de cette contribution	Comme pour C, mais en montrant ce que les travaux antérieurs ont fait ou pas ; progression logique vers le sujet	Exposé intéressant et élaboré de la raison d'être de ce sujet et des questions à traiter, avec un avant-goût de la contribution originale.
Points de pourcentage		4-20	24-28	32-38	42-50
Argumentation	50%	Quelques points pertinents dans les listes de descriptions, consistant principalement en éléments pour ou contre	Liste des points les plus pertinents et des avantages et des inconvénients, mais difficulté à présenter un argumentaire convaincant.	La plupart de (tous les) points pertinents sont traités ; utilise une structure appropriée pour résoudre les question avec une argumentation convaincante	Similaire à B, mais présente un argumentaire original et personnel, bien étayé par des ressources/références allant bien au-delà de la littérature courante
Points de pourcentage		2-4	7-11	13-17	18-20
Résumé et conclusions	20%	Le résumé est réduit à une liste d'éléments pour ou contre aboutissant à une conclusion bancale	Le résumé distingue les différents éléments mais dans recul, la conclusion est faible et hâtive	Le résumé est équilibré et mène à la conclusion avec un bon raisonnement.	Le résumé aboutit à une conclusion surprenante ou originale qui soulève de nouvelles questions.
Points de pourcentage		1-3	5-7	9-11	13-15
Références	15%	Peu d'éléments, peu de preuves de compétences bibliographiques, mise en forme incorrecte	Preuve d'une certaine capacité de recherche bibliographique, références standard dans un format généralement correct	Complet, montrant le soin apporté au traitement de la question à résoudre, format correct et présentation claire	Similaire à B, mais utilise des références inhabituelles pour étayer un argument original.

TABLE 4 – Exemple de grille critériée avec points de pourcentage pour une évaluation qualitative multicritère.

éléments d'appréciation sont ici définis selon la taxonomie SOLO (voir section 3.3), c'est-à-dire en traduisant une progression structurelle des niveaux d'acquisition. Une proposition de grille critériée correspondant à cette situation est présentée TABLE 4.

De plus cet exemple montre comment il peut être possible de pondérer les différents critères en fonction de leur importance au moyen de pourcentages, avant de combiner leurs notes respectives pour obtenir la note finale. On enrichit pour cela la grille critériée au moyen de points de pourcentage utilisés pour effectuer les recombinaisons et le calcul de la note finale. Sur l'exemple, les quatre critères sont respectivement pondérés par les pourcentages 15%, 50%, 20% et 15%.

Les lignes « *Points de pourcentage* » montrent la manière qui a été retenue pour faire correspondre un nombre de points de pourcentage à chaque niveau de chaque critère. Par exemple le critère « Introduction » ayant reçu la pondération 15%, les points sont attribués à chaque note sur une échelle de 0 à 15, en l'occurrence 14 pour A, 10 pour B, 6 pour C et 2 pour D. La note F reçoit implicitement la valeur 0 et les modulations se traduisent par des variations de un point autour de la valeur de la note (par exemple autour de A les notes A+ et A− reçoivent respectivement les valeurs 15 et 13). L'écart entre les notes est volontairement plus grand que l'écart entre les niveaux au sein d'une même note, afin de souligner que l'obtention d'une note est plus importante que l'obtention d'une « bonne note » au sein d'une même note.

Lors de la notation, une note est tout d'abord attribuée qualitativement à chaque critère en fonction des éléments d'appréciation de la grille critériée. Par exemple, si l'introduction décrit le sujet, fait référence à des travaux antérieurs en les évaluant de façon superficielle et se poursuit par l'exposé du cas présent, sans progression logique du sujet, cela correspond aux critères C en général, disons C+. Cette note est ensuite convertie en un nombre de points de pourcentage en utilisant la table de correspondance, ici 7.

Chaque critère est évalué de cette manière et les scores de tous les critères sont ensuite additionnés entre eux. Une seconde table de correspondance (voir TABLE 5) est alors utilisée pour reconvertir ce nombre de points de pourcentage en une note littérale. Supposons qu'un étudiant obtienne un total de 67 pour cette AE. La valeur la plus proche de ce total dans la table de correspondance est 68, soit B qui sera la note globale attribuée à cette activité.

≤ 45	46-50	52	55	60	65	68	70	75	80	> 80
Échec	D	C−	C	C+	B−	B	B+	A−	A	A+

TABLE 5 – Table de correspondance entre points de pourcentages et notes.

Cet exemple nécessiterait probablement d'être adapté en fonction du champ disciplinaire ou du contexte de l'enseignement mais le principe général demeure valable. Les choix qui ont été effectués sont assez arbitraires, mais ce sera toujours le cas lorsqu'on utilise des nombres pour quantifier des données qualitatives.

Il peut ensuite y avoir différentes manières de combiner les notes des différentes activités d'un enseignement pour obtenir la note finale d'un enseignement (ou d'AAV comme cela est évoqué section 3.4 et TABLE 3), la plus simple étant de repasser par une échelle numérique telle l'échelle GPA en vigueur aux USA (voir section 5).

4.2 Le modèle COGS [Green & Emerson]

Cette section présente le modèle décrit dans l'article : Kris H. Green and W. Allen Emerson. A new framework for grading. Fisher Digital Publications : `fisherpub.sjfc.edu/math_facpub/3`*, 2007.*

Ce second exemple décrit une méthode d'évaluation littérale pour lequel le principe de combinaisons des notes de plusieurs AE ne nécessite pas de passer par une conversion numérique (voir fiches pour une description détaillée). Elle repose sur un double postulat de départ destiné à permettre d'uniformiser toutes les évaluations :

— chaque AE peut être notée selon exactement trois catégories de critères (voir exemples TABLE 6) ;
— chaque catégorie de critères est appréciée selon seulement deux niveaux de validation : E qui correspond au niveau attendu («*expected*») pour valider, et I qui traduit un niveau remarquable («*impressive*»), ce qui correspond par conséquent à une échelle à trois notes si l'on ajoute une note traduisant l'échec.

Chaque catégorie de critères sera évaluée plusieurs fois tout au long du cours ce qui permet de faire un retour qualitatif à l'étudiant qui peut apprécier la manière dont il progresse dans cet enseignement.

Cours	Catégorie de critères	Description
Cours de mathématiques pour les étudiants en commerce	Mécanique et Technique	Les définitions et calculs mathématiques de base du cours, ainsi que les techniques informatiques (Microsoft Excel, en grande partie) nécessaires au cours
	Analyse et raisonnement	La planification de solutions à des problèmes complexes et le développement logique d'analyses pour des problèmes commerciaux réalistes
	Communication et professionnalisme	La rédaction et la présentation des solutions aux problèmes, ainsi que les attitudes et comportements des étudiants (assiduité, travail en groupe, etc.)
Cours d'histoire	Sources	La collecte, l'évaluation et l'incorporation des sources d'information
	Analyse	L'utilisation des preuves pour présenter un argument logique
	Communication	La grammaire, le style et la présentation de l'argumentation

TABLE 6 – Exemples de décomposition d'enseignements selon trois critères d'évaluation.

L'outil servant à structurer la notation d'un enseignement est par conséquent un tableau à trois lignes (les catégories de critères) et deux colonnes (les niveaux de validation). L'enseignant met dans chaque case entre trois et dix critères formulés de manière positive et coche ensuite les critères auxquels le travail répond, par exemple avec un « X » pour une réussite complète et un « / » pour une réussite partielle. Il peut ensuite déterminer si le résultat à l'AE se situe dans le niveau de performance attendu ou remarquable dans chacune des catégories. La matrice ainsi remplie fournit à la fois une note et un moyen efficace de retour d'information à l'étudiant.

La seconde partie de la méthode décrit une manière d'agréger toutes les notes de chaque catégorie sans recourir à une conversion quantitative. Pour chaque catégorie, la note globale (E ou I) est attribuée en fonction de la majorité des notes qu'il a obtenues à l'ensemble des évaluations (majorité de E ou de I). Les notes globales des trois activités sont enfin elles-même combinées pour retrouver une notation plus classique à cinq niveaux. La condition d'admission est que chaque catégorie soit validée avec au moins un E et la proposition de correspondance est la suivante :

Notes globales aux trois catégories de critères	Note finale
III	A
IIE	B+
IEE	B
EEE	C
n'obtient pas au moins E partout	D ou F

Ce modèle est assez contraignant quant aux hypothèses de départ (exactement trois catégories évaluées sur seulement deux niveaux) et la justification de ces contraintes est plus de nature combinatoire (obtenir un nombre convenable de notes, c'est-à-dire proche de 5) pour l'agrégation finale que pour des raisons pédagogiques. La possibilité de généraliser sa mise en œuvre peut donc interroger. Mais cet exemple a cependant le mérite de montrer comment on peut réaliser une évaluation qualitative qui permette de faire aux étudiants des retours riches d'informations sur leurs évaluations et de proposer un principe pour agréger des notes sans recourir à une conversion quantitative.

5 Agrégation de notes exprimées qualitativement

Si l'on se concentre sur un système de notation à quatre ou cinq lettres, le système américain d'utilisation d'une échelle GPA (*Grade Point Average*) permet de produire assez finement une moyenne permettant :

— soit d'agréger les notes des différents cours pour produire une note moyenne finale ;
— soit de produire une moyenne plus précise pour répondre à un besoin de classement.

Par exemple avec un système à cinq lettres, chaque lettre est associée à un chiffre de 0 à 4 (0 pour F jusqu'à 4 pour A). Les modulations autour des lettres se font ajoutant ou retranchant 0.3 points (voir TABLE 7).

La moyenne est alors calculée par la conversion numérique de la note de chaque cours. Rien n'interdit de pondérer ces notes par des coefficients, et la méthode peut également *a priori* être utilisée pour agréger plusieurs notes au sein d'un même cours.

On peut noter que ce calcul de moyenne ne présente pas les biais du calcul d'une moyenne de notes sur 20. En effet les distances entre les notes respecte une appréciation qualitative des résultats et il est donc acceptable de leur faire jouer le rôle d'une mesure (contrairement aux notes sur 20).

Échec	D	C−	C	C+	B−	B	B+	A−	A	A+
moyenne finale	1	1.7	2	2.3	2.7	3	3.3	3.7	4	4.3

TABLE 7 – Le système *grade-point average* (GPA) de calcul de moyenne à partir de notes qualitatives littérales.

6 Conclusions

Ce premier livrable précise les objectifs du projet EQAE en les situant par rapport aux travaux et avancées en sciences de l'éducation et pédagogie universitaire. Il met en avant les principes et constats suivants :

— un système d'évaluation des acquis doit permettre conjointement (1) de faire aux étudiants un retour sur la qualité de leurs acquisitions et (2) de les noter afin de valider la réussite à un enseignement ou à un programme d'étude ;
— le premier des deux modèles d'évaluation, basé sur le principe que l'on peut mesurer quantitativement le niveau des étudiants et les comparer entre eux, n'est pas adapté à l'objectif (1) et présente de nombreux défauts théoriques et méthodologiques qui biaisent les réponses à l'objectif (2) ;
— le second, basé sur l'appréciation individuelle des performances de chaque étudiant par référence à un ensemble de critères précis et connus, répond aux deux objectifs ; il soulève la difficulté pour remplir l'objectif (2) de pouvoir agréger des appréciations qualitatives pour produire une note générale, mais il existe différentes solutions pour y parvenir ;
— une mise en œuvre maîtrisée de l'évaluation qualitative repose sur l'application des principes de l'alignement pédagogique (ou alignement constructif).

Il dresse ensuite un rapide état de l'art des systèmes de notation universitaire dans le monde et synthétise brièvement ce qui leur est commun et ce en quoi ils diffèrent :

— il existe une forte convergence des modèles vers un système de notation basé sur quatre ou cinq niveaux, généralement associés à des lettres, avec le cas échéant des modulations permettant de majorer ou de minorer certains de ces niveaux ;
— ces notes rendent compte d'une appréciation qualitative du résultat obtenu ;
— la manière de conduire l'évaluation pour construire cette appréciation peut varier entre méthodes qualitatives et méthode quantitatives ;
— la plupart de ces systèmes proposent des échelles de correspondance numérique, souvent des pourcentages, associées aux notes qualitatives, et il existe une très grande variabilité dans la définitions de ces pourcentages, ce qui ne fait probablement que traduire la subjectivité que présente l'établissement de cette correspondance.

Ces différents éléments plaident pour rechercher des exemples de systèmes d'évaluation et de notation satisfaisant les deux objectifs suivants :

— appréciation de la performance des étudiants effectuée de manière qualitative ;
— traduction du niveau de performance démontré par une note exprimée sur quatre ou cinq lettres modulables ;
— possibilité d'agréger plusieurs notes pour produire une note générale à un enseignement ou une note finale à un programme de formation.

La dernière partie de ce livrable présente plusieurs exemples correspondant à ces objectifs et destinés à servir de base à la phase suivante du projet, celle de la coconstruction d'un modèle d'évaluation qualitative des acquis des étudiants qui pourrait être partagé et mis en œuvre par les formations de l'UB.