*À propos de l'évaluation des apprentissages, de la « mesure » et des notes**

Cécile D'Amour
Membre du Groupe de recherche-action
PERFORMA

La réforme de l'enseignement collégial entraîne des modifications dans la façon de concevoir et d'exercer la responsabilité de l'évaluation des apprentissages. Pour élaborer, dans chacun des cours, une stratégie générale d'évaluation des apprentissages qui soit pertinente, il y a lieu de clarifier des principes et de se doter d'une méthodologie et d'instruments qui soient appropriés. Il y a là un vaste chantier auquel des personnes et des groupes du réseau s'attaquent présentement, de façon plus ou moins concertée[1].

Pour élaborer la stratégie d'évaluation à utiliser dans un cours, il est utile, notamment, de connaître et de comprendre le cadre dans lequel s'exerce la responsabilité de l'évaluation des apprentissages, cadre dont certaines composantes sont communes au réseau collégial, et d'autres, propres à l'établissement. On peut ainsi saisir les contraintes et la marge de manœuvre, procéder à une application réfléchie et critique des règles du jeu, et, éventuellement, contribuer à améliorer ces règles qui, forcément, du fait que le chantier évoqué ci-dessus est en cours, ne sont pas toutes cohérentes avec les principes qui devraient guider l'évaluation des apprentissages dans un contexte d'approche par compétences.

Nous allons ici nous attarder à présenter et à critiquer une double balise relative à l'évaluation des apprentissages dans le cadre des cours : le choix de l'échelle en pourcentage et celui de la note 60 % pour représenter le seuil de réussite. Nous évoquerons également quelques avantages d'un autre choix possible : celui de recourir à une échelle par niveaux descriptifs. À partir des critiques et possibilités présentées, chacun pourra poursuivre sa réflexion sur des modalités d'évaluation qui seraient plus appropriées. Quant à nous, nous creusons la question et proposerons des avenues d'ici peu.

Une tradition à critiquer

L'utilisation de l'échelle en pourcentage et le choix de 60 % comme note de passage sont deux caractéristiques traditionnelles de l'évaluation des apprentissages au collégial. Elles sont maintenues par l'article 27 du Règlement sur le régime des études collégiales qui a été édicté par décret du gouvernement, le 14 juillet 1993.

En une douzaine de mots, l'énoncé central de l'article 27 exprime plusieurs décisions dont la pertinence, dans un contexte d'approche par compétences, est à tout le moins discutable. En effet, dire que « la note traduisant l'atteinte minimale des objectifs du cours est de 60 % », cela signifie que l'échelle adoptée est l'échelle en pourcentage, une échelle quantitative, donc, qui comporte un très grand nombre d'échelons et un maximum absolu ; cela signifie également que le seuil de réussite est fixe et qu'il est situé relativement bas sur l'échelle.

Nous allons examiner ces différents choix pour mettre en lumière les défauts qu'ils comportent, à notre avis. Il importe de souligner ici que les caractéristiques examinées qui constituent des défauts dans un contexte d'approche par compétences n'étaient guère souhaitables dans le contexte qui prévalait auparavant, mais qu'elles sont encore plus déplorables lorsqu'on vise le développement de compétences et qu'on doit attester les niveaux de ce développement.

Une échelle quantitative

L'expression de données sous forme quantitative amène à utiliser des opérations mathématiques comme le calcul de la moyenne, de l'écart-type ou de la cote Z. Ces opérations qui peuvent toutes être effectuées sur les nombres (ceux qu'on dit réels) ne peuvent pas toutes être utilisées pour traiter des données chiffrées traduisant des phénomènes concrets.

En effet, lorsqu'on traduit des données à l'aide d'une échelle, il importe de s'assurer que les opérations qu'on fait sur les notes ont du sens, *compte tenu des caractéristiques de la variable considérée*.

Dans certains cas, on peut mettre en ordre les différentes valeurs de la variable, comme pour la variable « racisme », par exemple, pour laquelle on peut définir des degrés. Dans ces cas, on dit qu'on utilise une échelle **ordinale**. Dans d'autres cas, il est seulement possible de distinguer les valeurs de la variable les unes des autres, comme c'est le cas pour la variable « sexe », par exemple. On a alors affaire à ce qu'on appelle une échelle **nominale**.

Pour certaines variables, on peut faire plus que d'en ordonner les valeurs : on peut assurer que deux écarts de même grandeur numérique correspondent bien à une même variation dans la grandeur du phénomène (on dit alors qu'il s'agit d'une échelle **d'intervalles**). Dans certains cas, on peut même établir des rapports

* Merci à trois collègues professeurs de mathématiques qui ont jeté leur regard critique sur cet article : René Chapleau et Jacques Dion, du collège Ahuntsic et Bernard Massé, du collège Joliette-De Lanaudière.

entre diverses valeurs de la variable ; on peut dire, par exemple, que telle valeur de la variable est égale au double ou au triple de telle autre valeur (on parle alors d'une échelle **de rapport**).

La température est un bon exemple de variable mesurée sur une échelle d'intervalles, sans qu'on puisse utiliser une échelle de rapport. En effet, un écart de température de 12 degrés correspond à une même variation physique de l'objet dont la température est mesurée[2], mais on ne peut pas dire que 12 degrés soit le double de 6 degrés ou, si l'on veut, qu'il fait deux fois plus chaud à 12 degrés qu'à 6 degrés.

Donc, quand on veut traiter ou faire parler des données obtenues à partir de situations concrètes, certaines opérations mathématiques ont du sens et d'autres n'en ont pas, selon le type d'échelle sur lequel la variable est mesurée[3]. Les quatre types d'échelle se classent dans l'ordre suivant pour ce qui est de l'utilisation d'opérations progressivement plus nombreuses et plus puissantes : l'échelle nominale, l'échelle ordinale, l'échelle d'intervalles et l'échelle de rapport.

Pourtant, dès que des nombres apparaissent, le commun des mortels est prêt à sortir l'arsenal des méthodes quantitatives. Réaction normale pour qui n'a pas eu l'occasion de saisir véritablement la portée et les limites des instruments mathématiques.

Quelques utilisations abusives

L'utilisation abusive des méthodes quantitatives dans le cadre de l'évaluation des apprentissages se manifeste principalement de deux façons.

La première est le recours à des instruments mathématiques dont la puissance ne s'accorde pas au type d'échelle utilisée. Par exemple, on utilise la moyenne arithmétique, qui exige une échelle d'intervalles, alors que les résultats de l'évaluation sont situés sur une échelle ordinale. En effet, si l'on peut assurer que tel résultat d'évaluation est plus (ou moins) élevé que tel autre, qu'il est plus (ou moins) élevé que le seuil de réussite[4] (ce qui signifie que la variable est repérée sur une échelle ordinale), on ne peut garantir qu'un même écart numérique entre deux notes correspond à une même différence de niveau de maîtrise (si nous pouvions le garantir, nous travaillerions avec une échelle d'intervalles). Ainsi, on ne peut dire que la différence entre 55 % et 72 % est la même que la différence entre 63 % et 80 %, même si, dans les deux cas, l'écart est de 17 points de pourcentage.

L'autre type d'utilisation abusive d'opérations mathématiques est le recours à l'opération d'addition (pondérée) pour combiner divers résultats d'évaluation en vue de traduire un jugement d'ensemble. Ce type d'abus se produit lorsqu'on construit la note finale d'un cours (ou même la note totale d'un examen) en additionnant des résultats partiels d'évaluation auxquels on a donné des poids identiques ou différents[5]. Cette façon de faire comporte, de façon intrinsèque, le risque que la note obtenue soit au-dessus du seuil de réussite même si des éléments essentiels de l'apprentissage visé n'ont pas été suffisamment maîtrisés.

De plus, si les divers résultats qui sont additionnés pour construire la note finale ont été obtenus à des moments différents, un autre risque de distorsion est présent : il peut arriver que la note d'un élève soit sous le seuil de passage alors que cet élève, au terme du cours, a maîtrisé l'apprentissage visé au-delà du niveau exigé pour obtenir la note de passage. Cette situation est vécue par bien des élèves qui font montre de persévérance mais dont l'apprentissage démarre lentement. Ils sont victimes du fait que les évaluations dont les résultats ont été utilisés pour construire la note finale auraient plutôt dû être de type formatif, puisqu'elles ont été effectuées durant la démarche d'apprentissage, alors que l'apprentissage était encore possible. En additionnant des résultats « datés », on invalide de façon fondamentale le jugement exprimé par la note finale ainsi obtenue.

Dans ce contexte, il serait plus prudent (pour éviter les manipulations non rigoureuses et le manque de signification des notes qui en découlent) de retenir une échelle où les échelons sont indiqués par des lettres plutôt que par des chiffres.

Une précision excessive

Même si tous les « producteurs et usagers » de notes résistaient à la tentation d'utiliser les instruments mathématiques de façon abusive, le choix de l'échelle en pourcentage comporterait encore un défaut majeur dans le cadre de l'évaluation des apprentissages : sa trop grande précision.

Cela n'a pas de sens d'utiliser une échelle « de mesure »[6] dont la précision est plus grande que celle dont peuvent nous assurer l'instrument de mesure et la façon dont cet instrument est utilisé. Or, il est clair que les épreuves qu'on utilise couramment et la façon dont on les juge ne peuvent nous permettre d'assurer que des notes qui diffèrent par un point de pourcentage reflètent bien des résultats d'apprentissage différents (par exemple, assurer que 73 % représente bien un apprentissage mieux maîtrisé que celui qui a permis d'obtenir 72 %). Cette évidence est encore plus nette quand on parle d'évaluer le degré de maîtrise d'un objectif multidimensionnel et intégrateur (comme le développement d'une compétence, par exemple).

On constate d'ailleurs quelques pratiques relativement courantes qui donnent à penser que cet excès de précision est reconnu. Dans certaines politiques institutionnelles d'évaluation des apprentissages, par exemple, on prévoit l'inscription au bulletin d'une seule et même note – 30 % – pour tous les étudiants qui ont obtenu une note finale entre 0 % et 30 %. On observe, par ailleurs, qu'en pratique, bon nombre d'enseignants ne laissent pas de notes finales entre 55 % et 60 %, considérant, avec raison, que la validité des instruments qu'ils ont utilisés et la précision des modalités de correction ne sont pas suffisantes pour garantir la précision de la note finale à 5 points près.

Notons qu'en attendant l'adoption d'une échelle plus appropriée, il est possible de pallier la précision excessive de l'échelle en pourcentage en choisissant de n'utiliser qu'un petit nombre d'échelons. Il s'agit simplement de s'entendre sur la « traduction chiffrée » des différents niveaux (assez peu nombreux) de l'échelle retenue.

Un maximum absolu

Rappelons qu'en évaluation des apprentissages, la « note » n'a pas de sens ni d'intérêt pour elle-même : elle en a uniquement comme représentation symbolique du degré de maîtrise de l'apprentissage visé – et non du degré de réussite à une épreuve (ou de la somme des résultats à différentes épreuves).

Le résultat d'une épreuve peut être parfait, mais le résultat espéré d'un apprentissage – surtout s'il s'agit d'un apprentissage complexe – ne peut comporter, lui, un maximum absolu. On ne peut pas dire d'un élève qui a réussi un examen de mathématiques ou une dictée à 100 % qu'il maîtrise à 100 % les apprentissages que cet examen ou cette dictée cherche à évaluer. On pourra dire toutefois que sa maîtrise des apprentissages visés appartient à une catégorie supérieure (notée par un symbole ou par un autre).

Et dire qu'un apprentissage est maîtrisé bien au-delà du seuil de réussite n'a pas la même connotation, en termes de perfection, que de donner une note de 100 % (ou de 20 sur 20 ou de 50 sur 50).

Les avantages d'une échelle par niveaux

En conséquence des critiques formulées précédemment, il apparaît préférable de choisir une échelle où chaque symbole représente une « zone de maîtrise » plutôt qu'une valeur unique. Il s'agit alors, en examinant le résultat d'une épreuve à l'aide d'une combinaison[7] de critères d'évaluation, de juger dans quelle zone de maîtrise se situe l'apprentissage effectué par un étudiant.

Signalons deux autres avantages du recours à une échelle par niveaux.

D'abord, on évite ainsi des faux débats comme celui qui tourne autour de la possibilité ou non de mettre une note parfaite en philosophie, en arts ou en composition, par comparaison à ce qui est possible en mathématiques ou en droit, par exemple. Dans tous les cas, le degré de maîtrise des apprentissages peut être plus ou moins grand (ou petit) et, dans tous les cas aussi, les épreuves utilisées pour porter le jugement reflètent une certaine compréhension de ce que signifie être « maître » dans la discipline ou le champ d'intervention considéré.

Le recours à une échelle par niveaux permet également de régler bon nombre des difficultés entraînées par le choix d'un seuil de réussite fixe comme 60 %, seuil toujours situé aux six dixièmes d'une échelle ayant un maximum absolu. Voyons un peu de quoi il retourne.

On peut penser que, dans un programme d'études, les apprentissages visés n'ont pas tous à être maîtrisés au même niveau.

Par exemple, selon la compétence professionnelle visée, telle ou telle compétence particulière devra être plus ou moins développée. Ainsi, le seuil minimal pour la compétence en diagnostic d'un médecin qui pratique les soins d'urgence devrait être plus élevé que pour un généraliste qui fait du bureau. Et le seuil minimal de compétence en relations humaines devrait être plus élevé pour la personne diplômée en soins infirmiers que pour celle qui se dirige en informatique, par exemple.

Il serait donc utile, pour représenter le seuil de réussite, d'avoir un symbole unique dont le sens général soit commun (niveau minimal de maîtrise des apprentissages qui permet de poursuivre les études ou de faire preuve de compétence suffisante pour entrer sur le marché du travail) mais dont le « contenu » précis et le degré d'exigence soient fonction du contexte.

En guise de conclusion

Lorsque ces critiques et avenues sont abordées avec des professeurs ou des conseillers pédagogiques, elles provoquent un remue-méninges certain qui amène une remise en question de ces choix d'échelle et de seuil de réussite.

Cette remise en question vient déranger des pratiques ancrées très profondément, notamment celles de construction de la note finale à partir d'une addition.

Puisque la réforme nous oblige à clarifier des principes et des pratiques, pourquoi ne pas en profiter pour faire le grand ménage et adopter des mesures qui seront pertinentes ? Ce qui nous éviterait, notamment, d'avoir à traficoter (les seuils, les notes, les opérations mathématiques…) pour obtenir une note finale qui reflète un jugement d'évaluation significatif et fiable.

En présentant ce qui précède, nous espérons avoir donné un coup de balai utile dans le cadre d'un tel grand ménage… ▨

NOTES ET RÉFÉRENCES

1. Notamment dans PERFORMA, deux groupes de travail sur l'évaluation : l'un sur l'évaluation dans les cours et l'autre sur l'épreuve synthèse de programme.
2. Il faut la même quantité d'énergie pour faire passer un liquide, par exemple, de 20 à 32 degrés, de 44 à 56 degrés, de 98 à 110 degrés.
3. Notons que, dans ce contexte, le terme « mesure » est pris dans un sens très large et que la mesure, alors, peut ne pas être quantitative du tout.
4. Et ce n'est pas là une mince affaire puisque sont en cause ici la validité de l'instrument d'évaluation et la fiabilité du jugement.
5. Pratique soutenue par l'instrument de planification de l'évaluation qu'on appelle « table de spécification ».
6. Nous utilisons les guillemets pour indiquer que tout n'est pas clair au royaume de la mesure en éducation. Ne serait-ce que parce que «mesure » évoque généralement « quantification » avec tous les dangers que ce rapprochement comporte.
7. Les critères sont combinés de façon à établir des niveaux descriptifs plutôt que d'être pondérés mathématiquement. Cette façon de faire est abordée par Ulric Aylwin dans « L'évaluation globale de la qualité des textes », *Pédagogie collégiale*, vol. 7, n° 4, mai 1994, p. 13 à 15.