Pas de fausse note

Daniel Reisz

L'article ci-dessous est extrait d'une lettre d'information de l'association Pénombre, datant du printemps 2002. À l'heure où la question de la note chiffrée semble revenir sur le devant de la scène, il nous a paru pertinent de redonner la parole à Daniel Reisz, qui aborde ici le sujet avec un regard scientifique.

Daniel Reisz est inspecteur pédagogique honoraire, membre actif de longue date de notre association.

Annabelle : Pierre, qu'est-ce que tu préfères, les beignets à l'ananas, ou les beignets à la pomme ?
Pierre : à l'ananas... mais j'aime mieux les nougats chinois. Les nougats chinois, je leur mets 7 ou 8,... plutôt 8.
Annabelle : moi, je leur mets 7.
Pierre : oui, mais toi, tu notes au hasard.
Annabelle : non, je dis ce que j'aime et ce que je n'aime pas.
Pierre : justement, tu ne peux pas noter comme ça. Imagine que tu sois une maîtresse : tu n'aimes pas un enfant ; donc tu vas lui mettre une mauvaise note. Non. Moi, si je mets 8, c'est parce que pour moi, les nougats chinois ont deux défauts, donc je leur retire deux points. Le premier c'est que les nougats, c'est comme la limonade, c'est très sucré ; plus on en prend, plus on a soif…
Annabelle : mais justement, pour moi, les nougats, ils ont trois défauts : le premier c'est comme tu dis ; le deuxième, c'est que je ne les aime pas trop…

Noter : action de mettre une note.
Note (chiffrée) : appréciation donnée selon un barème préalablement choisi (Robert) ; appréciation […] de quelqu'un, de son travail, d'un devoir (Encyclopédie Larousse).
Classer : mettre un certain ordre.

On retiendra surtout dans ces extraits de dictionnaires que *noter* c'est en quelque sorte *apprécier*, ce qui peut nous laisser sur notre faim pour une approche un peu scientifique de cette affaire. L'action de *noter*, de *classer* ne peut être sortie de son inscription sociale. On ne note pas, on ne classe pas sans but plus ou moins explicite. Et dans la procédure de notation, de classement, à côté des procédures techniques, il y a des questions de pertinence intellectuelle. Pourquoi une faute de grammaire est-elle plus grave qu'une faute d'orthographe ? Pourquoi l'ergonomie d'un caméscope compte-elle autant que la qualité de son objectif ? Même s'il peut paraître illusoire de vouloir isoler cet aspect, nous essayerons dans ce texte de nous intéresser aux procédures techniques de notation, de classement comme tentatives de *mathématisation du réel* dont l'ambition est de nous sortir d'une simple perception subjective par une démarche qui se veut plus rationnelle.

Noter, est-ce mesurer ?

Mesurer sous-entend qu'on ait défini *au préalable* la notion de *grandeur* à mesurer. Dans un ensemble d'objets de même nature (hôpitaux, copies, fonctionnaires, magnétoscopes…), introduire la notion de grandeur consiste à définir ce que sont des objets équivalents. Ce n'est évidemment pas suffisant. On peut très bien regrouper des objets ou des individus en classes d'objets équivalents (les femmes et les hommes, ou encore, comme le faisaient

les théoriciens du racisme, les aryens, les juifs, les slaves, les roms, sans parler des noirs, des jaunes, des métis...) sans qu'on puisse passer ensuite à une mesure ou à une notation. On peut parler ici de simples classes *nominales* : on peut *nommer* les classes (les hommes, les femmes, les noirs, les aryens, les juifs…). On peut aussi décider de les *classer* : les hommes devant les femmes, les aryens devant les juifs…

Au delà d'un simple classement, on peut introduire une notion *d'écart* plus ou moins grand entre deux grandeurs classées en les *repérant* sur une échelle numérique établie à partir de valeurs attribuées arbitrairement à des grandeurs stables et répétables (par exemple, la température d'un corps par rapport à celle de la glace fondante et de l'eau bouillante). On parlera ici d'un *repérage numérique* des grandeurs.

Enfin, au sens plein du terme, la *mesure* demande qu'on sache définir la somme de deux grandeurs de même espèce. De telles grandeurs seront *additives*. On choisit alors une certaine grandeur comme unité (sa mesure sera par définition 1) et on mesure les autres grandeurs par le rapport entre la grandeur à *mesurer* et la grandeur unité, selon une procédure de mathématisation (passage de l'addition au rapport) que nous ne préciserons pas ici.

Le processus de mesure le plus simple est le *dénombrement* appliqué à un groupe d'objets ou à une succession de phénomènes tenus pour identiques (par exemple les gouttes d'un liquide, le nombre de décès dans un hôpital, le nombre d'élèves présentés au bac).

Dans les sciences d'observation, tout processus de mesure est tributaire des instruments de mesure et des techniques mises en œuvre. En outre, il peut être troublé par des phénomènes parasites qu'il n'est pas toujours possible d'éliminer. On ne peut donc attribuer une signification objective à une mesure qu'à condition de définir le degré de précision et de fidélité de l'instrument de mesure ainsi que la marge globale des erreurs et incertitudes liées intrinsèquement à toute mesure.

En *docimologie* (le discours sur la notation), nous sommes loin de réunir toutes ces conditions. Sommes-nous en présence de grandeurs ? C'est-à-dire avons-nous défini *au préalable* ce qu'étaient des objets (hôpitaux, copies, fonctionnaires, magnétoscopes,…) équivalents, meilleurs, moins bons ? NON ! Regardons ensuite l'échelle de notation, par exemple la plus courante en France, celle qui consiste à noter de 0 à 20. La mise en place d'une telle échelle exige plusieurs conditions. En particulier de définir la grandeur nulle, c'est-à-dire pour chaque catégorie (services hospitaliers, lycées, copies, fonctionnaires, magnétoscopes…) l'ensemble des nuls, et la *grandeur parfaite*, c'est-à-dire l'ensemble des *parfaits* de chaque catégorie. Puis de définir les échelons de l'échelle (pour les températures, on décide qu'elles sont proportionnelles à la dilatation du mercure). Rien de tel pour une échelle de notes ! On fait même exactement l'inverse : on note d'abord et on définit ensuite, par l'intermédiaire de la note, ce que sont deux hôpitaux, copies, fonctionnaires, magnétoscopes… : équivalents, distants d'un point, de deux points. Disons que l'usage intensif de ces procédures de notation

finit par instaurer sur nos objets une notion de grandeur repérable, même si la démarche est viciée dès le départ (et sans parler ici de la pertinence des critères de notation).

Une échelle de notes, telle que nous venons de la décrire, n'est en réalité qu'un repérage et n'est pas, par essence, munie d'une structure additive. On ne devrait alors faire ni sommes, ni moyennes de telles notes. Mais c'est là qu'enseignants, notateurs, testeurs, évaluateurs, succombent presque tous à la magie du nombre, au charme du résultat global chiffré, à l'illusion du mesurable. On ne peut évidemment pas aller, sans contorsions multiples, jusqu'à la grandeur mesurable pour la notation. Et pourtant on y va bien souvent, c'est-à-dire qu'on va s'autoriser à faire des moyennes (pondérées ou non), de la même façon qu'on fait de l'eau tiède avec de l'eau chaude et de l'eau froide (on reste alors sur la même échelle de notes, par exemple de 0 à 20) ou à faire des sommes pondérées de notes (on se retrouve alors sur une autre échelle, de 0 à x, où x est la somme pondérée des « 20 »). Dans une démarche scientifique, on mesure, on compare, on additionne des grandeurs de même nature. Peut-on additionner, ou faire la moyenne, d'une note d'anglais et d'une note de mathématiques d'un élève, des notes des différents services d'un hôpital, sous le simple prétexte que les notes (et non les grandeurs notées) sont des nombres « de même nature » ? Cela a-t-il un sens intrinsèque ou cela a-t-il pris du sens parce que nous le faisons ? La réponse à cette question est loin d'être simple. Nous pouvons trouver une telle démarche totalement aberrante ou, à l'opposé, trouver qu'une moyenne représente un indicateur pertinent d'une « valeur moyenne » de l'élève, de l'hôpital, du caméscope, à partir d'une notation de différentes composantes.

Avec un bon barème, ce sera plus juste !

Admettons qu'on note quand même, malgré tout et parce que dans la société actuelle c'est pratiquement incontournable. Remarquons qu'il y a alors plusieurs démarches possibles (ici on ne s'intéressera pas à la pertinence des critères, mais uniquement à la stratégie de notation). On peut noter intrinsèquement, globalement, sans barème, en s'appuyant sur une compétence et/ou une expérience plus ou moins avérées : telle dissertation vaut 11/20, l'ergonomie de tel magnétoscope vaut 14/20, tel chirurgien vaut 16/20, tel Côtes-du-Rhône 13/20. Les docimologues appellent une telle note une *note d'estime*. Souvent on y ajoute un petit commentaire « *pour justifier la note* ». Mais parfois le notateur est lui-même soumis à des contraintes : un fonctionnaire n'est pas noté intrinsèquement, mais de telle façon qu'il se trouve dans une zone de notes qui lui assure tel avantage et lui barre tel autre en matière de promotion, d'avancement, de mutation. On pourrait parler de *notation indirecte* ou de *notation rétrograde* (non, ce n'est pas un mauvais jeu de mot !). Cette idée de notation rétrograde, c'est-à-dire d'une notation fabriquée pour donner les résultats désirés, est présente dans la plupart des processus de notation, ne serait-ce qu'au nom de la pertinence de la notation envisagée. Parfois cette idée est simplement diffuse, implicite (on pense réellement ne « mesurer » *que* la performance d'un élève, sans avoir

f*abriqué* une épreuve *ad hoc* dont on sait d'avance qu'elle va distribuer les notes « comme il faut »), parfois elle est explicite (quelles épreuves, quels barèmes fabriquer pour avoir 80 % de reçus au baccalauréat ?).

Souvent, pour « *être plus juste* », on se donne un barème. *Plus juste* : l'ambiguïté de cette formule est tout à fait remarquable parce qu'elle articule deux sentiments forts du notateur, du noté et de la société environnante : justice et précision scientifique. II y a plusieurs types de barèmes : tant de points pour une faute de grammaire, tant de points pour une faute d'orthographe. Tant de points si on saute 1 m 20, tant de points si on saute 1 m 30, etc. De tels barèmes peuvent se discuter quant à leur pertinence pédagogique, mais ont ensuite le mérite d'être indiscutables (ou presque) dans leur application (on parle de *note formalisée*). Nous sommes là dans la même situation que lorsqu'on repère la température grâce à la dilatation du mercure dans le tube du thermomètre. Plus utilisée est la pratique de barèmes tels que le style sera noté sur 5, le plan sur 8, l'argumentation sur 7, la première question sur 3, la seconde sur 5, etc.
Mais de tels « barèmes » ne sont que des démultiplications du processus de notation d'estime et produisent souvent des dérives importantes par effet cumulatif des habitudes individuelles des notateurs.

T'as quoi comme moyenne, toi ?

Dans le milieu scolaire, l'usage de moyennes à des fins diverses (bulletin, passage de classe, examen, concours, classement trimestriel pour parents et administration avides de ce genre d'information…) est monnaie courante. Cela dit, les moyennes ne se rencontrent pas qu'en milieu scolaire : un classement comparatif de services hospitaliers se fait à travers une moyenne d'indicateurs hétérogènes de par la nature des objets notes (notoriété, mortalité, soins ambulatoires, durée de séjour...) et de par la nature des indicateurs chiffrés utilisés (notes entre 0 et 3, pourcentages, nombre de jours, bascule « oui/non »...) et celui des caméscopes itou à partir de la note d'ergonomie, de celle du zoom, de celle de la mécanique… Les gens aiment les classements : faisons des moyennes pour pouvoir classer facilement.

Pour fixer les idées, nous allons nous placer dans le contexte scolaire. Que veut-on faire, que fait-on en général ?
- comparer des séries de notes (notes des élèves de différentes classes, notes des élèves d'une classe dans les différentes disciplines, notes des copies de différents correcteurs),
- comparer plus globalement, plus synthétiquement un ensemble de résultats par l'intermédiaire de moyennes de plusieurs séries de notes, dans le but, par exemple de valider, justifier telle ou telle décision.
On peut utiliser des moyennes dans plusieurs directions : faire la moyenne dans une discipline de l'ensemble des notes obtenues par les différents élèves ; faire la moyenne de chaque élève sur un ensemble de disciplines ; faire une moyenne à travers le temps des notes obtenues au sein d'une discipline par un élève donné.

Pour comparer des séries, on se contente très souvent de ne regarder que la moyenne sans s'occuper de leur dispersion autour de cette moyenne.

Lorsqu'un notateur note une production d'élèves (professeur dans sa classe, examinateur face à ses copies, interrogateur à un oral…), il produit une série N de notes que nous supposerons, selon l'habitude la plus courante, prises sur l'échelle [0 ; 20].

$$N = (x_1 ; x_2 ; x_3 ; \dots ; x_n)$$

Il y a ici n notes, c'est-à-dire n copies, n élèves.

Une façon classique de décrire synthétiquement cette série N est d'en donner une caractéristique centrale (moyenne ou médiane sont les plus utilisées dans la pratique) et une caractéristique de la dispersion (écart-type ou quartiles sont en général respectivement associés à moyenne et à médiane). Restons-en à la moyenne $\mu(N)$ et à l'écart-type $\sigma(N)$. Remplacer N par $\mu(N)$ et $\sigma(N)$ a les vertus et les défauts de toute synthèse : il y a simplification et amélioration de la lisibilité, que l'on paye par une certaine perte d'information.

Dans le quotidien du monde enseignant on s'occupe beaucoup des moyennes :
- la classe A est « meilleure » que la classe B, sous prétexte que $\mu(A) > \mu(B)$,
- le prof X est plus « vache » que le prof Y, parce que pour un même paquet de copies : $\mu(X) < \mu(Y)$,
- l'élève E passera en Première et l'élève F redoublera, sous prétexte que :
$\mu(F) < 10 < \mu(E)$.

On se préoccupe beaucoup moins de la dispersion des notes qui joue pourtant un rôle important.
Est-on bien conscient, dans un conseil de classe par exemple, que lorsqu'on fait la moyenne, pour chaque élève d'une même classe, de la note d'EPS (élément de la suite N(EPS) de moyenne $\mu(EPS)$ et d'écart-type $\sigma(EPS)$, affectée du *coefficient* 1) et de la note de français (élément de la suite N(F) de moyenne $\mu(F)$ et d'écart-type $\sigma(F)$, affectée du *coefficient* 3), le classement des élèves sera bien plus gouverné par la note d'EPS que par celle de français, malgré le jeu des coefficients, dès que $\sigma(F)$ est « nettement inférieur » à $\sigma(EPS)$?

Soyons conscients que, si le poids « technique » d'une discipline (il y a bien d'autres composantes dans le poids d'une discipline, ne serait-ce que la personnalité même du prof) est évidemment fonction du coefficient dont il dispose, il est aussi fonction de la dispersion de ses notes, dispersion qu'il peut influencer pour une bonne part.

Pour les moyennes, il faut être conscient d'un effet mécanique : faire des moyennes « rabote » la dispersion ! On est là aussi en présence d'un phénomène souvent mal perçu dans les conseils de classe ou dans des jurys d'examen : la contraction de la dispersion des notes lorsqu'on fait des moyennes. Pour bien comprendre le phénomène, considérons une situation tout à fait particulière, sachant que, dans la réalité, le phénomène de contraction est du même ordre de grandeur. Soit donc dans une classe de n élèves, les notes des k différentes disciplines :
$N(F) = (x_1 ; x_2 ; x_3 ; \dots ; x_n)$ les notes de français,
$N(HG) = (y_1 ; y_2 ; y_3 ; \dots ; y_n)$ les notes d'histoire-géographie,
$N(M) = (z_1 ; z_2 ; z_3 ; \dots ; z_n)$ les notes de musique.

Et supposons, *idéalement*, que les moyennes et les écarts-type de la classe soient les mêmes dans chaque discipline. Soit μ et σ cette moyenne et cet écart-type communs (le proviseur est très content

d'un tel consensus…). Supposons que la sacro-sainte moyenne de chaque élève soit calculée en appliquant le même coefficient à chaque discipline.

L'élève « 1 » aura donc pour moyenne : $m_1 = \frac{x_1 + y_1 + \ldots + z_1}{k}$ et ainsi de suite, ce qui produira la série des fameuses moyennes de chaque élève si déterminantes pour les décisions mettant en jeu leur scolarité ultérieure : M = (m_1 ; m_2 ; m_3 ; ... ; m_n).

Dans notre cas idéal, on aura évidemment $\mu(\mathrm{M}) = \mu$, mais combien de professeurs, combien de chefs d'établissement seront conscients que $\sigma(\mathrm{M}) = \frac{\sigma}{\sqrt{k}}$ c'est-à-dire qu'avec 9 disciplines l'écart-type est divisé par 3 et qu'avec 4 disciplines, l'écart-type est déjà divisé par 2 !

On m'objectera que je sous-entends que les séries de notes des différentes disciplines sont indépendantes les unes des autres et qu'ainsi je nie la notion de bon élève, d'élève moyen, de mauvais élève, c'est-à-dire que par-delà les disciplines il y a une certaine cohérence, une certaine corrélation entre les notes (les bons sont plus ou moins bons partout, les mauvais…). Certes cela est plus ou moins vrai et cela atténuera donc le phénomène de « rabotage », mais sauf à imaginer que chaque élève ait la même note dans chaque discipline, il y aura peu ou prou du rabotage. On peut ainsi avancer l'idée qu'en général le *contrôle continu* (note décisive comme moyenne de plusieurs notes de contrôles) sera moins discriminant en termes de dispersion des notes que l'épreuve unique. Cela ne préjuge en rien les avantages et inconvénients de l'une ou l'autre pratique.

Est-il normal que la distribution des notes soit normale ?

Autre question : est-il naturel, normal, souhaitable, indispensable que la dispersion des notes (sous réserve qu'il y en ait une certaine quantité) ait l'allure de la célèbre courbe de Gauss (courbe en cloche) ?

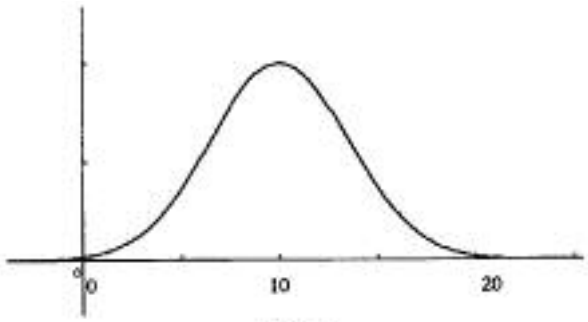

Figure 1

La distribution des notes selon une courbe de Gauss (pas forcément centrée sur la moyenne 10, comme pourrait le faire croire la figure) est-elle une loi naturelle ou un modèle social porté par la culture du notateur qui consciemment ou non fabrique épreuves et barèmes pour y aboutir ? II faut dire que le modèle a tout pour séduire : symétrie harmonieuse, cohérence avec les idées dominantes (petite élite de « forts », masse de « moyens », minorité de « faibles »).
La courbe de Gauss serait justifiée si les performances des élèves face à un travail et la notation qui en est faite suivaient « objectivement » une *loi normale* (au sens statistique et mathématique de cette expression).

Une telle loi a une définition précise et décrit des phénomènes aléatoires fréquents dans la nature. Mais nous n'avons aucune certitude quant à l'adéquation de la situation scolaire décrite précédemment à une telle loi.

On peut se demander si la référence gaussienne, invoquée pour rassurer les acteurs scolaires (élèves, professeurs, parents, jurys, institution) n'agit pas en fait de façon perverse. La littérature consacrée à la docimologie rapporte souvent l'expérience suivante : on fait corriger par plusieurs correcteurs (afin de lisser les travers individuels) un même lot de copies. On aboutit en général à une distribution des notes suivant une belle courbe de Gauss.

On garde ensuite de ce lot le quart des copies les plus faibles et le quart des copies les meilleures, qu'on réunit en un seul lot qu'on fait corriger à nouveau (par d'autres correcteurs !). La logique voudrait qu'on obtienne une distribution bimodale (courbe à deux bosses dont l'allure serait proche de la figure 2). Dans la plupart des cas il n'en est rien, les correcteurs recréent une distribution qui suit une courbe de Gauss.

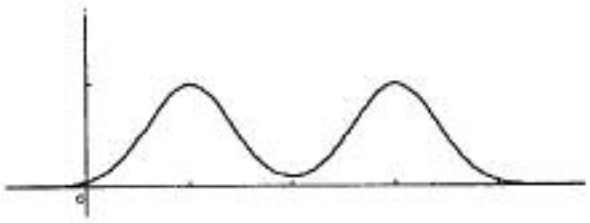

Figure 2

Au mieux elle sera plus étalée (écart-type plus grand) que la courbe initiale.

D'autres expériences de docimologie confirment cette tendance naturelle à produire des notes distribuées de façon gaussienne : derrière une loi statistique se cache sans doute aussi un biais comportemental de nature individuelle, sociale ou institutionnelle. Et d'autres expériences permettent de mettre en évidence d'autres biais comportementaux des correcteurs, très souvent induits par l'attente de l'institution, des parents, des élèves.

Les notes distribuées selon une courbe de Gauss répondent en général à une fonction plus ou moins explicite : séparer le bon grain de l'ivraie (ou plutôt *définir* le bon grain et l'ivraie) tout en gérant une masse de « moyens ». Lorsqu'on se propose simplement d'extraire d'une population une petite élite (certains concours par exemple), épreuves et corrections s'adaptent pour produire plutôt une distribution dont la courbe aura l'allure en « i » de celle de la figure 3.

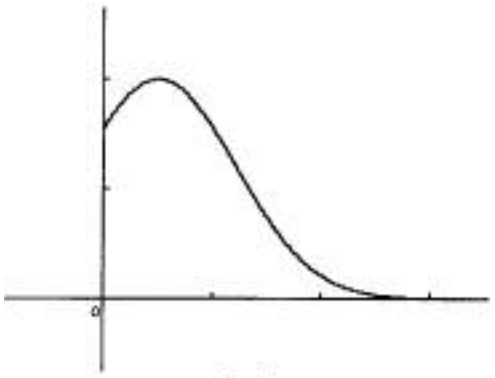

Figure 3

Au contraire, lorsqu'il s'agit d'un examen largement accordé (le brevet des collèges par exemple) on trouvera plutôt des distributions dont la courbe aura une allure en « j » (figure 4). C'est bien la fonction (de l'épreuve à noter) qui crée en quelque sorte l'organe (du notateur).

Figure 4

Conclusion

Noter, classer est un fait social, un besoin, une envie d'ordre socioculturel. Essayer d'objectiviser une telle procédure, de se donner les instruments pour réaliser cette objectivisation peut paraître une démarche naturelle, d'autant plus qu'intuitivement on a envie de rapprocher une procédure de classement / notation de procédures plus scientifiques, plus mathématiques : *mesurer* / *compter* / *ordonner*. Nous avons regardé dans ce texte quelques-uns des principes qui gouvernent cette modélisation, mais nous avons aussi essayé d'en montrer les limites et la relativité des démarches utilisées. Tout cela devrait inciter à beaucoup de prudence notateurs, classificateurs, sondeurs, enquêteurs, médiatisateurs… D'autant plus que nous n'avons pas du tout abordé un autre facteur qui devrait inciter à encore plus de prudence, c'est tout ce qui concerne la pertinence même des critères de jugement, la neutralité et la stabilité de jugement de l'évaluateur, l'usage de synthèses parfois bien réductrices, voire biaisées.

Tout cela me rend bien sceptique et je vais me consoler en buvant un coup de rouge qui a eu 14/20 dans le guide X et 17/20 chez Y, sur la terrasse d'un café « trois étoiles » d'une ville qui a 15/20 pour sa qualité de vie : le bonheur !