Sur l'évaluation de compétences en mathématiques

par Jacques BAIR

En éducation, l'évaluation occupe une place importante : elle a toujours été par le passé, est encore à l'heure actuelle et sera aussi certainement dans le futur, l'objet de nombreuses études. La tendance contemporaine d'axer l'enseignement sur le concept de compétences ne simplifie pas cette problématique.

Dans l'enseignement des mathématiques, les compétences disciplinaires sont de deux types, à savoir les compétences en relation avec les contenus mathématiques et les compétences en relation avec les processus mathématiques, qui doivent être pris en compte simultanément lors de l'évaluation de productions des élèves. Les performances de ceux-ci doivent tenir compte de connaissances, de savoir-faire et de capacités, dont l'évaluation peut se faire de façon assez traditionnelle, mais aussi de savoir-être, d'habiletés, d'aptitudes et d'attitudes dont l'évaluation est assurément moins coutumière et plus délicate.

Ce document comprend deux parties.

Tout d'abord, un texte est consacré à des généralités sur la mesure de compétences, ainsi qu'à la présentation, illustrée par des exemples concrets, de différents types d'échelles (universelle, descriptive, globale) et de grilles d'évaluation.

Enfin, ce travail se termine par la reproduction de diapositives (initialement réalisées à l'aide du logiciel Power Point) préparées pour accompagner une leçon sur l'évaluation ; l'exposé oral correspondant a été donné, dans le cadre de la FOPED (Formation Pédagogique), à plusieurs cohortes de stagiaires en mathématiques à l'Université du Luxembourg [1].

Première partie : Différents types d'échelles

Nous allons essentiellement mettre en évidence quelques types d' « échelles» qui peuvent être utilisées pour « mesurer » certaines variables comme celles évoquées ci-dessus. Il s'agit d'un outil qui est aussi fréquemment exploité dans certaines situations concrètes relevant de

[1] Au Grand-Duché de Luxembourg, la FOPED est organisée par l'Université du Luxembourg et concerne des jeunes enseignants, diplômés dans leur discipline et travaillant déjà dans des établissements du secondaire. Ces professeurs doivent effectuer des stages, suivre des cours en sciences de l'éducation et réaliser un mémoire afin de pouvoir devenir un enseignant-fonctionnaire dans le pays. C'est, grosso modo, l'équivalent de l'Agrégation de l'enseignement secondaire en Belgique.

sciences humaines telles que la psychologie, les sciences sociales, la médecine, le marketing, … Nous profiterons de cette occasion pour donner, souvent au sein de notes en bas de pages, des indications succinctes sur des techniques quantitatives pouvant s'avérer intéressantes dans ce contexte.

1. Généralités sur le concept de mesure

Avant d'exhiber quelques exemples de types d'échelles dans l'évaluation de compétences en mathématiques, nous allons émettre quelques considérations générales sur la « mesure » d'attitudes, d'opinions, de jugements, …

Nous nous intéressons ici à des productions d'élèves facilement observables, aussi bien qu'à différents processus mentaux, tels que des attitudes, jugements ou opinions, qui ne sont pas directement observables au travers de la manifestation de performances concrètes ou de comportements visibles, mais qui sont appréhendés par leur verbalisation au moyen de certains indicateurs particuliers. Par exemple, la motivation d'un apprenant peut être estimée par la perception que l'individu manifeste pour la valeur de l'activité concernée, mais aussi par la perception de sa propre compétence ou encore par celle de sa contrôlabilité face à la tâche envisagée [2].

Mesurer un processus décrit par un indicateur consiste, pour nous, à établir une relation entre le processus étudié et un symbole, choisi au sein d'un ensemble fini de possibilités ; ceci revient à construire une « échelle », dont les possibilités sont les « échelons » qui permettent d'avoir une idée sur l'appréciation concernant l'indicateur en question. Cette appréciation peut être, selon les cas, portée par l'enseignant ou par l'apprenant. Par exemple, lors de la résolution d'un problème de mathématiques, le professeur peut juger si un élève a été capable ou non de traduire un énoncé exprimé en langage naturel dans une formalisation mathématique correcte, tandis qu'un élève peut se déclarer très intéressé, indifférent ou pas intéressé par l'exercice proposé.

A un processus mental α étudié chez un apprenant peuvent être associés plusieurs indicateurs 1, 2, … qui sont « mesurés » par des *indicateurs* traduits par des symboles a_1, a_2, … respectivement choisis au sein d' « échelles » appropriées ; de la sorte, pour chaque critère est construite une *échelle* composée évidemment d'*échelons*. Ces derniers peuvent être des

[2] Voir notamment les travaux de Viau R. sur le sujet.

nombres, que nous choisirons le plus souvent entiers, qui peuvent être positifs, négatifs ou nuls selon les cas : ils seront nommés des « notes ». Signalons encore que ces notes traduisant les mesures sont quelquefois remplacées par des lettres A, B, …, ou par des symboles iconiques (« smiling faces ») ou encore par un point situé sur une droite graduée, …

En principe, plus les échelons sont nombreux, plus fine devrait être la mesure. A titre d'illustration, considérons une *échelle d'accord*, indiquant donc les degrés d'accord d'un élève face à une affirmation de l'enseignant ; ces derniers peuvent être en nombre variable ainsi qu'en attestent ces possibilités dont chacune affine certaine qui la précèdent :

- cas de 2 échelons : « d'accord ; pas d'accord »
- cas de 3 échelons : « d'accord ; indifférent ; pas d'accord »
- cas de 4 échelons : « tout à fait d'accord ; un peu d'accord ; un peu en désaccord ; tout à fait en désaccord »
- cas de 5 échelons : « tout à fait d'accord ; un peu d'accord ; indifférent ; un peu en désaccord ; tout à fait en désaccord »
- cas de 6 échelons : « tout à fait d'accord ; plutôt d'accord ; légèrement d'accord ; légèrement en désaccord ; plutôt en désaccord ; tout à fait en désaccord ».

Il convient toutefois de veiller à ce que les catégories retenues soient discriminantes : en effet, il est difficile, voire impossible pour un individu, de distinguer entre plusieurs niveaux d'appréciation fort proches ; par exemple, il peut être difficile pour quelqu'un de choisir la mention « légèrement d'accord » au lieu de celle « plutôt d'accord ».

Remarquons que les 5 échelles données ci-dessus à titre exemplatif semblent accorder une même importance aux mentions d'accord qu'à celles de désaccord. De la sorte, les notes associées aux différentes mentions peuvent être toutes positives : par exemple, l'échelle « tout à fait d'accord ; un peu d'accord ; un peu en désaccord ; tout à fait en désaccord » est traduite par les nombres respectifs « 4 ; 3 ; 2 ; 1 » ou encore par les nombres « 100 ; 75 ; 50 ; 25 » ; mais, pour tenir compte de la « symétrie » de l'échelle, elles peuvent former un ensemble de nombres positifs associés aux mentions d'accord et à leurs opposés pour les termes de désaccord correspondants : ainsi, l'échelle avec 4 mentions traitée ci-dessus peut être traduite par les notes respectives « 2 ; 1 ; -1 ; -2 » [3]. Dans une telle éventualité d' « échelles symétriques », mentionnons que celles avec un nombre impair d'items comprennent une mention « neutre » du type « ni en accord, ni en désaccord », ou encore « indifférent » ou « je ne sais pas », à laquelle est associée la note 0. L'absence d'un tel échelon 0, qui se retrouve

[3] Cet exemple met bien en évidence le fait que des moyennes, par exemple, sur de telles notes n'ont aucun intérêt.

donc dans les échelles avec un nombre pair d'échelons, oblige le sujet à devoir émettre un avis soit positif (avec un accord plus ou moins important) soit négatif (avec un désaccord plus ou moins prononcé) et ne peut dès lors pas s'abstenir ni émettre un avis d'indifférence.

2. Echelle universelle

Certaines échelles sont fort générales et peuvent être utilisées dans des situations variées.
Le cas le plus typique et aussi le plus courant est fourni par les *échelles numériques* : au phénomène envisagé est attribué une note, c'est-à-dire un nombre le plus souvent entier. En contexte scolaire, les systèmes cotation sont fort variables. Généralement, les performances des élèves sont évaluées à l'aide d'un nombre naturel variant entre un minimum (pour un mauvais résultat) et un maximum (pour un excellent résultat) ; ainsi, les Belges et les Français notent le plus souvent de 0 à 10 ou à 20, les Italiens de 0 à 30, les Suisses de Genève de 0 à 6, les Luxembourgeois de 1 à 60, les Autrichiens de 1 à 5, …
Mais les échelles peuvent également être *qualitatives*, lorsque les échelons se réfèrent à des « qualités ». En guise d'exemples classiques, voici quatre échelles prototypiques avec 4 ou 5 échelons :

Echelle d'excellence	**Echelle de fréquences**	**Echelle de satisfaction**	**Echelle de facilité**
Très mauvais	Jamais	Insatisfaisant	Très difficilement
Mauvais	Parfois	Plutôt satisfaisant	Difficilement
Moyen	Souvent	Satisfaisant	Facilement
Bon	Toujours	Très satisfaisant	Très facilement
Très bon		Sans avis	

De telles échelles générales sont encore qualifiées d'*universelles,* sont évidemment faciles à construire et à interpréter ; elles sont notamment exploitées dans des tests standardisés et conviennent bien pour figurer sur un bulletin scolaire. Mais cette universalité peut apparaître comme étant un inconvénient pour ce type d'instrument. En effet, même lorsque les jugements sont objectifs et concordants [4], ce qui est loin d'être garanti, les indications fournies par ces échelles ne sont pas toujours des plus intéressantes pour les sujets concernés : ainsi, lorsqu'un élève apprend que son activité est médiocre ou qu'il progresse difficilement

[4] Les mêmes élèves peuvent être évalués dans des situations différentes et / ou par des juges différents. Si l'on ne tient compte que d'une des deux « facettes » (celle des situations ou celle des juges), la concordance des avis peut être analysée à l'aide d'indices statistiques adéquats, dont les plus courants sont l'*indice kappa de Cohen* dans le cas d'évaluations qualitatives et l'*indice de corrélation de Kendall* dans le cas d'évaluations quantitatives ou ordinales. La théorie de la *généralisabilité* permet de traiter simultanément toutes les facettes pouvant être considérées.

dans le développement d'une compétence, il ne reçoit de la sorte aucun renseignement concret sur la nature des lacunes constatées, ni sur la manière d'essayer de les combler. Nous verrons ultérieurement comment fournir aux apprenants un feedback plus instructif.

3. Grille d'évaluation

On est souvent amené à considérer simultanément plusieurs échelles pour évaluer un même phénomène : celles-ci forment alors une *grille d'évaluation*.

Mentionnons deux types de grilles fréquemment utilisées dans des enquêtes pour évaluer des attitudes et / ou des opinions de patients en psychologie, de clients en marketing, de malades en médecine, … mais aussi d'apprenants en pédagogie.

Considérons tout d'abord le *différentiel sémantique* (encore nommé *différentiateur sémantique*) qui a été initialement développé par Osgood *et al* (1957) pour comparer les associations entre des stimuli caractérisés.

Dans le contexte scolaire, il s'agit d'inviter les étudiants à se situer « quelque part » entre deux adjectifs « antonymes bipolaires », c'est-à-dire de sens tout à fait opposés, par exemple « lent / rapide », « concentré / distrait », …Par extension, les adjectifs sont quelquefois remplacés par des phrases complètes, par exemple « je me lance dans des calculs dès la lecture de l'énoncé / je réfléchis beaucoup après la lecture de l'énoncé et avant d'effectuer des calculs ».

Concrètement, une liste de paires d'adjectifs (ou de phrases) est soumise aux apprenants avec un certain nombre de croix dont l'une doit être choisie et, par exemple, encerclée. Comme illustration, envisageons le cas où un professeur souhaite s'intéresser à la perception par ses élèves d'une famille d'activités ; un début [5] de différentiel sémantique pourrait se présenter comme suit :

Le problème à résoudre était :

➢ inutile	x	x	x	x	utile
➢ difficile	x	x	x	x	facile
➢ ennuyeux	x	x	x	x	intéressant

L'élève doit encercler une des croix qui, implicitement, se réfèrent à une échelle universelle ; ainsi, le choix de la première (resp. deuxième ; troisième ; quatrième) croix à partir de la

[5] Dans la pratique, le nombre d'items pourrait évidemment être plus grand que le nombre de croix.

gauche de la première ligne signifierait que le sujet juge très inutile (resp. un peu inutile ; un peu utile ; très utile) l'exercice en question.

Cette technique du différentiateur sémantique possède le grand avantage d'être très facile à utiliser par les étudiants. Elle s'avère toutefois délicate à concevoir et à exploiter par l'enseignant, car, d'une part, celui-ci doit veiller à ce que les deux énoncés (adjectifs ou phrases) proposés pour chaque item se réfèrent bien au même phénomène, soient tout à fait contraires l'un de l'autre et forment un tout cohérent [6], et, d'autre part, les conclusions concrètes qui découlent des renseignements fournis par les élèves sont difficiles à tirer de façon incontestable.

Portons à présent notre attention sur un second type d'échelle couramment rencontrée dans la pratique et qui porte le nom de son concepteur, le psychologue américain Rensis Likert (1903 – 1981). Au départ, une *échelle de Likert* avait été conçue comme étant un outil susceptible de mesurer des attitudes ; elle était composée d'une échelle universelle d'accord, avec cinq échelons, et proposait au sujet d'indiquer sur cette échelle son plus ou moins grand degré d'accord (ou de désaccord) pour une série d'affirmations relatives à un même sujet.

Pour illustrer ces propos, donnons un exemple pédagogique. Considérons le cas d'un professeur s'interrogeant sur la perception par ses élèves de la valeur métacognitive d'une narration de recherche accomplie en groupes. Il pourrait notamment inviter ces derniers à indiquer leur degré d'approbation sur ces affirmations en choisissant un des symboles prévus à cet effet :

- Le fait de rendre compte de toute ma démarche favorise ma compréhension :

  -- - = + ++
- Devoir raconter en détail toute ma démarche aide à mieux structurer mes idées :

  -- - = + ++
- Cette façon de travailler me permet de toujours savoir où j'en suis dans mes raisonnements :

  -- - = + ++
- Le travail avec des condisciples m'a permis de mieux me situer par rapport à la matière :

  -- - = + ++

Par extension, le nom d'échelle de Likert est souvent attribué à toute échelle comprenant un nombre quelconque, généralement compris entre 2 et 7, d'échelons devant traduire l'avis d'un

[6] Il est possible de mesurer la fiabilité d'un ensemble d'items censés se rapporter à un même phénomène à l'aide de paramètres statistiques : l'indice *alpha de Cronbach* est souvent utilisé à cet effet.

sujet sur des propositions de l'enquêteur. Il est souvent recommandé de proposer au moins 3 items sur un même phénomène à étudier ; ceux-ci doivent évidemment former un ensemble cohérent. Il importe de plus qu'une telle grille fournisse des résultats similaires quand elle est utilisée dans les mêmes conditions par plusieurs juges [7] ; elle doit également être construite avec soin de manière à former une grille d'évaluation conforme à la situation d'enseignement et les compétences qui doivent être étudiées. A cet effet, il paraît indispensable d'envisager les points suivants [8] :

- Sélectionner une (ou des) compétence(s)
- Construire une famille d'activités susceptibles de pouvoir être exploitées comme situations d'évaluation
- Déterminer le type de grille
- Choisir le type d'échelle d'appréciation
- Déterminer le nombre d'échelons
- Formuler des éléments observables en relation avec les critères d'évaluation retenus
- Construire une échelle d'appréciation en précisant le niveau d'exigence, notamment en repérant des critères prioritaires
- Procéder à la mise en forme de la grille
- Mettre la grille à l'essai (pré-test).

Un tel programme, déjà très difficile à bien mettre en œuvre, ne constitue qu'une première étape dans l'évaluation, et probablement la moins difficile. De fait, une fois la grille remplie par les élèves, l'évaluateur (qui est en général le professeur) devra l'exploiter adéquatement aussi bien sur le plan formatif que sur le plan certificatif : et cela est extrêmement délicat !

4. Echelle descriptive

Les points faibles des échelles (ou grilles) universelles peuvent être, en partie, éliminés au moyen d'*échelles descriptives*. Comme leur appellation le suggère, il s'agit dans ce cas non plus d'émettre un jugement général, par exemple au moyen d'une échelle d'excellence, mais de décrire de façon détaillée, et pour chacun des critères envisagés, différents niveaux de performances pouvant être réalisées par les apprenants. En guise d'illustration concrète, retenons trois critères se rapportant à la résolution d'un problème mathématique posé en langage courant et pouvant être résolu soit par des essais numériques, soit au moyen de

[7] La fidélité d'une échelle peut être mesurée statistiquement, par exemple à l'aide d'un indice *kappa de Cohen*.
[8] D'après le site http://www.Csdps.qc.ca.

systèmes d'équations[9]. Ces trois indicateurs, à savoir la modélisation du problème, l'argumentation et la résolution, pourraient être évalués au moyen d'une même échelle d'excellence comprenant, par exemple, les trois niveaux « médiocre » , « acceptable » et « excellent » ; par contre, des échelles descriptives différeraient pour les quatre cas et pourraient notamment se présenter comme cette grille testée sur le terrain [10] :

Critères	Niveau « excellent »	Niveau « acceptable »	Niveau « médiocre »
Modélisation	L'élève introduit des variables et les utilise à bon escient	L'élève introduit des variables mais ne les exploite pas adéquatement	L'élève utilise uniquement des descriptions textuelles des grandeurs et n'introduit pas variables
Argumentation	L'élève explique et justifie toutes les étapes de sa démarche d'une manière claire et compréhensible	L'élève explique et justifie maladroitement les grandes lignes de sa démarche	L'élève enchaîne des calculs sans explication ni justification
Vérification	L'élève valide ses réponses tout en précisant et justifiant sa démarche	L'élève affirme, avec justification, que certains résultats sont vrais ou faux, mais ne le fait pas pour tous les résultats	L'élève ne met pas en doute ses résultats

Les échelles descriptives sont évidemment plus difficiles à construire que leurs homologues universelles, car elles réclament une bonne anticipation des réponses possibles par les élèves. Elles sont également d'un usage plus restreint puisqu'elles sont spécifiques de l'indicateur envisagé. Elles possèdent toutefois des avantages importants d'un point de vue pédagogique, notamment :

- Grâce à la référence explicite à des productions concrètes, chaque élève reçoit un feed-back concernant l'activité décrite et, en conséquence, peut utilement se servir de cette appréciation pour s'auto-évaluer, puis réguler ses apprentissages. De telles échelles descriptives conviennent donc particulièrement bien pour de l'évaluation formative.
- L'aspect descriptif de l'échelle permet de porter un jugement sur les élèves sans devoir les comparer entre eux : le jugement porté a ainsi un « caractère critérié, et non normatif » [11].
- Les tâches étant bien décrites, plusieurs évaluateurs utilisant la même échelle devraient avoir des jugements assez concordants.

[9] Par exemple, le problème baptisé le «Gros Dédé sur une balance » que l'on peut trouver sur le site : http://www.irem.univ-montp2.fr/groupeZEP/stage_zep/GROS_DEDE_SUR_UNE_BALANCE.doc. *: « En utilisant les informations données par trois dessins, déterminer combien pèsent le gros Dédé, le petit Francis et le chien Boudin ».*

[10] Cette grille a été construite par Schintgen, un étudiant de la FOPED à l'Université du Luxembourg, lors de son stage professionnel réalisé pendant l'année académique 2006-2007.

[11] D'après Scallon G. (2004), *L'évaluation des apprentissages dans une approche par compétences*, Editions De Boeck, Bruxelles.

5. Echelle descriptive globale

Grâce à la manière dont elle est construite, une échelle descriptive permet de repérer les points forts et les faiblesses d'un apprenant pour une tâche bien déterminée. Afin d'évaluer les étudiants pour des activités complexes, il est évidemment possible de faire appel à des grilles d'évaluation composées de plusieurs échelles descriptives, mais les résultats ainsi obtenus ne sont pas toujours faciles à exploiter, notamment parce qu'ils peuvent être discordants.
Une appréciation de performances pour des tâches complexes à l'aide d'une seule appréciation, qui donnera néanmoins à l'apprenant un feed-back suffisamment précis pour lui permettre de réguler son apprentissage, peut être réalisée par une *échelle descriptive globale*, encore appelée *holistic rubric* dans la littérature spécialisée américaine. Il s'agit de construire une seule échelle comprenant généralement 4 ou 5 échelons qui décrivent explicitement des qualités attendues. Un exemple est fourni par la taxonomie SOLO [12] qui propose cinq niveaux de sophistication qui peuvent être rencontrés dans les réponses d'un apprenant face à une tâche de nature académique :

- Niveau A : *préstructurel*. La tâche n'est pas approchée convenablement ; le problème n'est pas compris ; les éléments de compréhension et d'analyse sont utilisés à contresens.
- Niveau B : *unistructurel*. Un ou quelques aspects de la tâche sont épinglés correctement et utilisés, mai sils ne contribuent pas à son développement ou à sa résolution.
- Niveau C : *multistructurel*. Plusieurs aspects de la tâche sont considérés et ils couvrent correctement les diverses composantes de la tâche ; cependant, ils sont traités séparément et la tâche ne peut encore être exécutée (apprentissage de faits, de connaissances).
- Niveau D : *relationnel*. Les diverses composantes sont intégrées, chaque partie révélant bien sa contribution à la compréhension ou à l'exécution de l'ensemble (apprentissage des liens et des relations).
- Niveau E : *abstrait étendu*. L'ensemble obtenu lors du niveau précédent est mobilisé ou conceptualisé à un haut niveau d'abstraction, ce qui rend les connaissances et compétences acquises utilisables dans d'autres circonstances [13] ; de plus, le processus

[12] « Structure of the Observed Learning Outcomes », faisant l'objet de nombreuses citations sur internet.
[13] Il s'agit du principe du « transfert ».

suivi pour en arriver là est reconsidéré, ce qui rend la démarche plus efficace et plus disponible pour d'autres opérations [14].

Concrètement, une semblable échelle globale peut être obtenue en « assemblant » adéquatement les différentes échelles d'une même grille d'évaluation, de manière à former une nouvelle échelle descriptive. Nous allons illustrer ces propos en réalisant l'exercice de construire une échelle globale avec 5 échelons en partant des trois échelles explicitées dans la section précédente et relatives à la compétence de résolution de problèmes mathématiques : l'échelle globale pourrait se présenter comme suit :

- Niveau A. L'élève utilise uniquement des descriptions textuelles des grandeurs et n'introduit aucune variable ; il enchaîne des calculs sans explication ni justification ; il ne met pas en doute ses résultats.
- Niveau B. L'élève introduit certaines variables, mais pas toutes ; il justifie certains éléments de sa démarche, mais pas tous ; il affirme sans justification que certains résultats sont vrais ou faux.
- Niveau C. L'élève introduit des variables mais ne les exploite pas adéquatement ; il explique et justifie maladroitement les grandes lignes de sa démarche ; il affirme, avec justification, que certains résultats sont vrais ou faux, mais ne le fait pas pour tous les résultats.
- Niveau D. L'élève introduit des variables et les utilise de manière efficace [15] ; il présente une solution complète, mais comportant parfois des erreurs relatives aux règles et conventions du langage mathématique.
- Niveau E. L'élève introduit des variables et les utilise de manière efficiente [16] ; il justifie systématiquement toutes les démarches de manière claire et compréhensible ; il valide ses réponses tout en précisant et justifiant sa démarche.

Une telle échelle descriptive globale possède l'avantage de rassembler divers indicateurs pour ne fournir qu'une seule appréciation facile à exploiter. Elle n'est toutefois guère facile à construire de façon cohérente ; pour cela, il convient de bien choisir les échelons des échelles descriptives de départ, ce qui se fait généralement en supposant que ces dernières sont bien corrélées, cette hypothèse n'étant pas souvent vérifiée dans la pratique. Pour mettre en évidence ces propos, reprenons l'exemple présenté ci-dessus : au départ de 3 échelles simples à 3 niveaux, nous avons construit une échelle unique à 5 niveaux ; en réalité, nous avons

[14] Ce point se rapporte à la « métacognition ».
[15] Une stratégie est qualifiée d'*efficace* si elle mène à un résultat adéquat.
[14] Une stratégie est dite *efficiente* si elle est efficace tout mais aussi économique en ce sens qu'elle privilégie une démarche concise pour atteindre le résultat.

choisi 5 des $3^3 = 27$ combinaisons possibles des divers échelons originaux ; il va sans dire que certaines de ces 27 combinaisons, pourtant souvent rencontrées [17], ne se retrouveront pas dans les 5 retenues et il faudra donc que l'évaluateur estime de façon subjective quel est le niveau de l'échelle globale le plus proche de la situation réelle.

Malgré de tels inconvénients, une échelle descriptive globale s'avère néanmoins particulièrement adaptée pour une évaluation normative, puisqu'elle livre une appréciation unique pour caractériser la prestation d'un élève : par exemple, le professeur pourra décider qu'un élève dont le niveau de compétence est A ou B se situe en dessous du minimum requis lors de la résolution de situations-problèmes comparables à celles traitées. Elle convient encore fort bien pour une évaluation formative, puisque l'apprenant peut dégager de cette échelle globale ses qualités et ses défauts et dès lors réguler son apprentissage concernant la résolution de problèmes.

6. Conclusion

La connaissance des types d'échelles nous semble être un pré-requis, nécessaire mais loin d'être suffisant, pour évaluer des compétences en mathématiques. En effet, l'enseignant ne pourra réellement commencer son travail sur le terrain que lorsqu'il connaîtra les outils mis à sa disposition.

Concrètement, pour évaluer une compétence, il s'agit de proposer aux apprenants la résolution de tâches-problèmes nécessitant la mise en œuvre de la compétence sélectionnée ; de telles épreuves doivent être critériées, ce qui suppose des critères pertinents, indépendants et peu nombreux évalués à travers des indicateurs [18]. Un *critère* est un point de vue auquel on se place pour évaluer, c'est une qualité attendue d'un objet. Les critères doivent être indépendants, de manière à ne pas évaluer plusieurs fois le même phénomène. On peut distinguer [19] diverses qualités que devraient avoir les critères :

- La *pertinence* ou adéquation de la production à la situation
- La *correction* ou utilisation correcte des concepts et outils
- La *cohérence* ou utilisation d'une démarche logique
- La *complétude* ou caractère complet de la réponse (exemple : pneu crevé)

[17] C'est le cas par exemple quand un même élève reçoit des notes contrastées selon les critères (et cela se produit souvent !).

[18] Voir le livre *Didactique de la géographie : organiser les apprentissages*, par Merenne B. (2005), Editions De Boeck, Bruxelles.

[19] Voir le livre *Une pédagogie de l'intégration : compétences et intégration des acquis*, par Rogiers X. (2004), Editions De Boeck Université, Bruxelles.

Il importe encore de construire des familles de tâches adaptées à chaque situation d'enseignement [20]. D'après des recommandations du BIEF [21] , les épreuves d'évaluation devraient être élaborées en veillant à ce que les situations respectent les paramètres de la famille et puissent être considérées comme équivalentes tout en suivant la règle des 2 / 3 : les apprenants doivent se montrer compétents lors de deux situations sur trois qui leur sont proposées. Ces trois situations équivalentes pourraient être réalisées comme suit : la première en groupe pour générer éventuellement des conflits socio-cognitifs et favoriser une certaine régulation de l'apprentissage, la deuxième individuellement avec accompagnement de l'enseignant puis une correction collective, et la dernière individuellement en tant qu'évaluation formative débouchant si nécessaire sur une remédiation. Au surplus, les évaluations devraient être d'un niveau plus faible que les situations travaillées en classe, car il ne semble pas raisonnable de demander à un élève de résoudre, en un temps limité, un problème complexe et inédit [22].

Quant à l'exploitation des échelles d'évaluation, à leur traitement scientifique, à leur interprétation et aux conclusions que peuvent en tirer les enseignants et les apprenants, c'est une toute autre histoire qui mériterait assurément des études ultérieures !

Deuxième partie : Leçon sur l'évaluation [23]

- **Evaluation : pourquoi ?**

2 visées

- ***formative*** : au service de la régulation de l'apprentissage
- ***sanctionnante*** : au service de la régulation de la société (certification ou sélection)

- Evaluer quoi ?

[20] C'est ce que fait notamment la Commission des Outils d'Evaluation mise sur pied par le Ministère belge de l'éducation : voir, par exemple, le site internet à l'adresse : http://www.enseignement.be.

[21] D'après l'article L'évaluation des compétences à travers des situations complexes, par Gérard F.M. (2005), *Actes du Colloque de l'Admee-Europe*, IUFM Champagne-Ardenne ; cette note peut être trouvée sur le site du BIEF (Bureau d'Ingénierie pour l'Education et la Formation), à l'adresse : http://www.bief.be/index.php?s=3&rs=17&uid=76

[22] A ce sujet, voir les deux livres rédigés par Antibi A. aux Editions Math'Adore : *La constante macabre ou comment a-t-on découragé des générations d'élèves ? (*2003) et *Les notes : la fin du cauchemar ou en finir avec la constante macabre* (2007) ; voir aussi la deuxième partie ci-après.

[23] Il s'agit d'un fichier (réalisé en Power Point) accompagnant un exposé oral sur l'évaluation (FOPED, Université du Luxembourg).

Les objets d'observation (de mesure) peuvent relever du niveau de complexité :

de performances isolées (habiletés) → de performances intégrées et complexes (compétences), dans les domaines

- cognitif (des savoirs)
- sensori-moteur (des savoir-faire, skills)
- affectif (des attitudes, savoir-être)
- métacognitif (des jugements, régulations sur soi)

- Une évaluation : c'est quoi ?

« Evaluer désigne une conduite supposant l'adoption d'une norme, pour laquelle une personne (l'enseignant, le formateur, un observateur) donne une information synthétique, parfois une mesure, à l'aide ou non d'une technologie (test, examen, grille, interrogation, enquête …) sur la valeur d'une personne en formation (un élève, un stagiaire), sur son comportement, sur un enseignement ou tout autre élément d'un système éducatif » (Viallet et Moisonneuve)

« Evaluer, c'est porter un jugement de valeur argumenté dans le but de prendre une décision » (Courtillot et Ruffenach)

- Quelques premières réflexions (Merenne; Courtillot-Ruffenach)
- Part de subjectivité
- Liée aux finalités de l'enseignement
- Pour évaluer une compétence, il s'agit de proposer aux élèves la résolution d'une tâche-problème nécessitant la mise en œuvre de la compétence sélectionnée
- Les épreuves doivent être critériées, ce qui suppose des critères pertinents, indépendants et peu nombreux évalués à travers des indicateurs
- Construire les évaluations en amont permet de fixer les objectifs d'enseignement
- A propos des critères (Gérard)
- Un critère est un point de vue auquel on se place pour évaluer, c'est une qualité attendue d'un objet ; il est construit à parti d'un cadre de référence (programme, décret, …)
- Les critères doivent être indépendants (pas 2 fois la même chose)
- 2 (qualités de) critères incontournables :

- Pertinence ou adéquation de la production à la situation
- Correction ou utilisation correcte des concepts et outils

- A propos des critères (suite)
- Deux critères importants :

- Cohérence ou utilisation d'une démarche logique

- Complétude ou caractère complet de la réponse (exemple : pneu crevé)

• Critères de perfectionnement : variables en fonction des objectifs poursuivis

• Les trois critères pour construire une évaluation (Mager)

1. L'action attendue de l'élève : on demande (on évalue ce qu'un élève peut faire après apprentissage et qu'il n'était pas capable de faire avant)
2. Les conditions de réalisation : on donne
3. La performance minimale à atteindre : on exige

• Indicateur

• Un indicateur est un élément recueilli dans la production, qui permet de se prononcer sur la façon dont les attentes (traduites en critères) sont satisfaites.

• Si les critères sont en nombre restreint et relativement transversaux, les indicateurs, eux, sont différenciés selon les champs d'application et peuvent être plus nombreux.

• Classifications des évaluations

• Formative / certificative

• Quantitative / qualitative

• Par objectifs / par compétences

• Continue / occasionnelle (sommative)

• Questions ouvertes / questions fermées

• Théorie / exercices / problèmes

• Interne / externe

• …

• La trilogie de l'évaluation

• L'évaluation <u>diagnostique</u> : elle se situe *avant* l'enseignement

• L'évaluation <u>formative/trice</u> : elle se situe *pendant* l'enseignement

• L'évaluation <u>sommative</u> : elle se situe *après* l'enseignement

• Quel est votre style d'évaluation ? (P. Musch)

• Quel est votre style d'évaluation ? (suite)

• Taxonomie de Gagné (domaine cognitif)

1. Etre capable de citer des faits
2. Etre capable d'utiliser des concepts
3. Etre capable d'appliquer des principes et/ou des lois
4. Etre capable d'appliquer des procédures
5. Etre capable d'inventer des procédures cognitives
6. Etre capable de mettre en œuvre des gestes

7. Etre capable d'avoir une attitude adaptée

- 10 caractéristiques d'une évaluation sommative (Courtillot-Ruffenach)

1. Elle se situe à la fin d'une séquence d'enseignement
2. Elle est notée
3. Elle est valorisante pour l'élève (conçue pour sa réussite)
4. Les questions sont hiérarchisées par ordre et difficulté croissante
5. Le barème est établi sur des critères simples et communiqué à l'élève

- Evaluation sommative (suite)

6. L'évaluation concerne toutes les compétences mises en jeu lors de l'enseignement
7. Il convient de n'évaluer qu'une compétence à la fois et qu'une fois chaque compétence
8. L'évaluation porte à la fois sur des acquis cognitifs et méthodologiques
9. Elle sollicite des activités intellectuelles variées (en situation contextualisante et/ou non)
10. Elle est multiforme (réponse à rédiger, dessin à faire, tableau à compléter, …)

- Un phénomène didactique :
  « la constante macabre »

« Si à un devoir ou à un examen, un certain pourcentage d'élèves n'est pas en situation d'échec, l'évaluation est considérée comme non crédible, anormale. Cette proportion constante d'élèves qui doivent ainsi, quoi que l'on fasse, se retrouver en situation d'échec sera qualifiée de *constante macabre* » (A. Antibi)

« Si vous souhaitez motiver vos élèves : faites-les réussir » (A. Rieunier)

- Comment les sujets de contrôles génèrent la constante macabre ?
- Difficulté des questions posées
- Des sujets trop bien « équilibrés »
- Des sujets trop longs
- Barème
- Rigueur dans la rédaction
- A la recherche d'un beau sujet
- Désir de « balayer » le programme du contrôle
- Comment ? (suite)
- La question cadeau
- La question réservée à l'élève Musclor
- Une drôle de générosité
- La constante macabre vraiment involontaire

- Décalage entre l'importance accordée à la note et l'imprécision dans la manière d'élaborer les sujets
- Suggestions pour combattre la constante macabre

1) (Antibi, 1988) Pour chaque contrôle de connaissances, la moitié de l'épreuve environ serait constituée par des exercices tout à fait analogues à ceux du contrôle précédent, l'autre moitié portant sur la partie de programme traitée entre les deux contrôles. Les élèves seraient bien sûr prévenus.

- Résultats empiriques
- Amélioration des notes
- Mise en confiance des élèves
- Incitation au travail (car récompense)
- Amélioration des relations professeurs-élèves
- Certains élèves découragés ont eu une sorte de déclic
- Intérêt pédagogique : revenir plusieurs fois sur la même matière
- Cas de la Finlande (Pisa) : pas de redoublement
- Suggestions pour combattre la constante macabre (suite)

2) EPCC : Evaluation Par Contrat de Confiance (Antibi, 2007)

- annonce du programme du contrôle (liste de questions traitées et corrigées en classe)
- Séance de question-réponses pré-contrôle
- Contenu (4 points sur 20 : question inédite) et correction du sujet

N.B. *Il convient bien sûr d'informer les élèves dès le début de l'année scolaire du type d'évaluation auquel ils seront soumis*

- Résultats empiriques
- La CM semble supprimée
- Un vrai climat de confiance voit le jour
- Les moyennes de classe augmentent
- Les élèves travaillent beaucoup plus
- Travail meilleur (contrat didactique plus clair)
- Plus de répartition gaussienne des notes (effet Posthumus), mais une loi bimodale, voire trimodale
- Des avantages aussi pour l'apprentissage
- Expérimentations positives en France, Belgique, Congo, …
- Quelques questions
- Liaison avec le reste de l'établissement ?

- Sentiment de culpabilité ?
- Favorise un apprentissage par cœur ?
- Pourquoi des exercices identiques ? (petite variation = obstacle potentiel)
- Et les tâches complexes ?
- L'EPCC évalue-t-elle des véritables compétences mobilisables ?
- Deuxième partie
- Essai d'une étude sur l'évaluation de compétences
- Quelques références
- SCALLON (2004) *L'évaluation des apprentissages dans une approche par compétences* (édition De Boeck Université)
- REY B., CARETTE V., DEFRANCE A., KAHN S. (2003) *Les compétences à l'école: apprentissage et évaluation* (édition De Boeck)
- BECKERS J. (2002) *Développer et évaluer des compétences à l'école : vers plus d'efficacité et d'équité* (édition Labor)
- PAQUAY L., CARLIER G., COLLES L., HUYNEN A.M., *L'évaluation des compétences chez l'apprenant : pratiques, méthodes, fondements* (2002), (Presses Universitaires de Louvain)
- …
- Avertissements
- *Comment identifier et décrire les ressources mentales mises en œuvre pour résoudre non pas un problème mais une classe de problèmes (…) Il n'y a pas de réponse qui n'exige des centaines de pages, toutes remplies d'hypothèses non prouvées, de modèles provisoires, de disputes théoriques* (de Montmollin, 1984)
- Que des questions ou suggestions; aucune certitude !
- Qu'est-ce qu'une compétence ?

C'est un système intégré de savoirs (savoirs, savoir-faire, savoir-être, …)

- qui permet d'aboutir à une performance
- mobilisable dans plusieurs contextes

=> *Etre compétent, c'est être performant dans une multitude de contextes*

- Difficultés
- Importance de la « famille de situations », mais *cette notion est problématique : elle n'est ni opérationnalisée, ni conceptualisée* (Crahay, 2003)
- Tautologie : *une compétence s'applique à une famille de situations et une famille de situations est caractérisée par la mobilisation d'une même compétence* (Chenu, 2004)

- En évaluation, c'est la famille de situations qui définit la compétence, mais le fait de considérer des situations comme identiques ou différentes est subjectif (par exemple : transplantation cardiaque, résolution de problèmes, …)
- Difficultés (suite)
- Inférer la maîtrise d'une compétence à partir du traitement de quelques situations est délicat
- Se contenter de constater que les élèves ne savent pas réaliser ce qu'ils n'ont pas appris !
- Acceptabilité sociale difficile (profs, élèves, parents)
- Niveau taxonomique le plus élevé
- …
- Evaluation de performances (savoirs et savoir faire), mais aussi « mesure » de savoir-être, d'attitudes, d'habiletés/aptitudes, …
- Concept de « mesure » d'attitudes
- Mesurer un processus consiste, pour nous, à établir une relation entre le processus et un symbole choisi au sein d'un ensemble fini de possibilités
- Echelle variable selon les processus et des indicateurs (nominale – ordinale – d'intervalle – de rapport ; universelle *vs* spécifique)
- Echelons variables en nombre, en nature (descriptifs, numériques, littéraux, iconiques, …)
- Echelle numérique
- Grande variété dans les notes possibles en contexte scolaire
- Homogénéité des échelons – caractère linéaire de l'échelle
- Que représentent les nombres observés ?
- Comment les traiter (statistiquement) ?
- Un 1er exemple : échelle d'accord
- Nombre d'échelons : en général de 2 à 7
- Parité ou imparité du nombre d'échelons
- Pouvoir discriminant des échelons
- Autres exemples d'échelles universelles
- Avantages / inconvénients
- S'appliquent à des situations variées
- Faciles à construire
- Faciles à interpréter

- Les jugements peuvent être subjectifs et non concordants (indices kappa de Cohen et de corrélation de Kendall ; théorie de la généralisabilité)
- Feedback peu instructif
- Grille d'évaluation
- Evaluation d'un même phénomène au moyen de plusieurs échelles
- Deux cas fréquents en sciences humaines
- *Différentiel sémantique* d'Osgood (1957) : lent / rapide, facile / difficile, … (indice alpha de Cronbach)
- *Echelle de Likert* (1903-1981) : échelle d'attitudes = > toutes échelles numérique (de 2 à 7 échelons ; indices alpha et kappa)
- Construction d'une grille (Csdps.qc.ca)
- Sélectionner une (ou des) compétence(s)
- Construire une famille adéquate d'activités
- Déterminer le type de grille
- Choisir le type d'échelle
- Déterminer le nombre d'échelons
- Formuler des éléments observables adéquats
- Construire une échelle d'appréciation en précisant le niveau d'exigence
- Procéder à la mise en forme de la grille
- Mettre la grille à l'essai (pré-test)
- Soumettre la grille, puis exploiter les résultats
- Echelle descriptive : un exemple concret
- Suite de l'exemple (par Schintgen)
- Avantages / inconvénients
- Usage plus restreint
- Plus difficiles à construire que les échelles universelles (anticiper les possibilités des élèves)
- Feedback pus précis (=> auto-évaluation facilitée, => pour une évaluation formative)
- Un jugement peut être porté sur les élèves, sans devoir les comparer (caractère critérié, et non normatif)
- Jugements assez concordants si plusieurs juges
- Echelles descriptives globales :
  deux exemples
- 1 : taxonomie SOLO

- Niveau A : préstructurel
- Niveau B : unistructurel
- Niveau C : multistructurel
- Niveau D : relationnel
- Niveau E : abstrait étendu

• 2 : globalisation des échelles « modélisation – argumentation – vérification » pour la résolution d'un problème (par Schintgen)

NB. 5 combinaisons parmi les 27 possibles

• Avantages / inconvénients
• Difficile à construire de façon cohérente
• Dans la pratique, on est souvent confronté à des niveaux non prévus par l'échelle
• Convient bien pour une évaluation normative
• Convient bien pour une évaluation formative (régulation)
• Quantification des échelles ordinales
• Linéarité ?
• Symétrie ?
• Valeur attribuée à un échelon « neutre » ?
• Méthode MACBETH (par VANSNICK, …)
• Détermination des poids par la méthode des cartes (par SYMOS)
• Statistique multivariée : méthodes descriptives ou explicatives; types de variables (cfr, par exemple, *Market : études et recherches en marketing* par EVRARD – PRAS – ROUX)
• Agrégation de plusieurs échelles
• Problème multicritère de type « tri »
• Un exemple ordinal simple : 3 critères avec 3 niveaux => 10 possibilités en 5 niveaux
• Problème mathématique relatif à des valeurs numériques
• Exemple 1 : épreuve du Bac national au GDL pour les V200
• Exemple 2 : Commission des Outils d'évaluation du Ministère Belge
• Une situation : p. ex. « démontrer en géométrie »
• 4 critères: les 2) et 3) sont « prioritaires »

1) traduction des données (4 points)
2) mise en œuvre d'une méthode de démonstration (6 points)
3) Argumentation (8 points)
4) présentation de la démonstration (2 points)

- Autres méthodes de R.O.

L'aide multicritère à la décision (par Ph. VINCKE, Editions de ULB et Ellipses)

- MAUT (Multi Attribute Utility Theory, par FISHBURN, …)
- ELECTRE (ELimination Et Choix Traduisant la REalité, par B. ROY, …))
- …
- **Pour conclure :**
  quelques conseils pratiques
- Respect de la règle des 2 / 3 (BIEF, 2005)
- Les évaluations devraient être d'un niveau plus faible que les situations travaillées en classe (BIEF, 2005)
- 3 situations équivalentes : 1) en groupe, 2) individuelle avec accompagnement de l'enseignant puis une correction collective, 3) individuelle (BIEF, 2005)
- Indication de la difficulté ou du niveau d'acquisition des compétences (voir exemples Courtillot-Ruffenach)