Les expériences existent depuis bien longtemps, mais tout se passe comme si elles n'avaient jamais eu lieu. Soit on ne les connaît pas, soit on n'y "croit pas" ! (Voir: Biblographie sur la docimologie)

Second ou avant dernier ? Admis à l'agrégation ou éliminé ?

En 1930 le professeur Laugier sème un malaise pernicieux dans les milieux universitaires, en effectuant une expérience de multicorrection de copies d'agrégation d'histoire puisées dans les archives. 166 copies ont été corrigées par 2 professeurs travaillant séparément, sans connaître leurs appréciations respectives.Tous les deux avaient une longue expérience et corrigeaient méticuleusement. Les résultats furent surprenants. La moyenne des notes du premier correcteur dépassait de près de deux points celle du second. Le candidat classé avant dernier par l'un était classé second par l'autre. Les écarts de notes allaient jusqu'à 9 points. Le premier correcteur donnait un 5 à 21 copies cotées entre 2 et 14 par le second ; le second donnait un 7 à 20 copies cotées entre 2 et 11,5 par le premier. La moitié des candidats reçus par un correcteur était refusée par l'autre.

Cette expérience caractéristique de docimologie a amené des chercheurs de plus en plus nombreux à s'interroger sur les sources d'erreurs des procédures d'évaluation traditionnelles.

DIFFÉRENTS TYPES D'ERREURS

F. Bacher (1969) distingue trois sources d'erreurs.

1°) "La première source d'erreurs est due à **l'évaluateur lui-même** qui note autour d'**une moyenne plus ou moins élevée** et qui **disperse plus ou moins ses notes autour de cette moyenne**.

De plus les évaluateurs ne classent pas dans le même ordre une même série de travaux ou de réponses.

2°) La seconde source d'erreurs tient, dans les examens traditionnels, **au choix du sujet même** de l'examen. Il est en effet peu satisfaisant de généraliser à l'ensemble des candidats une constatation "ponctuelle", fondée sur l'un seulement des innombrables sujets qui auraient pu lui être proposés.

3°) La troisième source d'erreurs vient **des élèves**. D'un jour à l'autre, d'un moment à l'autre, se produisent des fluctuations aléatoires de la capacité qu'il s'agit d'évaluer. De plus une certaine forme d'évaluation peut défavoriser de façon systématique un type d'élèves (la question s'est posée notamment à propos des épreuves orales)"(Reuchlin, 1.67 p. 215)

PREMIÈRE SOURCE D'ERREURS : L'EVALUATEUR

Les évaluateurs ne sont pas d'accord entre eux.

- * **En 1932** eut lieu une des premières enquêtes docimologiques. **La Commission Carnégie** effectua une expérience de multicorrection en prélevant, au hasard, cent copies dans les archives du baccalauréat à Paris. Ces copies furent distribuées à 6 groupes de 5 examinateurs. Les disciplines concernées étaient : le français, la philosophie, le latin, les mathématiques et la physique. On demanda aux examinateurs de noter les copies et de fournir un rapport sur

- les qualités exigées

- le classement des dites copies - la méthode de notation utilisée.

Les résultats montrèrent une forte dispersion des notes attribuées à chaque copie par les correcteurs. Aucune copie ne reçut deux fois la même note. L'écart maximum des notes dépassa les prévisions.

Une copie de fran çais est notée 3 et 16 ; en philosophie et en latin l'écart maximum est de 12 points.

Les mathématiques et la physique, réputées pour des sciences exactes, ne sont pas épargnées : **l'écart maximum est respectivement de 9 et 8 points.**

D'autre part, **le classement des copies variait fortement d'un correcteur à l'autre.**

- * En 1975 **l'Institut de Recherche sur l'Enseignement des Mathématiques de Grenoble** entreprend une expérience analogue de multicorrection.

Un échantillon de 6 copies photocopiées de mathématiques (niveau B.E.P.C.) est soumis à 64 correcteurs, **avec un barème**, très précis, **sur 40 points.**

Les résultats confirment ceux de l'enquête précédente, effectuée 43 ans plus tôt. **La dispersion des notes atteint près de 20 points.**

- * **Le Centre International d'Études Pédagogiques de Sèvres, l'Association des Professeurs de Français, le Centre Pédagogique Régional de Toulouse, l'I.R.E.M. de Toulouse**, pour ne citer que les plus marquants ont mené des expériences du même type et sont arrivés aux **mêmes conclusions.**

- * De plus, Laugier et Weinberg ont montré que **la double correction est illusoire**. En effet il faudrait, pour obtenir une note "exacte" (Une note "exacte" étant une moyenne de notes telle que l'adjonction d'une autre note ne modifie pas sensiblement cette moyenne) .- 127 correcteurs en philosophie - 78 " en composition française - 28 en anglais - 19 en version latine - 16 en physique - 13 en mathématiques

Les divergences entre correcteurs ne sont pas seules en cause

En plus l'évaluateur n'est pas d'accord avec lui-même

Pas d'accord entre eux, les examinateurs ne se montrent pas davantage fidèles à eux-mêmes lorsqu'il s'agit de juger une seconde fois un devoir après un intervalle plus ou moins prolongé.

Les enquêteurs **anglais** en avaient été frappés. Laugier et Weinberg en **France**, avaient fait la même constatation. A leur demande un professeur de physiologie de la Faculté des Sciences accepta 37 copies -dactylographiées et anonymes - qu'il avait corrigées trois ans et demi auparavant.

Dans 7 cas seulement, il remit la même note au même devoir. Dans les 30 autres cas, il y eut des divergences comprises entre **1 et 10 points.**

L'admissibilité, avec ses nouvelles notations, aurait été modifiée ; la moitié des précédents admissibles aurait été refusée et la moitié des refusés déclarée admissible.

Le degré d'accord de ce professeur avec lui-même ne fut pas plus élevé qu'avec deux de ses collègues chargés de la même tâche : les coefficients de corrélation atteignant respectivement 0,58, 0,59 et 0,56.

On poursuivit plus loin l'expérience, et elle aboutit à un fait plus troublant encore.

On demanda **à une bachelière**, Paulette, intelligente mais **ignorant tout de la question traitée**, de noter à son tour ces compositions de physiologie, après les avoir lues une fois pour se faire une idée du sujet. Ses notes eurent une corrélation de 0,51 avec celles attribuées par les professeurs compétents. **La bachelière ne se trouvait pas plus en désaccord avec les spécialistes que ceux-ci entre eux". (in Science et Vie, 1968, n° 610).**

Avec des enseignants de l'I.R.E.M. de Reims, j'ai fait corriger 20 copies de B.E.P.C. par 30 professeurs de mathématiques après l'établissement d'un barème. J'ai corrigé aussi ces copies **sans les lire**, en regardant l'écriture et la présentation. Le maximum de l'écart pour une même copie était de 11 points sur 20. Ma note n'était jamais à une extrémité!

Cette expérience a provoqué un choc affectif considérable chez les enseignants participants: "Si c'est comme cela, je n'ai plus qu'a donner ma démission d'enseignant". Il ne s'agissait plus seulement d'information mais de connaissance (cognito-émotionnelle). Cette expérience pourrait être

faite avec profit dans les I.U.F.M.

""Prenons le cas d'un correcteur en face d'une copie anonyme d'examen. Il se forme dans la conscience du correcteur l'image d'un couple affectif ; il est là devant l'élève qu'il ne connaît pas mais qu'il essaie de voir à travers l'image d'un cancre ou d'un élève intelligent, ou d'un imbécile, ou d'un original... C'est en fonction de l'harmonie ou du désaccord de ce couple affectif que le correcteur va le plus souvent noter.

En corrigeant, il se forme d'une façon plus ou moins distincte une image d'élève sur laquelle il applique progressivement une étiquette type en fonction de sa répugnance ou de son admiration pour lesquelles peu de chose suffit, une jolie phrase, une de ses idées favorites qu'il retrouve exprimée, ou un cliché qui lui déplaît.

On constate des différences très grandes... à l'occasion de la même copie corrigée une deuxième fois à quelques jours d'intervalle, la première note ayant été effacée. Dans ce dernier cas, selon l'humeur du moment, la lecture de la veille, les soucis du jour , l'amabilité, la gentillesse ou l'insolence dont ont pu faire preuve les élèves dans la classe du jour précédent, les réactions sentimentales du correcteur ne sont pas toujours les mêmes et un nouveau couple affectif apparaît dans sa conscience, un couple dont les deux partenaires se sont trouvés changés."

(M. Marchand, Le couple de l'éducateur et de l'élève dans leurs relations concrètes. Université d'Alger

Les évaluateurs sont influencés

par des facteurs qu'ils ne soupçonnent même pas

Plus significatives encore sont les expériences de R. Rosenthal.

"Les expérimentateurs devaient étudier **le comportement de rats** et les noter, au cours d'un programme comprenant des techniques classiques (apprentissage du labyrinthe et boite de Skinner). Il fut dit à la moitié des expérimentateurs que les rats qu'ils devaient étudier avaient été spécialement choisis pour se comporter brillamment, à l'autre que les rats avaient été spécialement choisis pour être stupides et bien entendu, les rats avaient été choisis au hasard. Dans les deux études de ce type réalisées par Rosenthal les expérimentateurs croyant leurs sujets brillants obtinrent un bien meilleur apprentissage de leurs rats que ceux qui pensaient que leurs sujets avaient été choisis pour être stupides".

Rosenthal est allé encore plus loin en reproduisant une situation réelle de l'enseignement.

"Dans 18 classes d'une école primaire américaine, 20 % des élèves choisis rigoureusement au hasard, furent signalés à leurs

professeurs comme ayant eu des résultats particulièrement brillants à un test non verbal d'intelligence générale, permettant prétendument de prédire leur épanouissement intellectuel. Les enfants ne savaient rien, seuls les professeurs étaient au courant. Au bout d'une année, ces enfants avaient réellement un gain de quotient intellectuel franchement supérieur, en moyenne, à ceux du groupe de contrôle". (in Atome n° 242 article de H. Pequinot,).

Citons **les travaux de C. Chase** qui ont mis en évidence l'influence de la qualité de l'écriture sur les notes attribuées aux rédactions (The impact of some obvious variables, in journalof Educational Measurement, 1968) et ceux de Wilson qui ont montré que les enfants des zones géographiquement défavorisées sont cotés au plus bas, par le fait même qu'ils appartiennent à ces zones.

L'évaluation est fortement teintée par la personnalité de l'évaluateur.

L'enseignant s'évalue autant qu'il évalue ses élèves!

DEUXIÈME SOURCE D'ERREURS : LE SUJET

Le choix même du sujet d'un contrôle ou d'un examen, les conditions de passation, interviennent également dans une large mesure. Des éléments, choisis pour éviter toute variation intempestive, tels que barèmes et coefficients de pondération sont loin de remplir les conditions de fiabilité souhaitées.

Le barème

La plus grande partie des épreuves d'un examen fait intervenir un barème plus ou moins précis selon la matière. Ce barème a pour but d'uniformiser la codification des appréciations. Mais ce barème que les correcteurs sont "priés de respecter scrupuleusement" est-il un instrument fiable ? Plusieurs équipes de chercheurs se sont penchés sur ce problème.

L'I. R. E. M. de Rennes a fait corriger 22 copies

de mathématiques (c'est l'épreuve de mathématiques, où le barème a une place fondamentale, qui a le plus souvent été testée) du B.E.P.C. par 10 professeurs. **Cinq d'entre eux l'ont fait avec barème, les autres sans barème**. L'analyse des résultats a mis en évidence trois points

- les utilisateurs du barème ont corrigé plus sévèrement

- l'écart entre les notes extrêmes est moindre quand on tient compte du barème

- le barème ne supprime pas la dispersion des notes.

Devant la déception de voir le barème se limiter à un rôle vaguement modérateur, on a été amené à l'affiner, dans l'espoir d'obtenir une meilleure précision dans l'évaluation.

Le Groupe de Recherche de Montauban dirigé par Cransac et Dauvisis s'est penché sur ce problème en 1975. Il leur a fallu plus de15 heures de travail pour se mettre d'accord et élaborer un barème très précis (sur 120 points). Malgré cet effort" pour une copie dont la note varie de 4 à 13 sur 20, chaque correcteur concerné a justifié sa notation et il a été impossible de donner raison à l'un ou à l'autre. Certains ont donné tour à tour raison à l'un puis à l'autre". D'autre part, le barème devient illusoire quand dans "une réponse peu claire, l'un voit un bon raisonnement, l'autre un raisonnement faux ; ce qui est inquiétant"

"**le barème imposé est appliqué en fonction de la personnalité des correcteurs**.

Certains l'appliquent à la lettre en faisant fi de leurs sentiments et de leur expérience pédagogique, d'autres ne peuvent ou ne veulent le faire et modulent le dit barème." (Ronceray, De l'influence du barème, I.R.E.M. Rennes)

. Nouvelle déception. **L'I.R.E.M. de Strasbourg** reprend cette expérience. On raffine le barème. On demande aux correcteurs "un travail d'appréciation de l'écart des copies en les confrontant deux à deux afin d'améliorer leur évaluation." Le groupe de recherche espère ainsi "faire des observations des copies deux à deux". Résultats : la fiabilité de la correction n'est pas améliorée. Sans se décourager on fait une "tentative de codage des comportements de réponse" avec "détermination d'un barème associé à ce codage". Là encore c'est l'échec. Mais ces travaux remarquables amènent les chercheurs vers une autre direction : **l'analyse plus soigneuse des sujets proposés.**

Le choix du sujet

La présentation et la formulation du sujet d'une épreuve peuvent-être source d'erreurs et la note d'un candidat ne pas refléter sa valeur propre. En effet, peut-on évaluer sérieusement les qualités d'un élève qui n'a pas su faire un problème dont les questions sont dépendantes et la première piégée ? Peut-on évaluer sérieusement les qualités d'un élève qui disserte "hors sujet" quand, au cours d'une expérience de multicorrection, une copie de français a été annotée "sujet bien traité" par un correcteur et "sujet non traité" par un autre correcteur ?

La formulation d'une question portant sur un concept mathématique donné détermine souvent l'exactitude de la réponse, tout comme, selon Piaget, la façon de poser une question orale à un candidat influence la réponse de ce candidat.

On retrouve dans ces études sur le barème l'évaluateur

Le Groupe de Recherche d'Epernay de l'I.R.E.M. de Reims a effectué une enquête sur ce thème en 1975. Il a mis en évidence la variabilité de l'exactitude des réponses selon les différents items d'un concept donné.

Citons un exemple frappant : le concept de l'équation.

A la question :"On achète deux pains, on donne dix francs, la marchande rend une pièce de cinq francs et une de un franc ; quel est le prix d'un pain ?, on obtient 93,44 % de réponses exactes dans une population d'élèves de troisième.

A la question :"Résoudre dans R: 2x+6=10", le pourcentage de réponses exactes tombe à 81,49 %.

Enfin si la question prend la forme suivante "Résoudre dans R : 2000x+6000=10000." on n'obtient plus que 60,34 % de bonnes réponses.

Il suffit donc de multiplier des données numériques 2, 6, 10 par 1000 pour que un quart des élèves ayant bien répondu à la question précédente se trompe à la dernière. Les résultats de cette enquête incitent "à la prudence dans le choix des critères qui permettent de déterminer le degré d'acquisition et de compréhension d'un concept mathématique. On peut également se demander si la compréhension d'un cours ne dépend pas étroitement de sa formulation.

Enfin on peut s'interroger sur le désir inconscient d'échec ou de succès du professeur vis-à-vis de ses élèves.

La notation n'en est -il pas en partie le reflet?

C'est une question que je me suis souvent posée en tant que Président de jury de bac.

TROISIÈME SOURCE D'ERREUR : L'ÉVALUÉ

Une cause non négligeable de fluctuation réside dans deux facteurs difficiles à maîtriser et à estimer : la variabilité des candidats et des conditions extérieures lors du déroulement des opérations.

"**La réaction affective de chaque élève à la situation diffère selon sa stabilité émotionnelle du moment,** son état de santé, la pression familiale ; en outre, du fait du temps limité, un incident mineur, même le bris d'une pointe de crayon, constitue un handicap inaperçu par le professeur. Dans des épreuves comme la dictée, la place occupée dans la salle peut fausser les résultats. De tels facteurs d'irrégularité sont spécialement actifs au cours des périodes de rapide évolution physique ou intellectuelle ; en particulier, plusieurs auteurs les ont mis en évidence au moment de la puberté". (Hotyat, in Revue Française de Pédagogie, Janvier 1968)

"Ce qui frappe aussitôt, c'est **l'extrême variété et complexité des réactions individuelles des élèves confrontés aux épreuves** et aux estimations qu'en font les maîtres. Il apparaît alors à l'évidence que l'examen, voire la simple composition de contrôle, sont d'abord des "événements" à caractère historique, qui s'inscrivent dans le destin personnel de l'élève, dont il est difficile de les dissocier. Qu'on pense à tel enfant de sixième pour qui chaque épreuve scolaire marque l'heure d'une nouvelle et solennelle remise en question de son statut à l'égard du professeur, et plus encore de ses parents, tandis que tel autre, au même âge et dans la même classe, vit les examens comme une ennuyeuse corvée, dont il faut se tirer économiquement pour

conserver le minimum de quiétude indispensable à la survie en milieu d'école. Qu'on pense à la panique de quelques uns devant le baccalauréat, à l'excitation conquérante de quelques autres, à la conduite soumise et humblement séductrice d'un certain nombre, au fatalisme de ceux-ci, à l'ingéniosité défensive, parfois malhonnête, de ceux-là, et ainsi de suite. Il est clair que ces conduites pèsent lourdement sur le résultat final de l'examen écrit ou oral, même si l'on admet, par impossible, que le maître demeure, pour sa part, à l'abri de toute variation ou inconsistance dans l'appréciation qu'il porte. Si l'on considère alors que de tels effets ne sont intelligibles qu'en référence à la personnalité de l'élève, et que cette personnalité ne se comprend à son tour qu'en fonction d'un contexte socio-familial et d'un devenir d'ensemble qui lui est rigoureusement propre, on admettra qu'il y a de quoi inquiéter un esprit soucieux de "justice" et rendre quelque peu suspectes les tentatives de réduction statistique concernant la notation, encore que les règlements rendent obligatoires classements et notations. A la limite en effet, il n'y a plus, en cette voie que des cas rigoureusement individuels, et donc incomparables". (Guillaumin, Docimologie et éducation, Les sciences de l'éducation, avril-septembre 1969).

"Lors d'**un sondage belge** sur le travail après la classe, la question suivante avait été posée à des élèves de 6°, 4°, 3°, 1° d'enseignement moyen et de dernière année d'école normale (2031 sujets). "Lors d'un examen ou d'une interrogation écrite, devenez-vous nerveux au point de ne pas obtenir des résultats correspondant à vos moyens et à vos efforts ?" Trois degrés étaient prévus pour les réponses : rarement, souvent, presque toujours. Plus de la moitié des interrogés signalaient ressentir souvent ou presque toujours un tel handicap. Contrairement à ce que l'on avait pu attendre, les fréquences étaient

à peu près égales chez les garçons (54,2 %) et chez les filles (54,6 %) apparemment plus émotives. En revanche, les élèves étant répartis en trois groupes selon leurs résultats dans leur classe, les fréquences étaient respectivement les suivantes

38,2 % dans le tiers supérieur, 59,3 % dans le deuxième tiers et 67,5 % dans le tiers inférieur.

Cependant, **des difficultés ont rarement un caractère pathologique** : en Grande Bretagne, un relevé des cas soumis aux Child Guidance Clinics, entrepris dans une aire géographique couvrant 212000 écoliers de 5 à 16 ans, révéla que sur 5 705 sujets ayant eu besoin de consultations des services d'hygiène mentale, le psychologue ou le psychiatre avaient signalé les examens comme facteurs de trouble dans 41 cas seulement". (Hotyat, Revue Française de Pédagogie, Janvier 1968).

En résumé : la note n'est pas mesure mais message, avec toute l'ambiguité de tous les messages!

La prise de conscience de la "complexité" de l'évaluation doit amener à relativiser les résultats et à introduire une souplesse dans ses conséquences, en particulier dans l'orientation.

Deux dimensions sont à prendre en compte, nous semble-t-il, pour apporter cette souplesse:

- le temps: le contrôle continu nous paraît préférable à une évaluation instantanée, couperet. (Le Bac Pro

actuel en est un exemple)

- l'ouverture: les examens par cumuls de modules , de crédits (l'européanisation des diplômes va dans ce sens) sont favorables dans la mesure où ils laissent "ouverte" la possibilité, avec le temps, de compléter et de revoir l'orientation tout au long de la vie.