ABRÉGÉ-2

LES OUTILS DE JUGEMENT

(Gérard Scallon)

AVANT PROPOS

RAPPEL
L'abrégé-1 se rapporte à trois savoir-faire liés à la formation à l'évaluation dans une approche par compétences :

1.- TRADUIRE DES ÉNONCÉS DE COMPÉTENCE EN TÂCHES COMPLEXES.
2.- ANALYSER UNE TÂCHE COMPLEXE EN RESSOURCES MOBILISABLES.
3.- BALISER UNE PROGRESSION.

ABRÉGÉ-2
CET ABRÉBÉ SUR LES OUTILS DE JUGEMENT SE RAPPORTE À UN QUATRIÈME SAVOIR-FAIRE :

4.- UN DERNIER ASPECT DE L'ÉVALUATION DES COMPÉTENCES, ET NON LE MOINDRE, REJOINT LA CAPACITÉ DE JUGEMENT QUI. ELLE-MÊME. REPOSE SUR CELLE DE DÉVELOPPER ET D'UTILISER DES OUTILS DE JUGEMENT (GRILLES D'ÉVALUATION, LISTES DE VÉRIFICATION, ÉCHELLES DESCRIPTIVES GLOBALES) ET D'IDENTIFIER DES CRITÈRES D'ÉVALUATION.

Entre faits et opinions

Dans la vie de tous les jours, nous sommes invités à communiquer des résultats associés à quelque événement dont nous avons été témoins. Nous nous efforçons d'être objectifs autant que possible, mais il est parfois difficile de retenir une impression, une opinion, voire un jugement. Devons-nous parler d'un raz de marée en précisant, froidement, qu'il a causé 20 000 pertes de vie ? … ou en donnant notre point de vue en mentionnant qu'il s'agit d'une catastrophe humanitaire ? Est-il mieux d'informer les actionnaires d'une entreprise que les ventes ont chuté de 26 % ces derniers mois au lieu de les prévenir d'une faillite imminente ?

En matière de rendement scolaire les « façons de parler » sont tout aussi apparentes. Marie a passé un examen en science et technologie. Faut-il faire état de son résultat de 54 bonnes réponses sur 60 (ou 90 %), par exemple, ou affirmer qu'il s'agit d'une excellente performance ? Ou encore, qu'elle s'est placée au septième rang de sa classe pour cet examen ?

Ce n'est pas tout ! Au regard de ce qui se passe dans la vie de tous les jours nos opinions ou nos impressions peuvent être laconiques. Après-tout, nous ne sommes pas tellement concernés, surtout lorsque nous ne nous sentons pas experts.

En matière de rendement scolaire nous pouvons substituer un jugement plus nuancé, voire analytique, au « verdict » global. L'information sur la performance de Marie en science et technologie peut être accompagnée de la mention d'un ou de plusieurs points faibles. Et le rang qu'elle s'est mérité dans sa classe peut être mis en contexte en connaissant davantage la force du groupe d'élèves.

Le choix entre communiquer des faits ou livrer des opinions n'est pas anodin et fait partie des enjeux de l'évaluation des apprentissages. Enseignants et enseignantes, les premiers responsables de l'évaluation avec leurs groupes d'étudiants, doivent être capables d'observer et d'emprunter des procédés adéquats de collecte d'informations. Tests, examens, contrôles, épreuves de rendement, productions complexes, etc. sont de cet ordre. Mais, le processus d'évaluation n'est pas complété pour autant. Il leur faut communiquer les résultats de chaque démarche d'évaluation, auprès des étudiants et aussi auprès de ceux qui les soutiennent (p. ex. les parents).

État des lieux en matière de jugement

Dans le domaine du rendement scolaire la performance des individus peut être révélée à partir de deux sources : l'examen écrit composé de plusieurs questions et la production complexe.

L'examen écrit relève de ces procédés de « quantification » où la performance d'un individu s'exprime par un « score », c'est-à-dire un nombre de bonnes réponses dans les cas les plus simples. À un contrôle en biologie comportant 20 questions, Jean-Louis a répondu correctement à 12 d'entre elles. Son résultat ou score est 12, 12 points, 12 sur 20 ou 60%. Le jugement qui doit accompagner l'information à transmettre au sujet de cette performance peut être basé sur une interprétation « critériée » (sans égard à la performance d'autres étudiants) ou sur une interprétation « normative » (par exemple, le rang occupé dans un groupe avec cette performance de 12 sur 20).

Dans une approche par compétences, c'est autre chose ! Une compétence ne peut être inférée à partir d'un examen composé de plusieurs questions, que celles-ci soient à réponse brève ou à choix multiple. Une compétence ne peut être démontrée qu'en exigeant des individus une production élaborée qu'il leur faut structurer eux-mêmes. Le terme « production » est générique et peut se rapporter à des compositions écrites (récit, conte, dissertation) ou à d'autres formes de prestation (routine en gymnastique, interprétation d'une pièce musicale, exposé oral).

Dans une approche par compétences, le jugement pose alors des défis considérables. Il n'y a pas de somme de points sur laquelle se baser. La démonstration de chaque compétence est un phénomène complexe qu'il faut regarder au travers plusieurs « fenêtres » (dimensions ou critères). Dans une perspective d'évaluation formative, les personnes chargées de la formation doivent pouvoir signaler les points forts et les points faibles d'une performance et suivre la progression de chaque individu. Dans une perspective de certification (évaluation sommative) ces mêmes personnes ou d'autres personnes, responsables de l'évaluation, doivent « noter » ou « coter » c'est-à-dire exprimer des jugements de façon succincte. Notes ou cotes sont de cette mouture.

Outils de jugement contre liberté d'expression

Dans la vie de tous les jours, il nous arrive d'exprimer librement nos jugements. Par exemple, les réactions à un tremblement de terre peuvent être diversifiées à souhait : terrible ! de forte intensité ! du jamais vu! Il en est de même des façons de recommander un restaurant ou de vanter les mérites d'une nouvelle voiture.

Pour ce qui est des apprentissages, c'est autre chose. L'évaluation de productions complexes, par exemple, ne peut être laissée aux caprices sémantiques des personnes juges. Ni aux aspects très particuliers que chaque personne veut bien observer ou noter. Ce serait la subjectivité à son meilleur comme au temps de la méthode dite de l'appréciation générale des compositions écrites d'étudiants, sans critères connus.

L'idée de « standardiser » les jugements n'a pas d'origine précise. C'était pourtant la préoccupation des premières échelles d'attitude qui proposaient en quelque sorte aux personnes consultées un choix forcé dans une chaîne graduée d'expressions. Par exemple, au lieu de demander « Que pensez-vous de la peine de mort ? » et de laisser libre cours au répondant pour exprimer son opinion on lui demandera de choisir l'un des échelons suivants d'une **échelle d'appréciation** :

…en désaccord … plus ou moins d'accord …entièrement d'accord

Dans le cas de phénomènes plus complexes (composition écrite, gymnastique, etc.) les points de vue dont il faut tenir compte peuvent être suggérés ou imposés à la personne juge. C'est notamment le cas de la grille d'évaluation comportant plusieurs critères (points de vue, dimensions, aspects) chacun accompagné d'une échelle d'appréciation. Le fait de demander à la personne qui évalue de considérer chacun des critères de la grille est une autre forme de standardisation.

-

Deuxième état des lieux : pourquoi évaluer ?

Il faudrait rappeler ici les principales fonctions de l'évaluation. l'une formative, l'autre sommative ou certificaive.

Les fonctions formative et certificative

L'évaluation **formative** doit déboucher sur des correctifs ou des améliorations, que ce soit en reprenant l'enseignement de départ (enseignement correctif) ou au moyen de feed-back informatifs lorsque la situation d'évaluation le permet. Dans ce deuxième cas, l'approche peut être adaptée à chaque individu en particulier (p. ex. dans une démarche d'autocorrection).

L'évaluation **certificative** vise la reconnaissance des apprentissage réalisés ou encore l'attestation des compétences que les étudiants doivent démontrer au sortir d'un programme d'études. La décision à prendre est de l'ordre de la promotion, de l'octroi d'un diplôme ou d'un permis de pratique.

Pour ce qui est de l'évaluation de productions complexes devant servir à inférer des compétences, les deux fonctions de l'évaluation ont des retombées d'ordre méthodologique différentes.

L'approche analytique ou globale

D'une part, en évaluation formative, la démarche doit être **analytique** puisqu'il s'agit de souligner tant les points forts que les points faibles relevés dans la réalisation de tâches complexes par chaque étudiant ou étudiante. Le jugement peut alors être porté au regard de chaque dimension de la performance attendue sans déboucher nécessairement sur un résultat global.

D'autre part, en évaluation sommative ou certificative, le jugement doit être succinct, voire **global**, puisqu'il s'agit d'éclairer une seule décision à prendre au terme de la formation : faire réussir ou faire échouer, si on peut se permettre ce genre d'expression. Toutefois, la démarche d'évaluation peut se démarquer de cette décision en communiquant un résultat global pour chaque étudiant et pour chaque compétence, soit au moyen d'une note ou d'une cote (la distinction entre ces formes de communication de résultats d'évaluation est présentée en addenda à cet abrégé --- l'addenda-1). Il revient à d'autres personnes de se servir de ce résultat pour prendre les décisions qui s'imposent.

-

EN BREF

LES OUTILS DE JUGEMENT POUR APPRÉCIER DES PRODUCTIONS COMPLEXES OU POUR INFÉRER DES COMPÉTENCES PEUVENT ÊTRE ABORDÉS SOUS DEUX ANGLES COMPLÉMENTAIRES :

1.- CELUI DE LEUR FORME ET DE LEUR CONTENU (GRILLE D'ÉVALUATION, LISTE DE VÉRIFICATION, ÉCHELLE DESCRIPTIVE GLOBALE);

2.- CELUI DU RÉSULTAT COMMUNIQUÉ (PROFIL ANALYTIQUE OU RÉSULTAT GLOBAL --- NOTE OU COTE).

Les outils de jugement d'après leur forme et leur contenu

L'unité fonctionnelle qui est à la base de certains outils est **l'échelle d'appréciation**. Il s'agit essentiellement d'une suite de termes ou d'expressions de « qualité » formant une progression. Par exemple, l'échelle d'appréciation universelle suivante avec cinq échelons :

[ ]médiocre [ ]acceptable [ ]bon [ ]très bon [ ]excellent

Cette échelle est dite universelle parce qu'elle est applicable à la très grande majorité des critères d'évaluation dans une foule de domaines. Il existe plusieurs modèles de ce type, certains étant plus spécifiques à des caractéristiques précises comme le comportement avec d'autres individus ou l'exécution de tâches répétitives. Par exemple :

[ ]très impoli [ ]impoli [ ]plus ou moins poli [ ]poli [ ]très poli

[ ]très lent [ ]lent [ ]plus ou moins rapide [ ]rapide [ ]très rapide.

Les échelons de ce type d'échelle peuvent s'exprimer en lettres ou en chiffres :

E D C B A	ou	1 2 3 4 5

Il suffit alors d'associer les lettres ou les chiffres à une légende qui permet de retracer l'échelle d'appréciation d'origine.

-

La grille d'évaluation

C'est un outil de jugement qui se compose essentiellement de critères chacun accompagné d'une échelle d'appréciation.

Voici une façon de se représenter la structure de ce type d'instrument :

LA GRILLE D'ÉVALUATION

Critère 1
[ ] échelon 1 [ ] échelon 2 [ ] échelon 3 [ ] échelon 4 [ ] échelon 5

Critère 2
[ ] échelon 1 [ ] échelon 2 [ ] échelon 3 [ ] échelon 4 [ ] échelon 5

Critère 3
[ ] échelon 1 [ ] échelon 2 [ ] échelon 3 [ ] échelon 4 [ ] échelon 5

Critère 4
[ ] échelon 1 [ ] échelon 2 [ ] échelon 3 [ ] échelon 4 [ ] échelon 5

etc.

Dans une grille d'évaluation, les échelles d'appréciation peuvent être uniformes ou universelles (p. ex. l'échelle d'excellence citée précédemment). Elles peuvent être plus spécifiques en exploitant un même champ lexical et en utilisant des adverbes d'intensité. Par exemple, pour apprécier le degré de politesse d'un individu :

[ ] très impoli [ ] impoli [ ]plus ou moins poli [ ]poli [ ]très poli.

Enfin, dans certaines grilles d'évaluation, les échelles d'appréciation peuvent être descriptives et, de ce fait, spécifiques à chacun des critères. L'exemple qui suit pourrait s'appliquer à l'évaluation du résumé d'un texte informatif :

Grille d'évaluation d'un résumé avec échelles descriptives			commentaire
Intégralité des idées de l'auteur			On remarquera que la <mention des idées>, l'<exactitude> et la <répétition> sont des indices qui rendent les échelles à la fois descriptives et spécifiques à chacun des critères.
[ ] aucune ou une seule idée de l'auteur	[ ] il manque une seule idée	[ ] toutes les idées de l'auteur sont mentionnées	
Précision du résumé			
[ ] plusieurs idées sont inexactes ou imprécises	[ ] une seule idée est inexacte ou imprécise	[ ] toutes les idées de l'auteur sont exactes	
Concision			
[ ] texte redondant (beaucoup de répétitions)	[ ] une ou deux répétitions	[ ] aucune répétition dans le texte	

-

On remarquera que la grille descriptive (ou grille d'évaluation descriptive) est beaucoup plus précise que les grilles traditionnelles construites avec des échelles uniformes ou universelles. Avec ce dernier type, chacun des trois critères d'évaluation d'un résumé aurait pu être accompagné d'une même échelle comme celle-ci :

[ ] médiocre [ ] acceptable [ ] excellent.

Il faut souligner que la grille descriptive, une fois complétée par une personne juge ou par l'étudiant ou l'étudiante (en auto évaluation) transmet beaucoup d'informations susceptibles d'amorcer des améliorations pour autant qu'elle n'est pas remplacée ou « masquée » par un résultat global. Son utilité en évaluation formative est indéniable.

La liste de vérification

Il s'agit d'un outil particulièrement utile dans certaines situations qui peuvent être décomposées en plusieurs sous-tâches ou en étapes bien identifiées. Rigoureusement, ce n'est pas un outil de jugement mais plutôt un instrument de consignation de faits divers dont on signale la présence ou l'absence. Ce qui n'empêche pas que les observations retenues se résument à une vue d'ensemble conduisant, de ce fait, à un jugement au sujet d'une production ou d'une démarche.

Idéalement, la liste de vérification devrait être composée d'éléments marqués avec le moins d'interprétation possible : « a signé sa lettre », « a ajusté son rétroviseur avant de démarrer », « a remisé son crayon », etc. sont des exemples de ce genre de faits divers susceptibles de constituer une liste de vérification. Malheureusement certaines listes de vérification comportent des aspects qui exige une interprétation : « a écouté attentivement l'exposé », « a réagi de façon appropriée », « a respecté les consignes », etc. sont des exemples d'éléments qui peuvent être cochés comme s'il s'agissait de faits « objectifs » mais qui n'en demandent pas loin une certaine interprétation.

Le simple marquage « tout ou rien » de ce genre d'élément à interprétation n'a rien d'objectif bien qu'il en ait toutes les apparences.

-

ENJEU D'ORDRE MÉTHODOLOGIQUE

UTILISATION GÉNÉRALISÉE OU UTILISATION SPÉCIFIQUE
À DES PRODUCTIONS PARTICULIÈRES ?

LA GRILLE D'ÉVALUATION AVEC ÉCHELLES UNIFORMES (NON DESCRIPTIVES) ET LA LISTE DE VÉRIFICATION COMPOSÉE D'ÉLÉMENTS À « INTERPRÉTATOPM » PEUVENT ÊTRE UTILISÉES DANS PLUSIEURS SITUATIONS-TÂCHES DE MÊME FAMILLE ET NE SONT DONC PAS SPÉCIFIQUES À DES PRODUCTIONS PARTICULIÈRES. C'EST UN AVANTAGE RECHERCHÉ !

EN REVANCHE, LE RECOURS À DES ÉCHELLES DESCRIPTIVES OU À DES ÉLÉMENTS FACTUELS * CONDUISENT À DES INSTRUMENTS DONT L'UTILISATION EST SPÉCIFIQUE ET LIMITÉE À DES PRODUCTIONS BIEN DÉFINIES. L'AVANTAGE RECHERCHÉ EST PLUTÔT DU CÔTÉ DE LA FIABILITÉ DES JUGEMENTS.

* Ce n'est pas toujours évident dans le cas de la liste de vérification.

Le modèle de production comme liste de vérification

La liste de vérification peut prendre une forme particulière lorsque la production demandée, tout en étant complexe et élaborée, doit contenir des éléments précis. La résolution de certains problèmes concrets de nature professionnelle (problème juridique ou médical, par exemple) entre dans cette catégorie. La réaction que les étudiants peuvent manifester à l'égard de certains événements (opinions justifiées) peut être jugée avec cette méthodologie. La démarche décrite ici est inspirée du ***performance assessment*** visant ce que les auteurs américains appellent les ***higher order skills***. Pouvons-nous nous en servir pour inférer des compétences ? La réponse est affirmative si la structure de la tâche complexe présentée aux étudiantes et aux étudiants est telle que ceux-ci doivent mobiliser (utiliser spontanément et en toute autonomie) leurs ressources ou des ressources externes. Le sujet traité ici est complexe. Un exemple, développé et utilisé par l'auteur de ces lignes, pour illustrer ce dont il est question est présenté à l'addenda-2. Il ne s'agit pas d'une compétence au sens strict, mais l'exemple montre que la méthodologie peut s'accorder à des réponses divergentes tout en étant acceptables.

-

L'échelle descriptive globale

Les échelles utilisées pendant plusieurs années ne se sont rapportées (en principe) qu'à une seule qualité ou dimension, c'est-à-dire à chacun des critères dans une grille d'évaluation. C'est la structure même de ce premier outil de jugement qui a été présenté dans cet abrégé.

L'échelle descriptive globale (inspirée des ***rubrics*** des écrits anglo-saxons) présente des échelons sous la forme de paragraphes descripteurs qui se rapportent à plusieurs critères (ou qualités) traités simultanément. Ce type d'échelle peut être facilement illustré avec la calligraphie en contrastant cet outil de jugement avec la grille d'évaluation.

Des individus ont été invités à copier, à la main, le texte suivant :

Pour écrire, il faut bien s'appliquer.

La grille d'évaluation, avec ses critères et une échelle uniforme pour chacun d'eux pourrait être la suivante :

Pente des lettres :
[ ]médiocre [ ]acceptable [ ]bon [ ]très bon [ ]excellent

Formation :
[ ]médiocre [ ]acceptable [ ]bon [ ]très bon [ ]excellent

Linéarité de l'ensemble :
[ ]médiocre [ ]acceptable [ ]bon [ ]très bon [ ]excellent

Espacement des mots
[ ]médiocre [ ]acceptable [ ]bon [ ]très bon [ ]excellent

Une grille d'évaluation avec une échelle descriptive spécifique à chaque critère, serait sans aucun doute une amélioration à apporter à cet outil de jugement. Cependant, s'il s'agit non pas de noter mais de « coter » (1, 2, 3, ou 4) un spécimen d'écriture, l'échelle descriptive suivante pourrait être utilisée :

1	2	3	4
L'écriture, dans son ensemble, laisse à désirer. On observe très peu de régularité : pente, hauteur, éloignement de la ligne de base. L'ensemble du texte est difficile à lire.	La pente des lettres est variable et plusieurs d'entre elles s'éloignent de la ligne de base. Leur hauteur n'est pas constante et certaines lettres sont mal formées au point d'être illisibles.	L'écriture n'est pas parfaitement régulière du point de vue de la pente et de la hauteur des lettres. Les lettres de certains mots sont séparées et quelques unes d'entre elles ont été formées à la hâte.	Les lettres sont penchées vers la droite de façon régulière. Les lettres sont bien formées et leur hauteur est constante. La base des lettres suit une ligne droite. Les mots sont séparés par un espace approprié.

-

Cette échelle descriptive globale présente des caractéristiques qu'il faut souligner :

1.- Les quatre échelons de cette échelle descriptive ont été rédigés en faisant varier simultanément les critères retenus (hypothèse d'une corrélation entre ces critères --- par exemple, l'hypothèse qu'une mauvaise pente est associée à une malformation des lettres).

2.- Les chiffres « 1, 2, 3 et 4 » ne sont pas des rangs (la meilleure performance reçoit ici la valeur 4 --- et non pas le rang 1).

3.- Il peut exister beaucoup de différences entre les spécimen d'écriture, différences qui ne peuvent être considérées avec seulement quatre échelons. Un tel outil de jugement doit donc être utilisé avec beaucoup de réserve.

Petite conclusion sur cette présentation des outils de jugement

Les outils de jugement qui viennent d'être présentés pourraient servir tels quels dans une perspective d'évaluation formative à cause de la précision du feed-back fourni aux étudiants et aux étudiantes. Tous les outils ne sont pas d'égale qualité sur ce plan. La grille d'évaluation descriptive, avec une échelle spécifique à chaque critère, se classe bonne première. Son caractère analytique (critères traités séparément) lui permet de rendre compte des forces et des faiblesses observées au sortir de chaque production. Sous réserve que la grille est bien construite, les échelons descripteurs indiquent avec précision les défauts à corriger ou les améliorations à apporter. Au regard d'une suite de productions de même nature, il y a là emprise pour le **suivi de la progression** de chaque étudiant ou de chaque étudiante.

La liste de vérification se rapproche « à sa manière » de la grille d'évaluation descriptive. On ne peut en dire autant de la grille d'évaluation avec échelles uniformes et encore moins de l'échelle descriptive globale, deux types d'outils de jugement qui n'ont pas été conçus pour fournir un feed-back dans une perspective d'évaluation formative.

Lorsqu'un résultat global doit être consigné ou communiqué

À certains moments de la formation des individus un jugement « syncrétique » plutôt qu'analytique s'impose. Après avoir consigné les principaux aspects d'une production complexe (grille d'évaluation ou liste de vérification) et également, de plusieurs productions, il faut reconnaître la maîtrise d'une ou de plusieurs compétences chez chaque étudiant ou étudiante. Le jugement doit alors être global et refléter le plus validement possible les différences entre les performances des individus. Dans l'esprit de plusieurs personnes, le verdict « succès-échec » (ou pass-fail des écrits américains) ne répond pas adéquatement à cette demande d'évaluation sommative ou certificative. D'où ce besoin de noter (ou de coter) pour rendre compte de certaines nuances.

-

Le cas le plus simple de tous est une addition de « points », le point étant une unité commode pour « accréditer » les diverses qualités d'une production ou d'une performance.

La grille d'évaluation (avec échelles uniformes ou échelles descriptives) se prête bien à cette arithmétique pour autant que les échelons de chacune des échelles soient accompagnés d'une valeur chiffrée comme dans l'exemple suivant :

médiocre [__]1 acceptable [__]2 bon [__]3 très bon [__]4 excellent [__]5

Au regard de la production ou de la performance d'un individu, chaque critère reçoit donc un nombre de points et c'est la somme de tous les points qui tient lieu de note chiffrée globale. Avec plusieurs productions à évaluer, c'est la même arithmétique qui s'applique pour établir la somme des notes globales comme le veut une longue tradition

La liste de vérification (faits divers ou composantes d'un modèle de réponse) se prête également à un dénombrement d'éléments observés (en allouant un point par élément marqué, par exemple).

Une pratique discutable

On comprendra que, dans cette forme de bilan pour un ensemble de productions, le jugement posé sur les qualités de chaque production disparaît pour faire place à une mécanique arithmétique. C'est cette pratique qui est remise en question dans une approche par compétences.

Ce qui fait problème dans certains cas c'est le modèle dit « compensatoire » où un aspect réussi compense pour un aspect échoué. Un exemple fort simple permettra de justifier cette critique. On a demandé à des élèves de désigner le destinataire d'une carte postale à envoyer. Voici trois spécimen de productions :

élève 1	élève 2	élève 3
M. Mme Jean Drolet rue des Métairies BelleEau (Qué.) G3C 1N9	Mme Julie 45 rue de Milot Santerre (Qué.) N6R 2T3	M. François Dupé 2567 rue Fortier Mortier-Ville (Qué.)

En prenant comme éléments à dénombrer dans une liste de vérification : 1) le nom complet du destinataire, 2) le numéro de rue, 3) la rue, 4) la ville et 5) le code postal, chaque élève se mérite quatre points (il manque toujours un seul élément --- numéro de rue, nom complet ou code postal). Ces productions sont-elles de même qualité ? **Oui**, si les éléments sont d'égale importance (nom complet ou code postal, par exemple). **Non**, si certaines omissions comme celle du nom complet sont plus graves que d'autres (code postal ou numéro de rue).

-

Tel est l'enjeu que pose la simple somme arithmétique d'éléments reliés aux aspects ou aux qualités diverses d'une production.

Une solution à envisager

C'est pour cette raison que les échelles descriptives globales ont été créées, pour ce qui est d'obtenir un résultat chiffré global sans passer par une somme d'éléments. Les échelons de ce genre d'échelle sont construits de façon telle que des éléments, plus importants que d'autres, vont être cités en premier pour que l'individu obtienne la meilleure cote. Dans l'exemple des cartes postales, un groupe de personnes pourrait établir l'ordre de priorité suivant : le nom complet est un élément incontournable suivi de près de la ville, du nom de rue et du code postal. C'est une question de validité quant aux informations absolument nécessaires pour que la carte postale se rende à destination.

Pour terminer cet exemple, le paragraphe descripteur de l'échelon le plus élevé (cote = 4, par exemple) pourrait mentionner tous les éléments d'une adresse complète. L'échelon qui suit dans l'ordre descendant des cotes (cote = 3) pourrait mentionner tous les éléments d'une adresse complète moins le code postal (si cette omission est jugée banale). Et ainsi de suite. Dans notre exemple, c'est l'élève 3 qui recevrait la cote 3, un résultat non confondu avec celui d'autres élèves.

Une autre solution à envisager

L'utilisation d'échelles descriptives globales n'est pas de tout repos lorsque le nombre d'éléments (aspects ou critères) dont il faut tenir compte dans la rédaction des échelons est élevé. Le recours à des valeurs « décimales » comme 2,5 ou 3,5 pour coter des performances qui ne cadrent pas parfaitement avec l'un ou l'autre échelon en témoigne.

Comme autre solution, il faut retourner à la somme des points mais en pondérant différemment cette fois certains critères et ce, pour corriger les effets indésirables du modèle compensatoire. Ainsi, nous pouvons doubler (voire tripler) le nombre de points alloués aux échelons de l'échelle accompagnant certains critères. Dans cet exemple, le critère B reçoit plus d'importance que le critère A :

Critère A
médiocre [__]1 acceptable [__]2 bon [__]3 très bon [__]4 excellent [__]5

Critère B
médiocre [__]2 acceptable [__]4 bon [__]6 très bon [__]8 excellent [__]10

-

Les effets de cette pratique sur la validité des jugements sont méconnus. En pratique, il nous faut s'exercer sur des exemples concrets pour apprécier les résultats obtenus et nous faire une meilleure idée. Poursuivons l'exemple de la carte postale adressée par trois élèves. Le tableau suivant présente une liste de vérification avec un poids différent à accorder à certains éléments ainsi que la note (ou score) obtenue par chaque élève.

Liste de vérification	élève 1	élève 2	élève 3
Nom complet __/ 4 Numéro de rue __/ 1 Nom de rue __| 1 Ville __/ 2 Code postal __/ 2	9	6	8

Le phénomène de compensation n'a pas joué en faveur de l'élève 2 puisque son omission du nom complet du destinataire lui a été coûteuse. Est-ce valide ? C'est toute la question qu'il faut soulever et le moyen d'y répondre avec discernement est d'avoir en mains un corpus de productions variant en qualité (ici des adresses de cartes postales).

La construction ou la rédaction d'outils de jugement.

Au regard de programmes par compétences, les instruments ou les outils « prêts à porter » n'existent pratiquement pas et ce, à tous les niveaux d'enseignement. La formation professionnelle n'échappe pas à cette réalité. C'est donc dire qu'enseignants et enseignantes, formateurs et formatrices sont contraints à développer eux-mêmes leurs outils d'évaluation. Il s'agit là s'un savoir-faire incontournable qui doit être développé.

Quel type d'outil privilégier ? Grille d'évaluation ? Liste de vérification ? Ou échelle descriptive globale ? Le recours à ce dernier type d'outil est une tendance non négligeable et il est trop tôt pour porter un jugement critique éclairé sur ce type d'outil.

L'enjeu est celui de rendre compte de la qualité de chaque performance observée dans un groupe d'individus en formation et ce, d'une façon telle que deux ou plusieurs personnes juges arrivent aux mêmes résultats, à peu de chose près.

Le savoir-faire dont il est question dans cet abrégé devrait se développer à même des situations avec lesquelles nous sommes tous familiers, même si ces situations ne se rapportent aucunement à la formation dispensée dans un programme d'études.

-

Des groupes de personnes pourraient alors « s'attaquer » à des objets comme : l'évaluation d'un site WEB, l'appréciation « chiffrée » d'un véhicule de promenade, la critique d'un restaurant, etc. Choix de critères ou d'indices, construction d'échelles d'appréciation ou modèle de réponse comme liste de vérification (site WEB idéal, spécimen de véhicule ou service attenduj dans un restaurant) pourraient être objets de discussion et d'échange. On ne doit pas chercher une bonne réponse dans ce genre d'exploration, mais le feed-back pouvant émerger de ces discussion constitue fort probablement une base valable d'apprentissage à l'évaluation.

Pour ce qui est de se pratiquer avec de véritables compétences quelques conseils utiles méritent d'être signalés.

Conseils pratiques pour l'élaboration
d'outils de jugement

1.- Avant d'entreprendre la construction d'un outil de jugement (grille, liste ou échelle) il fortement conseillé de structurer une tâche ou une famille de tâches d'évaluation (situations de compétence ou tâches complexes). À la question: « quel est votre instrument d'évaluation ? » il faut pouvoir montrer et un spécimen de tâche et l'outil de jugement. Ce sont deux éléments soudés et l'outil de jugement, à lui seul, ne constitue pas une « réponse acceptable » à la question posée.

2.- La structuration de la tâche nous amène à préciser ce que les étudiants doivent accomplir pour démontrer telle ou telle compétence. Il faut prévoir une consigne ainsi que les données qui seront accessibles tout en respectant la définition même qui a été donnée de la compétence : la capacité de mobiliser. Trop de sous-questions enlève cette effort de mobilisation pourtant essentiel à la notion de compétence.

3.- Dans la construction ou la rédaction d'un outil de jugement, la perfection n'est pas au rendez-vous du premier coup. Loin de là. Au mieux, il faut disposer d'un ensemble de productions concrètes, bonnes et mauvaises, productions qui vont servir de source d'inspiration pour les critères, les échelles d'appréciation ou le modèle de réponse. Cet ensemble de productions pourrait être obtenu lors des premiers essais de la démarche d'évaluation, par exemple lors d'un trimestre. Ce sera aussi l'occasion de mettre à l'essai l'outil de jugement pour y apporter des améliorations en vue d'une prochaine utilisation à un autre trimestre. Tel est l'esprit avec lequel il faut travailler.

-

ADDENDA-1
RÉSULTATS CHIFFRÉS :
SCORES, NOTES ET COTES

Le résultat obtenu à un examen s'appelle « **score** » dans les écrits francophones européens. Ce résultat correspond habituellement au nombre de bonnes réponses à un examen objectif. Dit autrement : c'est le « nombre exprimant le résultat d'un test » (Legendre, 2000, 1145). Nous sommes dans l'ordre des procédés de quantification c'est-à-dire qu'il y a « comptage » ou dénombrement d'éléments. Pour donner un sens au nombre obtenu ou à ce résultat, deux modes d'interprétation ont été identifiés dans le domaine de l'évaluation pédagogique : l'interprétation normative et l'interprétation critériée.

Tout n'est pas examen. Il existe des habiletés ou des savoir-faire qui ne se laissent pas observer par une succession de questions précises, que celles-ci soient à choix de réponse ou à réponse brève. Ces habiletés sont inférées en placant les élèves devant des tâches complexes qu'ils doivent accomplir, par exemple : une composition écrite, le suivi d'une recette, l'exécution d'un mouvement en expression corporelle.

Pour ce qui est d'apprécier ou de juger ce genre de production la **note** vient en premier lieu. Il faut bien distinguer la valeur numérique directement attribuée à une performance et celle obtenue en comptant des bonnes réponses comme dans le cas des examens objectifs. Notes et scores se ressemblent mais ne sont pas le fruit d'un même processus. Dans certains pays comme en France, il fut un temps où les compositions écrites étaient directement notées sur 20, c'est-à-dire sans qu'il y ait eu un calcul arithmétique auparavant. Selon cette démarche, une personne juge peut attribuer directement une note dont la valeur se situe entre 0 et 20, un « registre » utilisé autrefois en évaluation de compositions écrites. Par exemple. à la lecture du récit rédigé par un élève, un enseignant peut lui attribuer la note « 17 » sur 20. Les critères d'évaluation, s'ils existent, nous sont inconnus, ce qui rend cette démarche d'appréciation hautement subjective, voire « idiosyncratique ». Il en va de même dans l'appréciation d'un film de la part de certains critiques. Après avoir décrit un « navet », certaines personnes n'hésiteront pas à lui accoler un « 2 » sur 10. Nous devons comprendre que cette valeur n'est pas le fruit d'un calcul, ni une somme. L'attribution d'une note telle que décrite s'inscrit dans un processus dit de **notation**. C'est ainsi que Legendre (2000, 904-905) décrit la notation en l'associant à l'attribution d'une **cote**.

Il n'est pas facile de distinguer finement entre note et cote. La note peut être exprimée en pourcentage ou sur un maximum relativement élevé (comme sur 20 pour les compostions écrites). La cote renvoie plutôt à un ensemble de chiffres ou de lettres en nombre très réduit. Les cotes peuvent s'échelonner de 1 à 5 ou de A à E, par exemple, et font partie intégrante d'une échelle d'appréciation composée de quelques échelons. Voici un exemple d'échelle qui pourrait être utilisée pour apprécier la ponctualité :

non ponctuel **[1]** plus ou moins ponctuel **[2]** ponctuel **[3]** très ponctuel **[4]**

Les valeurs chiffrées 1, 2, 3 et 4 sont des cotes et ne proviennent d'aucun calcul. Et on ne saurait les qualifier de résultats de mesure. Surtout pas ! Il nous reste à préciser dans quelles circonstances ou avec quels outils d'évaluation les échelles d'appréciation et les cotes qui leur sont associées sont utilisées. Dans une grille d'évaluation, chaque critère est accompagné d'une échelle d'appréciation, ce qui en fait une démarche analytique. Dans le cas de certaines performances complexes ou dans une approche par compétences, les personnes juges peuvent recourir à une échelle unique d'appréciation ou échelle descriptive globale.

-

GRILLE D'ÉVALUATION POUR UN RÉSUMÉ

Choix du titre (rapport au contenu du résumé)
aucun rapport **[1]** faible rapport **[2]** évocateur **[3]** très évocateur **[4]**

Justesse des idées de l'auteur
aucune idée **[1]** quelques idées **[2]** la plupart **[3]** toutes les idées **[4]**

Concision (nombre de répétitions)
reprise textuelle **[1]** plusieurs **[2]** quelques unes **[3]** aucune **[4]**

Qualité de la langue
médiocre **[1]** acceptable **[2]** bonne **[3]** excellente **[4]**

Les échelles d'appréciation qui composent une grille d'évaluation peuvent conduire à un résultat global pour autant que les cotes attribuées d'un critère à l'autre puissent être additionnées. Pouvons-nous appeler ce résultat note ou score ? Il n'est pas facile de trancher. Étant le résultat d'un calcul, la somme des cotes pourrait bien être appelée « score ». Nous devons comprendre que, selon cette façon de procéder, une faiblesse marquée à l'un des critères peut être compensée par une cote élevée à un autre critère. Ainsi, par exemple, deux individus peuvent se mériter un résultat total de 11 sur 16, d'après la grille d'évaluation donnée en exemple, sans forcément réussir aux mêmes critères. C'est tout le problème d'interprétation que pose ce procédé de simple addition des cotes obtenues aux divers critères d'évaluation d'une production.

C'est dans le but de corriger cet état de fait que l'échelle descriptive globale a été développée comme modèle de procédé d'évaluation de performances ou de productions complexes. Pour divers aspects de sa performance ou selon les qualités de sa production traitées simultanément, un élève ne reçoit qu'une seule cote qui lui est attribuée directement et globalement.

Il est important de noter que l'objet de cet addenda est de faire état de diverses façons d'exprimer un résultat chiffré, qu'il s'agisse d'un **score**, d'une **note** ou d'une **cote**. Cependant, nul ne peut préciser la nature d'une valeur chiffrée comme fruit d'une évaluation. Il faut pouvoir retracer le procédé qui a conduit à ce résultat. Par exemple, un élève a obtenu « 18 » en multiplication. Est-ce un score ou une note ? Pour ce qui est des cotes, elles se distinguent facilement des autres modes d'expression, leur registre étant limité à quelques valeurs de résultat (entre 1 et 4 ou entre A et D, par exemple).

-

ADDENDA-2
Le modèle de réponse comme
liste de vérification : exemple

Dans un cours gradué en évaluation formative (hiver 2002), les étudiants et les étudiantes ont appris ce qu'est la mesure au sens strict à même des exemples pris dans les sciences physiques. En sciences humaines, le comptage d'éléments pris comme unités n'est pas aussi rigoureux. Pour évaluer ce que les étudiants retirent de toutes ces considérations et comment ils peuvent faire appel à des notions théoriques, la tâche d'évaluation qui leur a été posée à l'examen terminal est présentée ci-dessous. Le premier encadré s'adresse directement aux étudiants (consigne, données et question soulevée). Il s'agit d'une question parmi plusieurs qui composaient l'examen.

<table>
<tr><td colspan="2">Voici deux exemples d'épreuves, l'un en arithmétique, l'autre en français, épreuves utilisées avec des élèves du primaire :</td></tr>
<tr><td>Effectuer les opérations suivantes:
a) 12 + 15 = ____
b) 2,3 + 1,45 + 0,04 = ____
c) 7 ÷ 2 = ____
d) 2 2/3 x 14 3/4 = ____
e) 14,23 ÷ (- 7,4) = ____</td><td>Accorder le mot souligné, s'il y a lieu:
1- Carole et Pierre <u>mange</u> beaucoup.
2- Des ciels <u>bleu -clair</u> .
3- Ils se sont <u>laissé</u> prendre.
4- C'est correct!, <u>pense</u> -t-il.
5- Il a acheté deux <u>Picasso</u>.</td></tr>
<tr><td colspan="2">La somme des tâches réussies à l'une ou à l'autre épreuve est-elle véritablement un résultat de mesure? Dans une école, les opinions sont partagées et les personnes difficiles à convaincre pour ce qui est de changer leur position. Quel est votre point de vue et comment arriveriez-vous à le justifier ? (NOTE: le nombre de tâches ou de questions n'a pas d'importance ici).</td></tr>
<tr><td colspan="2"></td></tr>
<tr><td colspan="2"></td></tr>
<tr><td colspan="2"></td></tr>
<tr><td colspan="2">Vous pouvez ajouter une ou deux lignes au verso de la feuille-réponse.</td></tr>
</table>

Liste de vérification (modèle de réponse)

<table>
<tr><td>Réponse négative explicite (« ce n'est pas un résultat de mesure ») ..... /2 [__]
Justification:
…tâches ne correspondent pas à des unités (d'égale longueur) ..... /1 [__]
…exemple : problèmes différents (addition, division, etc.) ou hétéroclites..... /1[__]

_____ou ____

Réponse positive au conditionnel…(ce serait, ce pourrait être) ..... /2 [__]
avec postulat (supposition explicite) que ce sont des unités… ..... /1 [__]
…si les problèmes de nature différente étaient de même difficulté..... /1[__]

RÉPONSES NON ACCEPTÉES
…ne se prononce pas sur le cas précis qui est présenté (reprend la théorie)
…chaque épreuve ne comporte pas assez de problèmes ou de questions</td><td>__/4</td></tr>
</table>

-

REMARQUE TRÈS IMPORTANTE

La démarche d'appréciation qui vient d'être décrite est fondée essentiellement sur les qualités d'une réponse-produit. Est-ce suffisant pour inférer que l'individu observé a su mobiliser toutes ses ressources. Peut-être que oui, s'il s'agit de savoir-faire ou de stratégies (auxquelles ressources des savoirs particuliers sont dédiés). Peut-on en dire autant des savoir-être ? Pas sûr ! Par exemple, il faudrait que l'habitude d'auto-réflexion ou le souci de précision ou encore des préoccupations d'ordre éthique laissent des traces. Il s'agit là d'un problème de taille dans le traitement de tâches complexes pour inférer une ou des compétences.