INSTITUT DE RECHERCHE SUR L'ENSEIGNEMENT DES MATHÉMATIQUES

UNIVERSITÉ DE FRANCHE - COMTÉ

Objectifs et Evaluation

Fascicule 1

Généralités

Groupe "Evaluation"
animé par Antoine Bodin

Irem de Besançon

OBJECTIFS et ÉVALUATION

FASCICULE 1

GÉNÉRALITÉS

Groupe "ÉVALUATION" animé par Antoine BODIN

Mai 1983

Réédition 1994

Introduction

Cela fait plus de cinq ans que, sous l'impulsion d'Antoine BODIN, l'I.R.E.M. de Besançon s'est résolument engagé dans une recherche sur l'évaluation en mathématiques, pour ce qui concerne l'enseignement des collèges.

Et il s'agissait bien d'une recherche. Il y a cinq ans, en effet, les travaux sur l'évaluation étaient encore peu nombreux et le groupe de l'I.R.E.M. se lançait dans un domaine relativement neuf. D'autre part la méthode employée a consisté à s'appuyer sur une expérimentation constante permettant un contrôle rigoureux des résultats obtenus.

Le groupe s'est fixé comme objectif l'élaboration de tests pour les classes du premier cycle. Il était nécessaire d'expérimenter les diverses épreuves auprès d'un grand nombre d'élèves et, pour cela, le groupe a bénéficié de l'aide de nombreux collègues qui ont utilisé dans leurs classes les documents produits. C'est ainsi que, progressivement, les tests ont pu être repris, rectifiés et, aussi, étalonnés de manière à les situer dans un cadre moins relatif et à leur donner une validité indépendante des appréciations subjectives et, autant que possible, liée aux savoirs et savoir-faire raisonnablement exigibles à tel ou tel niveau du premier cycle.

Le travail persévérant poursuivi par le groupe a, peu à peu, permis la constitution d'un ensemble complet de tests pour les quatre classes du premier cycle qui ont fait l'objet de publications isolées. Mais il nous est apparu que cet ensemble méritait d'être mis, dans sa totalité, à la disposition des enseignants, non seulement sous la forme d'un produit fini, mais aussi accompagné de quelques uns des détours imposés par la méthode expérimentale et qui permettent de mieux comprendre les choix effectués. C'est ainsi que la publication de deux brochures consacrées l'une aux classes de 6e et 5e, l'autre aux classes de 4e et 3e, a été décidée : il s'agit des fascicules 2 et 3 du document que nous présentons aujourd'hui. Ils concernent, certes, les mathématiques, mais la méthode employée, largement décrite ici, peut probablement présenter un intérêt pour les autres disciplines ; c'est du moins le souhait que nous formulons.

Il eût été regrettable, nous a-t-il semblé, de ne pas trouver dans ce document quelques uns des repères théoriques sur lesquels la recherche entreprise s'est constamment appuyée. En effet, les travaux dans le domaine de l'évaluation sont désormais très nombreux et de multiples publications sont disponibles ; le groupe de l'I.R.E.M. a largement puisé dans cette vaste documentation et en a présenté quelques aspects dans des articles du "Bulletin de liaison de l'I.R.E.M.

de Besançon". Il a donc été décidé de rassembler quelques unes de ces considérations théoriques qui sont susceptibles d'intéresser nos collègues, mathématiciens ou non ; c'est l'objet du fascicule 1 dans lequel on trouvera quelques éléments généraux sur l'évaluation et dont la lecture est un préliminaire indispensable à l'utilisation des tests pour les diverses classes.

Ces trois fascicules constituent un tout auquel nous ne souhaitons pas qu'une valeur définitive soit attribuée. Bien au contraire, notre espoir est qu'ils suscitent critique et, surtout, désir de l'engager dans des actions voisines grâce auxquelles l'évaluation formative pourra montrer tout son intérêt pour l'enseignement et l'aide qu'elle peut apporter dans la lutte contre l'échec scolaire.

Jean-Claude FONTAINE.

REMERCIEMENTS

De nombreux collègues ont contribué à l'élaboration de ces documents, les uns en participant aux réunions, les autres comme correspondants dans les établissements de l'Académie.

Certains tests ont été passés par plus de 150 classes, les résultats-élèves qui ont été adressés à l'IREM nous ont permis de construire les étalonnages. Les commentaires qui les accompagnaient ont permis de faire un certain nombre de modifications que nous espérons positives. Que tous ces collègues, connus ou inconnus trouvent ici l'expression de nos remerciements.

Citons plus particulièrement ceux qui ont participé de façon régulière à nos réunions au cours des années 79-80, 80-81 , 81-82 et 82-83 .

GROUPE DE TRAVAIL DE L'IREM

Antoine	BODIN.....	Collège ORNANS
Suzanne	BOUTONNET.....	Collège BELFORT
Jean	CESAR.....	Collège VOUJEAUCOURT
Samuel	FAIVRE MACON.....	Collège MAICHE
Jean-Claude	FONTAINE.....	Directeur de l'IREM
Claude	FRELET.....	Collège DOLE
Lucien	GIGNET.....	Collège BAUME-LES-DAMES
Chantal	GOVIN.....	Collège CHAMPAGNOLE
Jean-Paul	GOVIN.....	Collège CHAMPAGNOLE
Pierre	GSELL.....	Collège BELFORT
Marie-Claire	KURRY.....	Collège BESANCON
Michel	MAGNENET.....	Lycée BESANCON
Bernard	ORIAT.....	Université BESANCON
Michelle	VOISIN.....	Collège DOLE
Michel HENRY		Université BESANCON
Nicole PELLETIER		Collège DOLE
Anne Marie PIERRE.....		Collège BESANCON
Antoine FENEUX		Collège BESANCON
Denise COFFE		Collège LUXEUIL
Clotaire PERNELLE		L.E.P Saint Amour .

GROUPE DE TRAVAIL DU COLLEGE D'ORNANS

Alain CONCE
Michèle FAIVRE
Bernard HUGONNOT
Bernard MATTEY
André MOYNE
Jacky PUGIN

Ce groupe a plus particulièrement travaillé à l'opérationnalisation des objectifs de la classe de sixième.

La frappe de ce document a été réalisée par Madame Elisabeth VUILLEMENOT ainsi que par Mademoiselle Annie SALOMON , et le tirage fait par Monsieur VRANA .

Aux collègues qui voudraient utiliser les fiches-élèves ou les tests récapitulatifs

Pour faciliter le travail de reprographie, on pourra se procurer à l'I.R.E.M. des batteries d'épreuves non reliées. Cette diffusion sera toutefois restreinte et il ne sera pas en principe possible de fournir les épreuves en nombre.

Se renseigner à l'I.R.E.M.

0

0 0

Aux collègues qui voudraient utiliser les fiches-élèves ou les tests récapitulatifs

Pour faciliter le travail de reprographie, on pourra se procurer à l'I.R.E.M. des batteries d'épreuves non reliées. Cette diffusion sera toutefois restreinte et il ne sera pas en principe possible de fournir les épreuves en nombre.

Se renseigner à l'I.R.E.M.

0

0 0

PRESENTATION GENERALE DES 3 FASCICULES

Depuis 1979, un groupe de travail sur l'évaluation se réunit régulièrement à l'I.R.E.M. Ce groupe s'est donné pour tâche d'identifier, d'expliciter et d'opérationnaliser les objectifs de l'enseignement des mathématiques au collège. Dans le même temps l'I.R.E.M. a organisé de nombreux stages et colloques sur l'évaluation.

Nous avons rassemblé sous le titre "OBJECTIFS et EVALUATION" l'ensemble des documents produits par le groupe de travail ainsi que les idées et réflexions qui ont guidé notre action et qui ont été exposés et débattues au cours des réunions et des stages.

Ces documents sont répartis dans trois brochures :

FASCICULE 1 : GENERALITES.

On trouvera dans ce fascicule à la fois des idées générales sur l'évaluation et la façon dont ces idées ont été prises en compte dans le travail de l'I.R.E.M. Notre réflexion a été grandement influencée par les travaux de chercheurs Français et Etrangers (voir bibliographie), par notre participation à des séminaires inter-I.R.E.M. et à des stages de formateurs I.N.R.P. Il est difficile de rendre à chacun ce qui lui revient, nous ne prétendons pas avoir fait une oeuvre originale, tout au plus espérons-nous avoir fait une synthèse utilisable par nos collègues.

Pour des raisons matérielles, ce fascicule est publié après les fascicules 2 et 3 qui contiennent les instruments d'évaluation et les référentiels. Nous souhaitons vivement que les utilisateurs actuels et futurs de ces épreuves ne négligent pas l'exposé de notre problématique. Nous avons certainement commis des erreurs et oublié de prendre en compte certains points qui peuvent paraître importants. Pour en juger il convient de ne pas se référer aux seules épreuves mais de prendre en compte nos intentions et la fonction essentiellement formative que nous entendions assigner à notre évaluation.

Ce fascicule est divisé en trois parties :

1ère PARTIE : Exposé de notre problématique et des objectifs poursuivis. Méthodes de travail.

2ème PARTIE : Développements de quelques points particuliers abordés succintement dans la 1ère partie.

3ème PARTIE : Articles divers sur l'évaluation déjà publiés dans le bulletin de l'I.R.E.M.

FASCICULE 2 : CLASSES DE SIXIEME ET CINQUIEME.

Ce fascicule contient les listes de micro-objectifs, les fiches-élèves correspondantes, les tests récapitulatifs (tests de validation) ainsi que leurs étalonnages sur une population représentative.

FASCICULE 3 : CLASSES DE QUATRIEME ET TROISIEME.

Contenu semblable à celui du fascicule 2.

PREMIERE PARTIE

Présentation de notre problématique, de nos objectifs et de notre méthode de travail.

Nous reprenons dans cette partie la présentation qui avait été faite dans la première brochure publiée en 1981. Nous avons préféré ne pas la modifier bien que sur quelques points nos conceptions se soient modifiés. On trouvera à la fin de cette présentation des remarques complémentaires permettant de l'actualiser.

> On pourrait déterminer les différents âges de la science par la technique des instruments de mesure. Chacun des siècles qui viennent de s'écouler a son échelle de précision particulière, son groupe de décimales exactes, et ses instruments spécifiques... nous voulons marquer la difficulté de déterminer les premières conditions de la mesure... On ne doit pas s'étonner de la prodigieuse variété des premiers thermomètres. Ils se trouvèrent bientôt plus nombreux que les mesures de poids. Cette variété est très caractéristique d'une science d'amateurs.
>
> G. BACHELARD (La formation de l'esprit scientifique)

I - Généralités

L'évaluation est un souci constant des enseignants. Il n'est pas possible d'enseigner une notion quelconque sans, en même temps se demander comment le message a été reçu. Dans la pratique, l'évaluation prend des formes très diverses selon qu'il s'agit de se renseigner sur l'efficacité de notre action, de renseigner les élèves sur leurs acquisitions ou de communiquer un bilan aux parents ou à l'administration. Il est bien connu que si nous sommes tous des évaluateurs, nos critères ainsi que nos pratiques diffèrent considérablement d'un individu à l'autre. Que signifie telle note attribuée à tel devoir ? N'avons nous pas eu des exigences excessives, ou au contraire trop modestes ? Et telle moyenne, où le même poids a été donné à des épreuves portant sur des sujets différents, passés dans des conditions différentes ?

Les recherches en docimologie et sur l'évaluation sont maintenant très abondantes, elles sont riches d'enseignements, mais restent pour l'instant l'apanage de quelques spécialistes.

Comment, concrètement, améliorer nos pratiques évaluatives ?

C'est pour tenter de répondre à cette question que nous avons été amenés à organiser au collège d'ORNANS pendant les années scolaires 78-79 et 79-80, une "recherche spontanée", puis en 79-80 un "groupe de recherche sur l'évaluation" à l'I.R.E.M. de BESANCON. A Ornans, l'ensemble des professeurs de mathématiques participe à ce travail. En ce qui concerne l'I.R.E.M., notre groupe réunit une dizaine de collègues du premier

cycle, du second cycle et de l'université. Beaucoup d'autres enseignants, directement ou non se sont associés à notre réflexion.

L'objet de cette première partie est de présenter cette expérience et de faire état de nos préoccupations générales et des considérations théoriques plus ou moins prises en compte dans ce travail.

- L'objet de l'évaluation : Que veut-on évaluer ? (Q U O I ?)
- Les buts de l'évaluation : Quelles finalités assigne-t-on à notre évaluation ? (P O U R Q U O I ?)
- Les méthodes utilisées : Modalités de l'évaluation (C O M M E N T ?)

Selon les réponses données à ces questions, diverses pratiques sont envisageables, chacune d'elles suppose des choix qu'il est important d'expliciter. Nous allons dans ce qui suit essayer de préciser les choix que nous avons fait.

II - Evaluation formative ou évaluation sommative ?

Le plus souvent, une évaluation conduit à une conclusion de type dichotomique : valable - pas valable, au niveau - pas au niveau, apte - inapte, reçu - collé. Lorsque l'on porte un tel jugement, dans la mesure où l'on ne prétend pas à l'infaillibilité, on court deux risques contradictoires :

- 1er risque : Déclarer apte quelqu'un qui est inapte
- 2ème risque : Déclarer inapte quelqu'un qui est apte

Il est clair que dès à présent se pose le problème des critères et des objectifs, nous y reviendrons. Pour l'instant, nous supposons que, dans le contexte, nous savons parfaitement ce que signifie "être apte" (ou être au niveau).

Par ailleurs, on peut distinguer au moins trois types d'évaluation :

- L'EVALUATION SOMMATIVE : qui intervient à la fin d'un cursus et qui est destinée à homologuer, à certifier, à sanctionner. Dans ce cas, la justice la plus élémentaire devrait conduire, en général, à minimiser le deuxième risque (pas toujours cependant ! Penser au recrutement des médecins).

- L'EVALUATION FORMATIVE : qui intervient en cours d'apprentissage et qui a pour but de renseigner le maître et l'élève, de mesurer le niveau d'adéquation entre les attentes de l'un et les capacités de l'autre. Cette évaluation rétroagit immédiatement sur la formation. Dans ce cas, c'est le premier risque qu'il convient de minimiser.

- L'EVALUATION DIAGNOSTIQUE ou encore PRONOSTIQUE qui est celle qui devrait essentiellement être prise en compte pour l'orientation. C'est aussi celle qu'il faudrait mettre en oeuvre avant d'aborder une question nouvelle pour savoir à quel niveau nous situer et ne pas sauter allègrement sur les prérequis. Le risque à minimiser dans ce cas dépend de la situation. Par exemple, en ce qui concerne l'orientation, il pourra dépendre de l'existence éventuelle de passerelles.

Quoi qu'il en soit, l'évaluation ne devrait pas être conduite de la même façon dans les différents cas. En particulier, il n'y a aucune raison pour que la somme d'évaluations formatives même bien conduites fournisse une évaluation sommative valable, ni pour qu'une succession d'évaluations sommatives soit profitable à la formation des élèves. Noizet et Caverni (psychologie de l'évaluation) soupçonnent les enseignants de ne pratiquer en fait que l'évaluation sommative : "L'évaluation sommative a pour résultat de différencier, l'évaluation formative a pour intention d'homogénéiser... les exercices en cours d'année qui devraient être envisagés dans une intention formative sont presque toujours pris dans une intention sommative". Les auteurs voient dans ce fait, la raison principale de l'intérêt quasi exclusif des élèves pour leurs notes.

Tout en reconnaissant l'importance des autres types d'évaluation, nous avons choisi de centrer notre expérience sur l'évaluation formative. On peut déduire de ce qui précède qu'il ne serait pas conséquent d'utiliser les tests que nous élaborons à des fins d'évaluation sommative, et encore moins pronostique.

III - Evaluation objectif par objectif

Avant de nous demander comment évaluer, il nous fallait savoir ce que nous voulions évaluer. Nous avons cherché, partant d'une part des programmes et commentaires officiels, d'autre part des opinions et habitudes pratiques des enseignants, à traduire les contenus de programmes en termes d'objectifs.

Par exemple, en classe de cinquième, le programme dit : "diviseurs d'un entier naturel", ce qui donne naissance à un objectif : "Etre capable de trouver ou de reconnaître un ou plusieurs diviseurs d'un naturel donné".

Un objectif étant ainsi défini, nous cherchons à l'opérationnaliser, c'est-à-dire à le traduire en termes de SAVOIR et de SAVOIR-FAIRE.

Par exemple, dans le cas ci-dessus, nous obtenons une liste du genre :

- Etant donné deux naturels a et b inférieurs à 100, tels que a soit inférieur à b, savoir dire si a est un diviseur de b.

- Etre capable d'écrire la liste des diviseurs d'un naturel inférieur à cent.
- Dès que a est supérieur à b (strictement), savoir dire que a n'est pas un diviseur de b.
- Etre capable de reconnaître une définition correcte parmi plusieurs définitions proposées.
- Etc...

Pour établir de telles listes, nous partons des propositions des collègues et des questions qu'ils ont l'habitude de poser. L'intersection des ensembles de MICRO-OBJECTIFS fournis par les uns et les autres nous donne ainsi une liste qui a quelque chance d'être reconnue comme minimum.

Nous avons choisi de faire porter notre évaluation sur les savoir et savoir-faire minima concernant des objectifs définis aussi clairement que possible.

Nous avons d'autre part, convenu de partager chaque programme en six ou sept classes d'objectifs liés par des relations de voisinage ou de dualité et, d'évaluer chaque classe d'objectifs séparément.

Exemple : Classe d'objectifs concernant "diviseurs, multiples, division euclidienne".

Nous dirons un mot plus loin des différentes TAXONOMIES permettant de classer les objectifs ou les exercices proposés. Nous avons pris comme base de réflexion, la classification N.L.S.M.A. et nous nous sommes ainsi intéressés aux objectifs du domaine cognitif et non (pas encore !) aux objectifs d'ordres affectifs ou psychomoteur malgré toute l'importance qu'ils revêtent et le développement des recherches dans ces domaines.

IV - Implications d'ordre pédagogique

De même que nous ne souhaitons pas intervenir sur les contenus, nous ne voulons pas nous immiscer dans les méthodes d'apprentissage utilisées. Tel d'entre nous utilise une méthode d'exposition linéaire, tel autre favorise plutôt l'activité de l'élève, a un enseignement plus éclaté, cherchant moins à cultiver la mémoire que l'esprit de recherche. Pourtant, l'un et l'autre pourront se retrouver au niveau des buts poursuivis, être en accord avec les objectifs proposés et leur traduction, et donc utiliser les mêmes instruments d'évaluation.

> Nous ne cherchons pas à évaluer des méthodes pédagogiques, encore moins la valeur des enseignants. Nous voulons seulement évaluer les connaissances et savoir-faire réels des élèves, et ce, indépendamment des moyens mis en oeuvre pour y accéder.

Il va de soi qu'une telle évaluation peut conduire chacun d'entre nous à des remises en questions et à une modification de nos pratiques pédagogiques ; cela se fera sous l'influence d'une meilleure connaissance des élèves et de leurs possibilités réelles.

Il convient ici de rappeler que des recherches, celles de REUCHLIN et de BACHER en particulier, ont montré que les professeurs ignoraient en général les possibilités de leurs élèves. Pour une question donnée, l'écart entre le taux de réussite attendu par les enseignants et le taux de réussite des élèves varie de 20 à 30 %, les enseignants surestimant systématiquement les capacités des élèves. Nous suggérons aux collègues qui auraient quelque doute concernant ces affirmations de faire le petit test suivant : lors du prochain devoir ou contrôle, après avoir choisi le sujet, estimer le taux de réussite des élèves, question par question. Il ne s'agit pas de prédire les notes qui seront attribuées, car consciemment ou non, elles prennent en compte le niveau général de réussite. Comparer ensuite les résultats prédits aux résultats réellement enregistrés.

On peut conjecturer qu'un léger décalage positif entre les attentes des enseignants et la réalité est pour l'élève une invitation à progresser (voir l'effet Pygmalion), un décalage trop important étant plutôt ressenti comme un abandon. Quoi qu'il en soit, il nous semble important de mieux connaître la réalité des possibilités, capacités, connaissances, savoir-faire, savoir-être de nos élèves, dans le but essentiel de mieux les aider.

Nous plaçant au niveau des savoir minima, il va de soi que nous ne voulons pas faire une évaluation élitiste. Nous souhaitons que pratiquement tous les élèves atteignent les objectifs fixés. Du moins construisons-nous nos instruments d'évaluation dans cet esprit ; nous ne faisons pas pour autant exception à la règle énoncée ci-dessus et le taux d'échec enregistré est toujours trop important à notre gré.

Nous nous plaçons ainsi dans l'optique d'une pédagogie de la réussite :

La réussite étant pour nous plus importante que l'échec, l'échec devrait être temporaire, analysé, pour faire l'objet de traitements appropriés et faire place en fin de compte à la réussite.

Dans ces conditions, nous construisons chaque test en postulant que :

- L'objectif correspondant n'est atteint par un élève que s'il a réussi à 70 % des questions posées. (Réussite au test).
- Dans l'ensemble, 70 % des élèves doivent réussir le test.

La courbe de GAUSS, courbe représentative de la loi normale est la courbe du hasard. Moyennant quelques précautions, dans une population donnée, les tailles des

individus adultes sont distribuées selon cette loi. Il est permis de penser que les aptitudes dans un domaine donné sont distribuées selon la loi normale. Cependant, pour donner un exemple, la courbe des salaires des français ne ressemble pas à la courbe de Gauss, elle a plutôt la forme d'un L. Il n'est pas raisonnable de penser qu'après un apprentissage, les savoir-faire des élèves sont distribués de façon tout à fait aléatoire. Certains auteurs ont pu parler, pour le dénoncer, du "mythe de la courbe de Gauss". C'est en effet ce modèle que nous avons généralement en tête lorsque nous évaluons.

Au modèle de la courbe de Gauss, nous voulons substituer celui de la courbe en J. Peu d'élèves échouent, beaucoup réussissent.

V - Prise en compte des théories et de l'état de la recherche

Les travaux de recherche concernant l'évaluation et la définition des objectifs se sont considérablement développés ces dernières années. En ce qui concerne les mathématiques, plusieurs I.R.E.M. se sont intéressés au sujet (Strasbourg, Toulouse, Rennes,...). Des thèses de doctorats y sont consacrées (celles de F. Pluvinage à Strasbourg, de R. Gras à Rennes), d'autres sont en préparation. Les recherches actuelles en didactique des mathématiques accordent une place privilégiée à l'évaluation.

Nous essayons dans la mesure de nos moyens, de prendre ces travaux en compte, de les intégrer à nos préoccupations, d'utiliser leurs conclusions. La tâche n'est pas facile, les domaines touchant de plus ou moins près à l'évaluation sont nombreux, citons en particulier :

- La psychologie
- Les sciences de l'éducation, en particulier la docimologie
- L'épistémologie génétique (travaux de J. Piaget)
- Les statistiques : statistique classique mais aussi des outils tels que l'analyse factorielle, l'analyse hiérarchique...
- La taxonomie

et bien sûr... la matière que l'on veut évaluer.

On ne peut prétendre faire une évaluation sérieuse sans avoir des connaissances dans chacun de ces domaines. C'est dire que l'élaboration d'un programme d'évaluation ne peut être l'oeuvre d'une personne isolée ; il s'agira nécessairement d'un travail d'équipe.

Outre des spécialistes de la discipline objet, il faudrait dans cette équipe, au moins un psychologue, un statisticien, un spécialiste des sciences de l'éducation. Nous n'avons malheureusement pas réussi à réunir une telle équi-

pe, mais c'est le souhait que nous formulons pour l'avenir, si du moins cette expérience doit, comme nous l'espérons se poursuivre au niveau des classes de quatrième et de troisième.

VI - Prise en compte de la réalité

Le travail effectué n'a pas vraiment le caractère d'une recherche, en ce sens que notre souci majeur n'a pas été de chercher à identifier les paramètres en présence, à les isoler, puis à formuler des hypothèses pour enfin mettre en place un dispositif expérimental destiné à les valider ou à les infirmer. Nos intentions étaient d'une certaine façon plus modestes ; nous voulions partir de nos pratiques et de l'état de nos connaissances, il s'agit donc d'une démarche pragmatique.

Ces intentions étaient en même temps plus ambitieuses, car nous désirions mettre au point un OUTIL qui soit recevable par nos collègues et immédiatement opérationnel. Certes, cette démarche est critiquable, mais nous avons d'illustres devanciers. Ce n'est autre en effet que celle suivie par BLOOM et ses collaborateurs.

Ne voulant pas nous placer dans une situation artificielle, nous avons été amenés à tenir compte des contraintes de tous ordres :

- Les contraintes de programme déjà citées
- D'autres contraintes institutionnelles, telles celles liées aux rythmes scolaires, à l'équipement des établissements concernant les moyens de reprographie...
- Celles liées à l'état des mentalités des enseignants, des élèves, des parents

VII - Méthodes de travail

L'équipe d'Ornans avait pris de l'avance en 78-79 en travaillant sur le programme de sixième. Peu à peu, un va et vient s'est instauré entre les deux équipes. La méthode de travail, plutôt fluctuante au début, tend à se stabiliser de la façon suivante :

- Lors d'une réunion, on choisit une classe d'objectifs que l'on traduit en termes de micro-objectifs, puis on élabore un projet de test les recouvrant.
- Ces documents sont repris par l'autre équipe, parfois par des collègues isolés, critiqués et amendés.
- On obtient une première version du test que l'on soumet alors à une classe. Les résultats obtenus sont étudiés item par item et les brouillons sont analysés ce qui permet de mieux repérer les difficultés ainsi que les erreurs dues à une formulation

ambigüe ou mal adaptée au niveau de compréhension des élèves. Dans certains cas, nous avons pu faire des passations individuelles, ce qui est à notre avis, la meilleure méthode, mais elle n'est pas facile à mettre en place.

- Au vu de ces premiers résultats, le test est remanié et l'on obtient une version définitive que l'on propose à tous les collègues intéressés.

Les tests de sixième et cinquième sont actuellement passés par plus de 2000 élèves répartis dans quelque 80 classes. Les résultats sont recueillis élève par élève et item par item, analysés à l'I.R.E.M. et fournissent un étalonnage du test. Analyse et étalonnage sont ensuite envoyés aux utilisateurs.

Chaque élève possède une "fiche individuelle d'évaluation" où son degré de réussite par rapport à chacun des objectifs est repéré par un code couleur. Des tests bis sont prévus que nous voulons rigoureusement isomorphes aux premiers, qui permettront à un élève n'ayant pas atteint un objectif d'être replacé dans les mêmes conditions ultérieurement. On pourra alors (et il pourra !) mesurer ses progrès par rapport à un objectif donné.

Par ailleurs, nous prévoyons aussi de construire des tests de perfectionnement, centrés sur les mêmes objectifs, et qui auraient pour objet de mettre davantage en jeu les niveaux supérieurs de la taxonomie (découverte, analyse, validation...). Mais nous n'avons guère avancé dans cette voie.

En ce qui concerne la construction des items, nous avons adopté plusieurs principes :

- La réussite ou l'échec à un item doit pouvoir être mis en relation avec l'objectif contrôlé, et uniquement avec lui (autant que possible). C'est-à-dire que nous cherchons à isoler au maximum les difficultés.

Par exemple, on évitera qu'une question concernant la compréhension d'un algorithme devienne en réalité pour l'élève une question sur la terminologie.

- Malgré cela, des dépendances continuent à exister entre les items ; nous essayons de les prévoir et nous chercherons à les vérifier en étudiant les corrélations.

- Nous utilisons aussi bien des questions ouvertes (l'élève doit construire sa réponse) que des questions fermées (la question commande la réponse, dans sa forme comme dans son fond). En ce qui concerne les questions fermées nous utilisons souvent des questions à choix multiples (Q.C.M.), l'élève devant alors choisir entre cinq ou six réponses proposées. Dans le meilleur des cas il aura à choisir parmi les 2^5 parties d'un ensemble à cinq éléments. La méthode utilisée nous permet de proposer des questions à alternative unique (VRAI - FAUX). On sait l'inconvénient habituel de cette méthode :

la réussite au hasard y est trop fréquente. De toutes façons, chaque fois que c'est possible, nous préférons ménager une troisième issue (par exemple : on ne peut pas savoir)

Dans la mesure où nous voulons que la personnalité du correcteur soit sans influence sur les résultats, ce qui est une condition essentielle, nous devons utiliser les questions ouvertes avec prudence.

Le plus souvent, nous regroupons plusieurs sous-questions dans un même item, la réussite à l'item n'étant acquise que si toutes les sous-réponses sont correctes. Il y a plusieurs raisons à cela :

- Aussi bien pour le correcteur que pour l'exploitation statistique, il nous a semblé qu'il n'était pas possible de dépasser vingt items par tests.
- Mais surtout, la réussite à une question isolée n'est pas toujours, loin de là, la preuve de la maitrise d'un micro-objectif. Voici un exemple :

  * On demande à un élève de cinquième de calculer : (-7) x (+3). Il répond correctement : -21
  * Plus loin, on lui demande de calculer : (-3) x (+5). Il répond alors : +15

  En réalité, cet élève croit que le signe du produit est le signe du nombre qui a la plus grande valeur absolue. Il n'a jamais qu'une chance sur deux de se tromper, et dans la notation traditionnelle peut espérer obtenir la moyenne, et être ainsi conforté dans son erreur. Nous pensons que ces deux question ne peuvent pas être séparées et qu'elles doivent prendre place dans un même item.

Nous avons appelé nos épreuves TESTS, d'une part par commodité, d'autre part parce que nous essayons de leur donner les qualités reconnues aux tests. Cette appellation est contestable, nombre de psychologues estiment qu'il faut deux années de travail pour construire un test. Rappelons que les qualités que l'on recherche dans les tests sont :

- LA VALIDITE : le test mesure-t-il bien ce qu'il est censé mesurer ?
- LA SENSIBILITE : quel est son pouvoir de discrimination ?
- LA FIDELITE : un même élève passant deux fois le même test, est-il assuré d'obtenir deux fois la même note ?

En ce qui concerne les deux premières qualités, nous n'avons pour l'instant, que des présomptions et il nous reste beaucoup à faire dans ce domaine. Pour le troisième point, il n'est bien entendu pas question de l'aborder en proposant deux fois le même test aux mêmes élèves, la première passation constituant elle-même un apprentissage, mais il y a des méthodes classiques basées sur les statistiques. C'est sans doute une

question que nous pourrons règler assez rapidement.

Pour ce qui est des modalités de passation, nous avons déjà dit que nous cherchions à être le moins contraignant possible. Nous demandons toutefois aux utilisateurs de respecter les consignes suivantes :

- Laisser aux élèves le temps qui leur est nécessaire, sans pour autant dépasser l'"heure de cours". Nous avons pu vérifier que le plus souvent les élèves les plus rapides avaient terminé au bout d'une demi-heure, les plus lents utilisant toute la séquence.
- Faire passer le test entre quinze jours et un mois après un éventuel apprentissage systématique. C'est donc la rétention à moyen terme qui nous intéresse et c'est avec cette idée que les tests sont construits. Malgré les apparences, ce point n'est pas en contradiction avec les conditions d'une évaluation formative.
- Nous demandons aux collègues de ne pas rendre les tests aux élèves, même pour la correction. Ceci est une condition nécessaire pour que les tests soient utilisables à des moments différents dans plusieurs classes d'un même établissement et éventuellement plusieurs années de suite. C'est aussi nécessaire si l'on veut que les tests bis ne s'éloignent pas trop des tests initiaux. Ce procédé a cependant quelques inconvénients, la copie étant traditionnellement un moyen de communication entre le professeur et les autres acteurs de l'action éducative. Nous nous demandons si en échange, il ne serait pas souhaitable de diffuser largement les listes de micro-objectifs, aux élèves (donc aux parents), à l'administration...

VIII - Conclusion

Nous avons essayé dans les lignes qui précèdent de préciser les choix que nous avons faits et la méthode utilisée. Au moment où nous écrivons, nous avons à peu près recouvert les programmes de sixième et cinquième. Nous ne pensons pas que ce travail soit parfait, bien des points doivent être revus, l'analyse statistique en est encore à l'état embryonnaire. La suite de ce travail et sa qualité dépendront beaucoup de l'accueil qu'il recevra parmi nos collègues et des échanges qui s'établiront entre nous.

Insistons encore sur deux points :

- Les batteries de tests proposées ne prétendent pas remplacer complètement les méthodes habituelles d'évaluation. Le devoir traditionnel garde sa valeur. Il sera simplement plus facile de l'utiliser pour ses qualités propres.
- Les tests proposés ainsi que les listes de micro-objectifs n'ont pas un caractère officiel. Il est possible d'y adhérer ou de les rejeter. Ils ne sont que l'expression des convictions d'un groupe d'enseignants qui cherche à être aussi repré-

sentatif que possible.

IX - Résumé

Nous voulons élaborer un OUTIL d'évaluation en mathématiques

- IMMEDIATEMENT UTILISABLE
- DANS LE BUT D'UNE EVALUATION FORMATIVE
- Portant sur LES SAVOIR ET SAVOIR-FAIRE MINIMA, A MOYEN TERME
- CONCERNANT DES OBJECTIFS CLAIREMENT DEFINIS
- TRADUIT EN TERMES DE MICRO-OBJECTIFS OPERATIONNELS
- Dans la perspective d'une PEDAGOGIE DE LA REUSSITE
- Sans imposer aux utilisateurs telle progression ou méthode d'enseignement
- En étant LE MOINS CONTRAIGNANT POSSIBLE quant aux modalités de passation des épreuves élaborées
- EN ASSOCIANT LES COLLEGUES, aussi étroitement que possible à la détermination des objectifs et à la construction des épreuves
- En cherchant à faire preuve du MAXIMUM DE RIGUEUR.

Le texte qui précède a été écrit en 1980 et a en quelque sorte servi de charte pour le travail qui a suivi. Trois ans plus tard, nous avons à peu près terminé l'opérationnalisation des objectifs de l'enseignement des mathématiques au collège. Tout du moins des objectifs de type cognitif qui semblent être habituellement poursuivis, qui sont de façon plus ou moins implicite, l'objet de l'évaluation traditionnelle et que les enseignants situent volontiers au niveau des savoirs minima. Peu à peu, nous nous sommes posés des questions concernant les objectifs de développement et d'expression. Il resterait beaucoup à faire en ce qui concerne certains objectifs difficilement opérationnalisables sous une forme épurée, c'est à dire non contaminée par d'autres objectifs : nous pensons en particulier aux objectifs qui visent surtout l'activité de l'élève indépendamment des savoirs mobilisés, aux objectifs concernant la communication, aux objectifs de transfert etc...
On peut penser qu'il s'agit là des objectifs essentiels, que les autres, ceux que nous avons évalué, ne sont souvent que des objectifs intermédiaires sans intérêt en soi. Il faut cependant constater qu'en pratique, ces objectifs essentiels ne sont évalués que par l'intermédiaire de savoirs et savoir-faire qui correspondent justement aux objectifs que nous avons essayé d'identifier. C'est d'abord l'absence de ces capacités qui est dans bien des cas la cause de l'échec scolaire, c'est leur présence qui dans l'état actuel des pratiques conditionne les possibilités d'accès aux capacités supérieures.

Une remarque nous semble fondamentale : au fur et à mesure que nous avons avancé dans notre travail, nous avons davantage pris en compte des objectifs mettant en jeu la compréhension, au détriment des automatismes et des savoirs spécifiques, tout en restant, à notre avis, au niveau des savoirs minima. Bien entendu, nos taux de réussite n'ont fait que diminuer, et peu à peu, nous sommes passés des courbes de la réussite aux courbes de l'échec. Nos savoirs minima se sont révélés être dans ce cas des maxima pour les élèves. Alors qu'une de nos ambitions était de participer à la lutte contre l'échec scolaire, nous nous sommes rendus compte qu'une évaluation rigoureuse pouvait avoir pour effet de mettre davantage en évidence les lacunes des élèves et en conséquence de renforcer l'échec. Est-ce à dire que nos objectifs ne sont pas raisonnables ? C'est possible pour certains d'entre eux, mais il serait trop facile de supprimer les objectifs encombrants et de nous limiter à ceux qui sont dans l'ensemble atteints par les élèves. Nous avons voulu faire un travail d'évaluation sans trop nous préoccuper des situations d'apprentissage ni même de la pertinence des objectifs par rapport à l'objet mathématique (pour simplifier, nous voulons dire que nous avons surtout recherché ce que les élèves devaient être capable de faire pour réussir dans

le système scolaire tel qu'il est), nous touchons sans doute aux limites d'une telle entreprise. Une approche différente s'appuyant à la fois sur une analyse de la nature de l'activité mathématique sous ses divers aspects et sur les connaissances actuelles en didactique des mathématiques s'impose maintenant à nous. Les outils d'évaluation qui pourraient en résulter ne manqueraient pas d'être en rupture avec les pratiques actuelles et risqueraient de n'être cohérentes ni avec les contenus actuels ni avec les méthodes d'enseignement. Quoi qu'il en soit, nous sommes conscients de n'avoir pris en compte qu'une partie de ce qui est évaluable, le système que nous proposons peut sembler être complet, en fait il ne l'est pas et il n'est sans doute pas souhaitable qu'il le soit. Citons A. de PERRETI qui, se référant à la loi dite "de Variété requise" énoncée par ASHBY pour la cybernétique et l'étude générale des systèmes, déclare :

> "Si la Variété des moyens et des processus de mesure et d'ajustement est insuffisante, à la place d'une régulation, on verra apparaître des mécanismes de réduction. La diffusion du savoir et des savoir-faire, l'entraînement des capacités seront réduits. Une structure élitique viendra contrarier des finalités démocratiques en introduisant des classements de ségrégation et en étouffant des possibilités d'orientation par des inerties de sélection. On peut dire également qu'une absence de variété favorise des blocages et des ruptures qui multiplient les effets pervers du système".
>
> Nous ne pouvons manquer de prendre en compte cet avertissement.

Remarques d'ordre pratique.

1) Les conditions de passation des tests telles qu'elles sont définies dans la première partie de ce texte ont en général été respectées pour la construction des étalonnages. Il va de soi que pour une utilisation libre, ces conditions peuvent être modifiées au gré de l'enseignant : par exemple, laisser à chaque élève le temps qui lui est nécessaire, faire passer l'épreuve en plusieurs fois, faire passer le test plusieurs mois après l'apprentissage etc... dans tous ces cas, l'étalonnage n'aura plus qu'une valeur indicative.

2) La consigne "ne pas rendre les tests" est devenue caduque. En effet, nous avons observé que les élèves souhaitaient conserver les épreuves, que celles ci leur servaient d'instruments de travail, de référence. De plus, la communication avec les familles s'en trouve améliorée. Le nombre d'épreuves en service, la possibilité d'utiliser en première passation les épreuves bis ou ter au lieu des épreuves initiales fait que les risques de préparation intensive et trompeuse d'une épreuve particulière sont très réduit. Cela reste vrai lorsque les épreuves sont passées à des moments différents dans plusieurs classes du même établissement.

3) Nous n'avons publié que quelques épreuves bis, beaucoup de collègues en ont construits en ne modifiant que légèrement l'épreuve initiale. Nous avons en effet abandonné l'idée de faire des tests isomorphes qui soient apparamment très différents des tests initiaux. Pour l'élève en difficulté, la moindre différence est une différence importante. Les élèves réalisent mieux leurs progrès et le chemin qui leur reste à faire lorsqu'ils sont soumis plusieurs fois à des épreuves semblables. Le risque de "bachotage" existe, mais d'une part un certain type d'entraînement peut être utile aux élèves et d'autre part ce risque est réduit par la variété des épreuves disponibles et par le type de questionnement utilisé.

D E U X I E M E P A R T I E

Développement de certaines des idées présentées dans la première partie.

I) LA DOCIMOLOGIE CLASSIQUE

Au cours de nos stages, nous avons fait une place jugée parfois trop important à la docimologie. Certains collègues pensent que la critique docimologique est bien connue et que ses conclusions sont maintenant acceptées et intégrées par l'ensemble des enseignants. L'important serait alors de promouvoir des procédures d'évaluation qui échappent à cette critique. L'évaluation englobe et dépasse largement le problème des notes, mais les notes elles-mêmes jouent un rôle considérable dans les processus d'orientation et d'éviction. Que ce soit dans les conseils de classe, les jurys d'examens, les commissions diverses, les rencontres avec les parents, elles sont omniprésentes et bénéficient largement de l'autorité de la chose jugée. Il n'est donc pas tout à fait certain que la cause soit entendue.

DEFINITIONS et REMARQUES

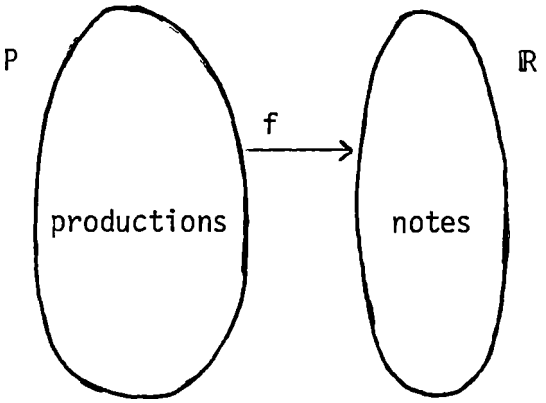
Nous emprunterons d'abord à de LANDSHEERE cette définition : "La docimologie est une science qui a pour objet l'étude systématique des examens, en particulier des systèmes de notation, et du comportement des examinateurs et des examinés". Nous étendrons simplement cette définition à d'autres situations que celle des examens. La docimologie peut être invoquée chaque fois qu'une production est notée et que cette note, échappant au couple évaluateur-évalué est utilisée par le système pour porter un jugement sur l'élève.

L'agacement provoqué par la référence à la docimologie est d'autant plus compréhensible que l'on s'est souvent contenté d'en rapporter les constats et qu'elle apparait alors essentiellement comme une mise en cause des pratiques des enseignants. "Les préoccupations qui l'ont inspirée se situent à peu près toutes dans le domaine de la CRITIQUE des systèmes actuels de notation pédagogique en honneur dans les différents examens et concours"(Guillaumin). Au delà des améliorations de détail qu'elle propose, elle incite en fait à une "modification profonde, à préciser des attitudes du corps enseignant à l'égard de la valeur en quelque sorte sacrée, accordée de manière générale à des chiffres établis dans l'équivoque" (Guillaumin).

LES ENSEIGNEMENTS DE LA DOCIMOLOGIE

Nous ne parlerons pas ici des méthodes utilisées (voir plus loin l'article sur le comportement de l'évaluateur, celui relatant une expérience de docimologie faite à l'I.R.E.M. ainsi que la fiche de travaux pratiques de J. KAUFMANN), nous nous contenterons de rappeler quelques résultats.

Etant donné un ensemble de productions (écrites ou orales) que nous appelerons P, NOTER c'est effectuer une application de P dans une partie de R (ensemble de notes). Soit f cette application.



Il convient de remarquer que dans la pratique on parle de NOTER LES ELEVES alors qu'en fait on ne note que leurs productions.

Xo désignant un élément de P, f(Xo) dépend certes de Xo, mais aussi du correcteur.

A chaque correcteur e correspond ainsi une application f_e.

Chaque correcteur se caractérise par un couple $(\bar{m},\sigma)$ que l'on appellera son EQUATION PERSONNELLE. $\bar{m}$ est la moyenne des notes attribuées par e pour l'ensemble des productions auxquelles il est susceptible d'être confronté. σ mesure la dispersion de ses notes autour de $\bar{m}$. Ainsi chaque correcteur place sa moyenne plus ou moins haut, et pour deux correcteurs ayant la même moyenne, les notes peuvent être dispersées de façon très différentes autour de cette moyenne. $\bar{m}$ et σ dépendent de la personnalité du correcteur, de son statut, de son âge, etc... Bien sûr un correcteur distinguera un ensemble P de productions de bonne qualité d'un ensemble P' de productions médiocres, les moyennes attribuées à P et à P' ne seront pas les mêmes, mais toutes deux sont attirées vers la moyenne personnelle.

Par ailleurs f_e (Xo) dépend de P tout entier et plus particulièrement de l'ordre dans lequel les éléments de P sont examinés.

On pourrait penser qu'il suffise de connaître les équations personnelles des correcteurs pour rétablir l'équité qui vient d'être mise en défaut. Malheureusement, le classement des productions dépend aussi du correcteur et il est clair qu'aucune translation ou transformation affine des notes ne pourra remédier à ce défaut.

Après une enquête sur le baccalauréat, H. PIERON pouvait conclure : "pour prédire la note d'un candidat, il vaut mieux connaître son examinateur que lui-même."

S'il s'agit de NOTER UN ELEVE et non plus seulement un produit, le problème se complique encore. En effet la note dépend alors de la tâche. Un élève sera plus à l'aise, manifestera plus facilement ses capacités devant une épreuve que devant une autre. Les enseignants le savent bien qui parlent de devoir facile ou de devoir difficile, mais ils savent aussi que pour deux devoirs différents les élèves ne se classent pas de la même façon.

La note dépend aussi de l'élève, et c'est heureux ! seulement c'est l'élève lui-même qui est fluctuant dans ses capacités. Au mieux on note un moment de l'élève, la permanence est plus difficile à saisir.

LES SOLUTIONS PROPOSEES PAR LA DOCIMOLOGIE

Les procédures proposées, appelées procédures de MODERATION visent en premier lieu à réduire l'effet de l'équation personnelle des correcteurs. Une transformation affine convenable ramenera chacun au même couple $(\bar{m}, \sigma)$. Nous avons vu plus haut que le classement lui-même dépendait du correcteur, cet aménagement ne serait donc pas satisfaisant. Il est cependant utilisé dans un certain nombre d'examen mais est rarement bien accepté par les correcteurs, avec raison d'ailleurs, en effet ce que l'on connait en général c'est le couple $(\bar{m}_p, \sigma_p)$ d'un correcteur pour un ensemble P donné. Ce correcteur peut légitimement penser que cet ensemble P n'est pas représentatif de l'ensemble des élèves ou des candidats. En tout état de cause, le couple (m, σ) réel du correcteur demeure inconnu et ne pourrait être approché que par le collationnement de toutes les notes qu'il a attribué depuis son entrée en fonction.

Une autre procédure qui peut d'ailleurs être composée avec la précédente concerne la multiplication des juges. Cette procédure est d'autant plus coûteuse que l'on a calculé qu'il fallait envisager plusieurs dizaines de correcteurs indépendants pour obtenir une note stabilisée. Dans certains cas les doubles ou triple corrections constituent cependant des palliatifs intéressants, surtout si les correcteurs conservent leur indépendance et ne s'influencent pas mutuellement. Mais c'est peut-être accorder trop d'importance à une production particulière en la considérant implicitement comme totalement représentative du sujet et on peut se demander si trois épreuves différentes notées par trois correcteurs différents ne seraient pas préférables à une épreuve unique soumise à trois personnes distinctes.

Enfin l'établissement systématique de barèmes précis permet de diminuer considérablement les divergences entre correcteurs, mais non de les annuler. Cette méthode suppose cependant des pondérations nécessairement arbitraires et empêche de valoriser des qualités particulières qui pourraient apparaître à l'intérieur de la copie.

Pour parvenir à une évaluation plus objective, la docimologie a aussi proposé de modifier la forme des épreuves. Ainsi les épreuves (ou TESTS) standardisées, les Q.C.M. (questionnaires à choix multiple), Q.R.O.C. (questionnaires et réponses ouvertes mais courtes), conduisent en général à des notes qui ne dépendent plus du correcteur. L'inconvénient est que bien souvent dans ces cas l'objet de l'évaluation s'est déplacé, on n'évalue plus la même chose : là où l'on cherchait à évaluer la capacité à construire une argumentation on n'évalue plus que la capacité à valider un résultat, là où l'on souhaitait évaluer la compréhension on n'évalue que des savoirs spécifiques mémorisés ou des automatismes. Nous verrons à propos de l'évaluation en général et de la question de l'opérationnalisation des objectifs comment éviter ces écueils.

Une autre proposition issue de la docimologie suggère de remplacer l'évaluation ponctuelle par l'EVALUATION CONTINUE. On en voit bien les avantages mais elle ne résoud pas tous les problèmes et en crée même de nouveaux. Dans la pratique où l'enseignant est maître à la fois des méthodes pédagogiques, du choix des épreuves d'évaluation et des barèmes de

correction adoptés il peut même y avoir en fin de compte détérioration par rapport à la situation d'examen.

LES ECHELLES DE MESURE

Les notes sont des nombres, c'est sans doute la raison pour laquelle il est si facile de les faire passer pour des mesures. Nous rappelerons (au moins aux scientifiques) que les moyennes de notes sont en fait des barycentres et que comme telles elles supposent une structure d'espace vectoriel.

Plus précisément, on peut distinguer :

1° L'ECHELLE ORDINALE

Dans bien des cas, avant de décider d'une note, on compare la production avec d'autres provenant d'élèves différents ou du même élève. Ce qui semble important dans ce cas c'est d'abord que l'ordre de classement des copies soit correct : "j'ai mis 15 à x, je vais devoir mettre davantage à y, sa copie est meilleure". Les seules transformations conservant la structure d'ordre étant les applications monotones croissantes, ce sont aussi les seules qui soient admissibles pour des notes mises dans ces conditions. La notion de distance d'une note à l'autre ainsi que celle de moyenne sont alors dépourvues de sens.

2° L'ECHELLE D'INTERVALLES (ou échelle affine)

Si l'on peut affirmer que la différence est la même entre 5 et 10 d'une part et entre 7 et 12 d'autre part (plus généralement entre a et a + c d'une part et entre b et b + c d'autre part, pour toutes valeurs possibles de a, b et c), alors on a affaire à une échelle d'intervalles. Il est clair que ni le correcteur, ni l'élève ne considèrent comme équivalents le passage de 2 à 4 et le passage de 9 à 11. Même en admettant qu'une échelle de notation soit une échelle d'intervalle, le zéro n'y serait pas défini et les seules transformations admissibles seraient

celles qui conservent les rapports d'intervalles, c'est-à-dire les applications affines.

3° L'ECHELLE DE RAPPORTS (espace vectoriel)

Si l'on peut affirmer que le rapport des notes a un sens, que 10 est le double de 5, que 16 vaut quatre fois 4... alors c'est que le zéro a aussi une signification intrinsèque. Dans ce cas et dans ce cas seulement on peut légitimement calculer des moyennes. Le seul cas sans doute qui satisfasse à ces conditions est celui des épreuves type Q.C.M. où la note est en réalité un SCORE (nombre de questions réussies) cela suppose bien sûr que les questions ne soient pas pondérées. En général, on est loin de ces conditions et il est tout à fait surprenant de voir des enseignants (même scientifiques !) calculer leurs moyennes trimestrielles au $1/_{100}$ près alors que leurs notes elles-mêmes sont données en points entiers. "Une précision sur un résultat, lorsqu'elle dépasse la précision sur les données expérimentales est très exactement la détermination du néant" (Bachelard).

Aux problèmes soulevés par la docimologie proprement dite s'ajoutent donc ceux relatifs à la réduction et la communication des résultats. Après un examen ou un contrôle continu, au moment d'un conseil de classe chaque élève pourrait se caractériser par une liste impressionnante de données, numériques ou non, qui devraient comprendre non seulement l'ensemble des notes attribuées, mais les épreuves auxquelles il a été soumis et les évaluateurs qui ont attribué ces notes. Toute méthode de réduction des données conduit à une perte d'information, le passage traditionnel à la simple moyenne est le moyen le plus sur d'en perdre le maximum possible.

LE MYTHE DE LA VRAIE NOTE

La réflexion docimologique repose en fait sur plusieurs postulats implicites.

Le premier est qu'un individu peut être valablement représenté par une de ses productions.

Le second est que toutes production a une valeur bien déterminée et qu'il convient de s'efforcer d'approcher au plus près cette vraie valeur. Le paragraphe précédent a sans doute montré que la Vraie note n'existait pas.

Le troisième est que les discordances entre évaluateurs sont dues au hasard. Les recherches des psychologues de l'évaluation ont montré que tel n'était pas le cas, les différences sont au contraire systématiques (voir plus loin "le comportement de l'évaluateur"). Cette dernière remarque prend une importance particulière dans le cas de l'évaluation continue. Les mêmes causes produisent toujours le même effet sur un évaluateur particulier mais produisent des effets différents sur des évaluateurs différents. De plus, l'évaluateur unique est conforté dans ses jugements par... ses propres jugements, son évaluation tend ainsi à devenir de plus en plus personnelle et à s'éloigner de plus en plus de celle que font ses collègues.

Au moment où l'évaluation continue se généralise, où l'on cherche à promouvoir des formes nouvelles d'évaluation, il peut sembler que la critique docimologique soit périmée. Nous pensons au contraire que son enseignement reste d'actualité et qu'il convient de le prendre en compte.

II) L'EVALUATION

Parmi les fonctions de l'enseignant, l'évaluation occupe une position privilégiée, elle permet à la fois de contrôler l'efficacité de l'action et de prévoir les formes qu'elle devra prendre.

L'évaluation a toujours été une préoccupation des enseignants, l'enseignement est un acte de communication, et dans toute communication l'émetteur a besoin de savoir si le récepteur a bien reçu le message. Dans un enseignement de type humaniste, destiné à une élite, où les valeurs à intégrer font l'objet d'un large consensus social et culturel, l'évaluation pose peu de problèmes. Dans l'enseignement de masse qui est le notre, et alors que les valeurs à transmettre ne font plus l'unanimité, les choses se compliquent. Il s'agit moins de contrôler l'acquisition d'une culture mal définie, que de permettre et de contrôler des acquisitions de savoir, savoir-faire, savoir-être dont l'intérêt résiste à l'analyse.

L'évaluation n'est pas une idée nouvelle, mais la réflexion qu'elle doit inspirer doit être actualisée pour prendre en compte les finalités actuelles de l'enseignement.

Voici d'abord deux définitions :

> "Dans son acception la plus large, le terme évaluation désigne l'acte par lequel, à propos d'un événement, d'un individu ou d'un objet, on émet un jugement en se référant à un (ou plusieurs) critère(s), quels que soient par ailleurs ce(s) critère(s) et l'objet du jugement."
>
> (NOIZET et CAVERNI)

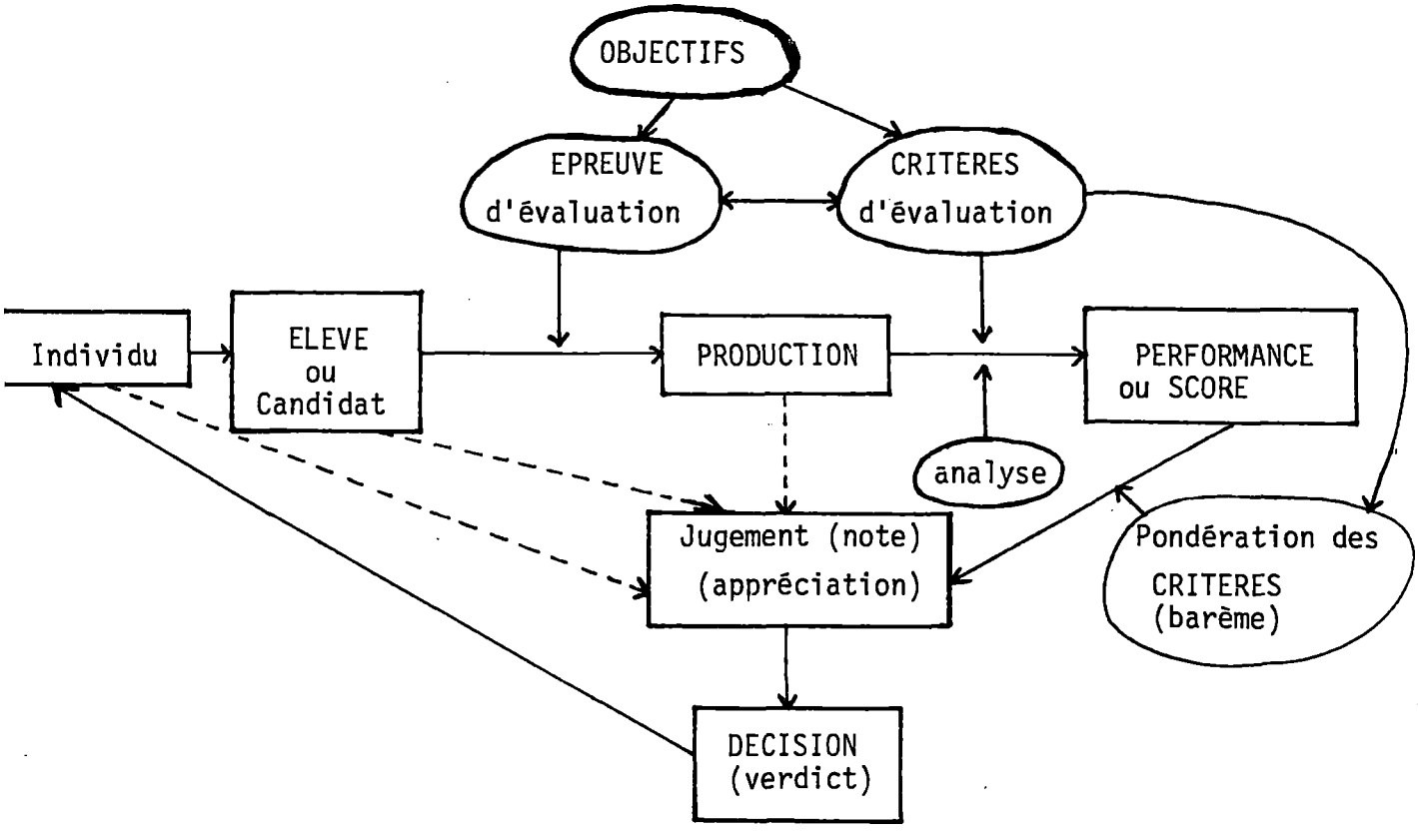
> "Evaluer signifie examiner le degré d'adéquation entre un ensemble d'informations et un ensemble de critères adéquats à l'objectif fixé, en vue de prendre une décision".
>
> (J.M. de KETELE)

Le mot "note" n'est pas cité, une évaluation conduit donc à un jugement qui peut ou non, selon les besoins, se traduire par une note. L'amalgame souvent fait entre noter, corriger et évaluer est source de bien des confusions.

Corriger un devoir ou une production c'est le rétablir dans sa cohérence, c'est le transformer pour qu'il puisse devenir une référence, un outil de travail pour l'élève. Corriger n'est pas nécessairement une action d'évaluation, cependant on en profite souvent pour procéder à une analyse de la production qui conduira à en dégager la performance.

La performance est un résumé significatif de la production, celle ci a été confrontée aux critères implicites ou explicites de l'évaluateur qui en retient alors les indices sur lesquels portera réellement son évaluation. Ce sera le nombre de fautes dans une dictée, la rigueur du raisonnement, le nombre et la nature des erreurs de calcul dans un travail de mathématiques, etc... L'intervention d'une pondération des critères qui se traduit souvent par un barème conduira alors à un jugement, éventuellement une note.

On peut ainsi proposer le schéma ci-dessous. Les filtres successifs y sont mis en évidence ; ainsi il est important de savoir si l'on note l'élève, l'individu ou une production. Selon que l'explicitation des critères aura lieu avant ou après l'épreuve (si toutefois elle a lieu) l'attitude de l'élève face à l'évaluation ainsi que le rapport maître-élève seront différents.



EVALUER, POURQUOI ?

L'école n'est pas faite pour l'école, la société a des droits sur elle, elle exprime ses demandes, parfois ses exigences par la voix du législateur puis par les instructions et programmes officiels. Les finalités, buts et objectifs qui nous sont ainsi assignés ne peuvent être atteints par la seule magie du verbe. Une régulation est nécessaire. L'évaluation permet à l'enseignant de contrôler son action et de procéder aux adaptations nécessaires. Cette évaluation peut être individuelle ou collective, mais même si elle est individuelle (de l'élève) elle ne peut manquer de renseigner le maître sur l'efficacité de son action.

Il n'est pas question de dire ici que les programmes sont toujours bons et qu'il ne faut jamais les remettre en question. Au contraire, une évaluation rigoureuse peut être à l'origine de certaines remises en cause. Il ne faut d'ailleurs pas confondre ce qui tient aux programmes, auxquels il est souvent utile de remonter, et ce qui est de l'ordre de la coutume ou des interprétations faites par les manuels.

Mais il y a bien d'autres raisons d'évaluer : on évalue pour motiver, pour diagnostiquer les causes de difficulté, pour certifier une formation, pour orienter, pour sélectionner... Les fonctions de l'évaluation sont donc diverses et elles ne devraient pas être confondues.

Il convient de distinguer :

L'EVALUATION FORMATIVE. C'est celle qui intervient en cours d'apprentissage avec l'intention de rétroagir sur la formation. Elle porte sur des savoir en cours d'élaboration et à ce titre valorise la rapidité d'acquisition au détriment de la qualité de la maturation. Tel élève apprend vite mais oublie tout aussi vite, tel autre apprend plus lentement mais ses savoirs sont durables. C'est dire que l'évaluation formative n'est pas adaptée à l'évaluation-bilan, elle est au service de l'élève, de ses apprentissages, elle devrait amener l'enseignant à différencier son action mais son but est d'homogénéiser et non de différencier les élèves. Il s'agit de situer les élèves par rapport à des objectifs, de repérer les insuffisances éventuelles, de chercher à remédier à ces insuffisances puis d'évaluer à nouveau.

C'est dire qu'une évaluation formative efface l'autre et que le cumul des résultats de ce type d'évaluation est un non sens pédagogique.

Donnons un exemple :

Après une séquence d'apprentissage portant sur l'addition des nombres rationnels, on demande aux élèves de calculer quelques sommes de fractions :

$$\frac{3}{7} + \frac{5}{3} \quad \text{etc...}$$

Certains élèves écrirons $\frac{8}{10}$ $\left(\frac{3+5}{7+3}\right)$

Si l'on note l'épreuve, la seule note acceptable serait 0, toute autre note risquant de les conforter dans des représentations érronées. Plus tard, après des actions de remédiation, ces élèves seront soumis à une épreuve de même type, que certains réussiront parfaitement. Il faudrait mettre 20 à cette épreuve. Mais quelle est alors la valeur de ces élèves : 0, 20, ou 10 ?

On sait pertinemment que si l'on répète cette même évaluation quelques mois plus tard, certains de ceux qui ont réussi du premier coup seront alors en situation d'échec tandis que certains de ceux qui n'ont réussi qu'au second essai réussiront encore au moment de l'évaluation différée. Pourquoi dans ce cas ne pas se contenter de l'information : "après apprentissage l'élève a montré qu'il savait additionner deux rationnels".

Certes, les habitudes sont telles, aussi bien chez les enseignants que chez les élèves qui attendent des notes et bien souvent se soucient plus d'elles que de leurs apprentissages, que le passage à une véritable évaluation formative ne pourra pas se faire facilement et que des étapes seraient sans doute à prévoir. D'autre part, dans la mesure où l'évaluation formative conduit naturellement à l'évaluation différenciée, des instruments d'évaluation adaptés, nombreux et variés, ainsi que des documents permettant le recueil des informations sont indispensables. En mathématiques le groupe de recherche de l'I.R.E.M. a essayé de travailler dans ce sens.

Dans le meilleur des cas, l'évaluation formative devient DIAGNOSTIQUE. Elle permet alors non seulement de localiser les difficultés des élèves mais elle permet aussi de repérer les causes de ces difficultés et de prévoir les remèdes spécifiques à apporter. Ainsi, une épreuve portant sur les aires peut permettre de diagnostiquer que certains élèves n'ont pas acquis le concept de surface, qu'ils ne peuvent pas concevoir une surface comme engendrée par le déplacement d'une ligne et donc qu'ils ne pourront pas résoudre certains exercices. Il devient donc inutile de les placer dans des situations où ils ne pourront qu'être en échec et il faudra leur proposer des situations adaptées qui leur permettra d'améliorer leurs représentations. L'ambition de l'évaluation formative est d'être diagnostique mais il ne faut pas se dissimuler la difficulté. Il est facile de constater qu'un élève ne sait pas, il est beaucoup plus difficile de comprendre pourquoi il ne sait pas.

Dans le contexte qui vient d'être défini, l'ERREUR n'est plus la FAUTE, ce n'est pas non plus le MAL. L'ERREUR jouit d'un statut privilégié, celui de révélateur des difficultés de l'enfant, des obstacles qu'il rencontre dans la structuration de son savoir. Plus que le SUCCES, l'erreur permet d'avoir accès aux représentations de l'enfant et de déceler les représentations erronées. L'ERREUR doit être prise en compte, analysée. Des recherches en didactique des mathématiques ont prouvé que certaines erreurs étaient le signe de perturbations annonciatrices de la proximité du passage à un niveau supérieur d'abstraction. Les élèves qui se trouvent encore loin de ce seuil ne connaissent pas ce conflit et, en utilisant leurs représentations du moment, solidement installées, ne commettent pas l'erreur.

> "En fait, on connait contre une connaissance antérieure, en détruisant des connaissances mal faits, en surmontant ce qui, dans l'esprit même, fait obstacle à la spiritualisation" (Bachelard, mais on pourrait aussi bien citer J. PIAGET).

L'EVALUATION PREDICTIVE OU PRONOSTIQUE est tant qu'à elle destinée "à prédire une performance dans une activité donnée ou à déterminer l'aptitude à réaliser certains apprentissages" (de Landsheere).
Il ne s'agit pas de faire le bilan des acquisitions de l'élève, il s'agit d'essayer de prévoir les acquisitions qu'il sera susceptible de faire. Cette évaluation porte donc sur des APTITUDES, sa fonction est d'ORIENTER.

Dans certains conseils de classe, il arrive que l'on fasse référence au MERITE de l'élève pour justifier des décisions d'orientation (nous entendons orientation au sens large, incluant les admissions en classe supérieure) : "il n'a pas assez travaillé, il ne mérite pas de passer en 4ème", "elle ne s'intéresse pas, elle n'a pas profité de son année scolaire, elle mérite de redoubler"... est-on bien certain que ces errements du passé sont des prédicteurs valables de l'avenir ? Les résultats scolaires traditionnels sont-ils eux mêmes de bons prédicteurs ? Nous renvoyons le lecteur à l'article "évaluation et liaisons 3ème - seconde" où l'on montre le peu de corrélation existant entre les résultats scolaires de 3ème et ceux de seconde. Certains tests utilisés par les psychologues et ne portant pas sur des acquisitions scolaires ont peut-être une valeur prédictive plus importante. Nous renvoyons à ce sujet à l'étude de J. KAUFMANN (voir bibliographie).

Avant une formation, il faut se demander quels sont les savoirs, les capacités que l'élève doit déjà posséder pour tirer profit du nouvel enseignement (pour pouvoir "suivre").
La reconnaissance de ces PREREQUIS suppose une analyse conjointe des objectifs, des méthodes pédagogiques utilisées et des contraintes didactiques. L'évaluation de ces prérequis procède à la fois de l'évaluation formative et de l'évaluation prédictive. Cette évaluation peut amener l'enseignant à repenser sa stratégie de formation, elle peut aussi conduire à des groupements différents à l'intérieur de la classe ou après brassage inter-classes (groupes de niveau par exemple).

Une pratique assez courante consiste à évaluer au début d'une année scolaire sur les acquisitions de l'année (voire du cycle) précédent . Cette évaluation qui mesure surtout l'étendue des oublis et des insuffisances permet le plus souvent de se conforter dans l'idée que les élèves "ne sont pas au niveau", elle permet de se rassurer en pensant que nous ne serons pas responsables des échecs inévitables, elle correspond à un "état des lieux" qui ne tiendrait pas compte de la nature des besoins. L'idée de l'évaluation des prérequis est tout autre, il s'agit de savoir pour agir et non de savoir pour juger. La reconnaissance des prérequis n'est cependant pas chose aisée. Quelles sont par exemple les conditions préalables à la compréhension d'un raisonnement hypothético-déductif ? Ceux nécessaires aux

acquisitions concernant les mesures de volumes, s'agit-il seulement de problèmes de multiplication ?

Même dans le cas de l'évaluation des prérequis, l'évaluation prédictive ne peut être mêlée à l'évaluation formative. Il est certain que l'on peut extraire de l'évaluation formative des indications ayant valeur prédictive, mais il y aurait alors lieu de leur donner plus ou moins d'importance selon l'orientation envisagée. D'autre part la confusion entre les deux types d'évaluations fausse la relation maître-élève, le maître apparaît toujours comme juge, comme censeur, et non comme guide, il devient alors naturel de tricher, d'essayer de faire croire que l'on sait, que l'on a compris alors que l'on a rien compris, de faire semblant. L'élève dans ce cas finit par se tromper lui-même, plus sûrement d'ailleurs qu'il ne trompe le maître, et cela au détriment de ses apprentissages.

L'EVALUATION SOMMATIVE est celle qui intervient à la fin d'une formation, c'est une évaluation bilan. Son rôle est de certifier une formation.

> "Alors que l'évaluation formative revêt, en principe, un caractère privé (sorte de dialogue particulier entre l'éducateur et son élève), l'évaluation sommative est publique : classement éventuel des élèves entre eux, communication des résultats aux parents par un bulletin scolaire, attribution d'un certificat ou d'un diplôme... (d'après BLOOM)."

La plupart des auteurs pensent que l'évaluation sommative ne devrait en aucun cas être contaminée par l'évaluation formative. Dans un contexte d'évaluation continue, il est clair que cela est difficilement réalisable, cela supposerait le rétablissement de véritables examens. Au moins conviendrait-il de ne retenir pour cette évaluation que des informations pertinentes et donc de rejeter celles qui auraient été manifestement recueillies en cours d'apprentissage et qui ne seraient pas significatives de savoirs établis. Que dire par exemple de la pratique qui consiste à prendre en compte pour l'évaluation sommative des notes de leçons dont on est bien conscient qu'elles rendent surtout compte de l'effort que l'élève à un moment donné a ou non accepté de fournir.

Parmi les raisons importantes qu'il y aurait de distinguer soigneusement l'évaluation formative de l'évaluation sommative, il y a celle qui est attachée à la notion de prise de risque. L'évaluateur qui doit décider que tel objectif est ou non atteint est soumis à deux risques contradictoires, le premier consiste à déclarer que l'élève sait alors qu'il ne sait pas (risque α), le second consiste à déclarer qu'il ne sait pas alors qu'il sait (risque β).
Pour des raisons d'équité, il est évident qu'il convient de minimiser le risque β dans le cas de l'évaluation sommative, alors que pour des raisons d'efficacité, c'est le risque α qu'il faut minimiser dans le cas de l'évaluation formative. Ceci entraîne que le questionnement ne peut être le même dans les deux cas.

QUI ou QUOI EVALUER.

Lorsque l'on veut évaluer la QUALITE d'une fabrication (cas d'une production industrielle par exemple) ou celle d'une entreprise prestatrice de services, il convient de diversifier l'observation, de ne pas se contenter de l'examen d'un seul produit. On sera amené à ECHANTILLONNER la production ou la prestation de façon à obtenir un sous-ensemble statistiquement représentatif. Les critères d'examen du produit devront rendre compte aussi complètement que possible de ce que nous entendons par QUALITE, si bien que la qualité elle-même pourra être définie par l'ensemble de ces critères. Prendre en compte l'esthétique de l'emballage par exemple, risque de fausser considérablement l'évaluation.

Le plus souvent, l'objet de l'évaluation pédagogique est l'ELEVE, il conviendra toutefois de ne pas perdre de vue que nous n'évaluons en fait que des produits (travaux écrits, prestations orales...), et que ces produits ne peuvent prétendre représenter l'élève que s'ils sont suffisamment diversifiés, s'ils constituent en quelque sorte un échantillonnage de l'ensemble de toutes les productions possibles de l'élève. S'il s'agit d'évaluer un NIVEAU en mathématiques, les critères retenus devront être pertinents et en tout état de cause c'est cet ensemble de critères qui définiront le niveau. Ainsi, prendre en compte la présentation, l'écriture, l'orthographe, la qualité de la rédaction, quand ce n'est la discipline dont fait preuve l'élève, c'est admettre implicitement que ces éléments participent d'une

façon ou d'une autre à la définition du niveau mathématique.

En évaluant l'élève, l'enseignant sait très bien qu'il s'évalue lui-même, il est directement impliqué, ce qui explique le coté souvent passionné, parfois dramatisé de l'évaluation. L'objet de l'évaluation est ainsi l'ENSEIGNANT qui s'évalue à travers les productions de ses élèves (AUTO EVALUATION). Pour dédramatiser cet acte on pourrait suggérer que l'enseignant évalue plutôt sa démarche pédagogique ou même l'adaptation d'une démarche pédagogique à un groupe d'élèves donné. Dans ce cas, l'EVALUATION COLLECTIVE anonyme peut être préférée à l'évaluation individuelle. C'est alors le GROUPE qui est l'objet de l'évaluation et le rapport maître-élève peut s'en trouver amélioré. On peut donc évaluer une méthode pédagogique, mais on peut aussi évaluer des objectifs. Par exemple des objectifs peuvent être considérés à priori comme intermédiaires, comme conditionnants l'accès à d'autres objectifs. Ce n'est qu'après évaluation que l'on saura si cette hypothèse est vérifiée.

L'évaluation dont nous parlons n'est pas nécessairement disciplinaire, quel que soit le mode de l'action, quelles que soient les intentions affirmées, l'évaluation est nécessaire. Par exemple, la mise en place d'un projet d'établissement devrait être précédée de l'évaluation des besoins et suivie de l'évaluation continue, puis à terme, de ses effets.

Prenons un exemple : un projet d'établissement est bâti autour de l'idée qui faut accroître l'intérêt scolaire des élèves (si l'on préfère, lutter contre le désintérêt scolaire) ce projet est nécessairement le résultat d'une évaluation préalable, l'objectif assigné ne serait pas pertinent dans n'importe quel établissement. Le problème des moyens et méthodes à mettre en place ne concerne pas directement l'évaluation, mais en fin de compte il faudra évaluer, c'est-à-dire savoir si l'intérêt a augmenté. Il est clair qu'une évaluation portant uniquement sur les savoirs n'est pas apte à renseigner sur l'intérêt. Aussi bien aux niveaux disciplinaires qu'au niveau transdisciplinaire il faudra constuire et utiliser des outils d'évaluation adaptés à cet objectif.

QUI EVALUE

Il faut d'abord distinguer l'évaluation INTERNE et l'évaluation EXTERNE.

Dans le cas de l'EVALUATION EXTERNE, c'est une instance extérieure qui évalue. En toute rigueur, cette instance ne devrait pas avoir de liens institutionnelle avec l'objet de l'évaluation. C'est le cas par exemple d'une commission de l'Assemblée Nationale chargée de l'évaluation d'une partie du système éducatif ou celui d'une équipe de sociologues ne dépendant pas de l'Education Nationale et effectuant un travail d'évaluation pour le compte de celle-ci. C'est sans doute aussi le cas de l'évaluation qui peut être faite par les collectivités locales (municipalités...) ou par les parents.

On admet en général qu'il y a aussi évaluation externe lorsqu'un examen est pratiqué par des enseignants qui n'ont pas eu auparavant de rapports avec le candidat (cas des jurys de brevet des collèges et baccalauréats par exemple), en réalité l'évaluation externe se fera vraiment lorsque l'élève ira monnayer son diplôme dans le monde du travail.

En ce qui concerne l'EVALUATION INTERNE, nous distinguerons l'auto-évaluation et l'hétéro-évaluation. Il y a AUTO-EVALUATION lorsque l'élève s'évalue lui-même, en général en utilisant des instruments d'évaluation fournis par le maître. De même lorsque l'enseignant s'évalue ou évalue sa démarche pédagogique. Il y a HETERO-EVALUATION lorsque le maître évalue l'élève mais aussi lorsque l'élève évalue le maître. Cette dernière forme d'évaluation est constante, implicite, elle gagnerait souvent à être objective et communiquée. Chaque fois qu'un individu est évalué par un autre individu, il y a donc hétéro-évaluation. Ce peut être l'évaluation de l'élève par un autre élève, celle d'un enseignant par ses pairs ou par un inspecteur, etc...

QUAND EVALUER

Une pratique courante consiste à évaluer à la fin d'une période d'apprentissage, lorsque les acquisitions ne sont pas encore stabilisées, en prévenant les élèves du contenu de l'évaluation, en leur demandant de réviser. Il est plus rare de pratiquer une évaluation à moyen terme ou à long terme. Quels savoirs restent disponibles trois mois ou un an après l'apprentissage ? L'évaluation à court terme favorise les attitudes "scolaires" et les apprentissages immédiats, elle ne doit pas être négligée pour autant mais nous avons vu qu'elle devait avoir une fonction formative. Les élèves ne se classent pas de la même façon dans une évaluation à long terme et dans une évaluation à court terme. Il est évident

que l'évaluation sommative devrait intégrer davantage d'évaluations à moyen ou long terme. Dans tout les cas, le moment de l'évaluation est important, et même si les élèves ne sont pas prévenus du contenu de l'évaluation, un minimum de préparation psychologique est indispensable. Chacun sait qu'une évaluation faite le premier jour ou le dernier jour du trimestre rend mal compte des savoirs ou des capacités des élèves. De même évaluer lorsque la réceptivité des élèves est insuffisante, pour calmer l'agitation du moment, voire pour les punir... n'est pas une pratique satisfaisante.

COMMENT EVALUER ?

Nous rangerons dans ce paragraphe aussi bien le problème des instruments d'évaluation que celui des critères à utiliser, le premier étant largement dépendant du second. Se poser le problème des critères, c'est se demander par rapport à quoi on veut évaluer. On peut évaluer par rapport au groupe ou par rapport à un groupe de référence, on peut évaluer au contraire par rapport à des objectifs.

L'EVALUATION NORMATIVE consiste à comparer la performance de l'élève à celle d'un groupe, c'est actuellement celle qui est la plus utilisée. Une épreuve étant donnée, on adaptera éventuellement le barème pour que les élèves moyens obtiennent une note moyenne. Même si cette adaptation n'est pas faite, on s'intéressera essentiellement à la distribution des notes. Le modèle de référence est celui de la courbe de GAUSS (ou courbe normale) et l'évaluateur à tendance à rapprocher sa moyenne de la moyenne normale acceptant seulement qu'elle soit supérieure ou inférieure selon qu'il juge sa classe forte ou faible. Toute épreuve dont les résultats ne rentrent pas dans ce cadre est considérée comme suspecte. N'importe quelle épreuve dispersant convenablement les notes est au contraire considérée comme valable. On cherchera si possible des épreuves originales, des épreuves qui n'auront jamais été données auparavant et qui ne seront sans doute jamais redonnées, l'idée étant que si, en moyenne, les élèves obtiennent des résultats corrects, l'épreuve est bonne et que les autres élèves sont insuffisants ou forts selon le cas. La pratique de l'évaluation normative, outre qu'elle décourage les élèves en difficulté a l'inconvénient de ne pas rendre compte

des savoirs et capacités réelles des élèves. Le sommet est atteint lorsque cette évaluation est faite en utilisant des EPREUVES NORMALISEES, nous avons pu voir, dans une académie voisine, une évaluation de ce type faite sur toutes les classes de troisième, la normalisation conduisait à déclarer "bons" en mathématique des élèves qui obtenaient une note de 5 sur 20, les "moyens" ayant entre 3 et 5. Ce procédé évite en partie de s'interroger sur la valeur de l'épreuve ainsi que sur la qualité de la formation reçue.

Dans l'EVALUATION CRITERIEE, il s'agit au contraire de rapporter les performances de l'élève à des objectifs clairement reconnus et communicables, elle amène à répondre à la question : l'élève a-t-il atteint tel objectif ? La notion de niveau prend alors un sens par rapport à un ensemble d'objectifs à atteindre. Nous aborderons plus loin la question de la définition des objectifs, de leur opérationnalisation ainsi que celle des qualités exigées des épreuves d'évaluation correspondantes. Notons simplement que l'originalité des épreuves ne constitue plus une qualité, celles qui auront fait la preuve de leur validité seront appellées à resservir, éventuellement après amélioration. Les objectifs pouvant être très diversifiés, faire appel tantôt à la production convergente, tantôt à la production divergente, ce sont eux qui pourront être originaux, non pour cultiver l'originalité mais pour traduire les différentes formes de fonctionnement de l'intelligence. On objectera peut être que le professeur doit se renouveler et qu'il n'est pas acceptable d'utiliser plusieurs fois les mêmes épreuves ; c'est peut être que l'on confond situations d'évaluation et situations d'apprentissage. A objectifs égaux, instruments d'évaluation égaux (nous dirons aussi isomorphes), par contre les voies permettant d'atteindre ces objectifs peuvent être différentes.

Ce type d'évaluation présente de nombreux avantages. Le fait que les critères soient communicables permet un meilleur dialogue avec l'élève, il permet aussi au niveau sommatif une meilleure harmonisation. Les progrès de l'élève sont valorisés de même que ses qualités personnelles qui peuvent être très différentes de celles du groupe. A chaque instant l'enseignant sait où l'élève se situe par rapport à l'ensemble des acquisitions jugées nécessaires, l'élève le sait aussi et sait sur quels points il convient de faire porter ses efforts. Si l'on prend la précaution de faire apparaître que les objectifs ne sont pas définis par l'enseignant lui-même, qu'ils font l'objet d'un accord entre enseignants ou qu'ils émanent d'une instance supérieure, bref si l'élève n'a pas

l'impression qu'il lui faut satisfaire à quelque "dada" ou lubie particulière à un enseignant, alors celui-ci sera plus facilement considéré comme un guide et son côté juge sera sans doute mis au second plan. Dans l'évaluation critériée, les problèmes d'ordre docimologiques sont à peu près résolus, ils sont simplement remplacés par celui de la pertinence des objectifs et ceux de la validité des instruments d'évaluation.

Dans ce qui précède, nous avons souvent parlé d'épreuves, il conviendrait plutôt de parler de mise en situation, il s'agit en effet pour chaque objectif de créer les conditions nécessaires à l'observation. Les objectifs ne sont pas tous de type cognitif, certains et non des moindres concernent les attitudes. Par exemple : "devant une situation mathématique nouvelle, l'élève aura un comportement actif, il recherchera le sens des mots nouveaux, il élaborera des schémas, il énoncera des conjectures". Peu importe ici que l'élève parvienne à maîtriser la situation, seule l'observation de son action permettra de savoir si l'objectif est atteint. L'instrument d'évaluation pourra alors être une simple liste de contrôle (cheek-list).

- A cherché le sens des mots nouveaux....
- A relevé la liste des hypothèses....
- A construit une figure....
- A modifié (n) fois cette figure.
- A écris une hypothèse (conjecture).

etc....

Enfin, l'évaluation critériée permet une évaluation différenciée, elle est le complément indispensable de toute pédagogie différenciée.

LES OBSTACLES

Dans ce qui précède, nous avons peut-être donné l'impression de penser que tout était simple et qu'il suffisait de remplacer l'évaluation traditionnelle par une double évaluation critériée : formative puis sommative.

Outre le fait que les référentiels d'évaluation (liste d'objectifs) ainsi que les instruments correspondants fait défaut, il faut tenir compte du poids des habitudes. Ainsi l'évaluation, formative aux yeux du maître, tend à rester sommative dans l'esprit de l'élève. Les familles, l'administration réclament des notes. La tradition veut que les notes trimestrielles soient des moyennes de notes effectivement attribuées. Corriger un devoir, c'est aussi se donner une note de plus, ne pas compter cette note peut donner l'impression d'avoir travaillé pour rien. L'élève lui-même, s'il admet facilement qu'un mauvais résultat ne soit pas comptabilisé, a du mal à l'admettre s'il s'agit d'un bon résultat. Nombreux sont les élèves qui ont l'habitude de tenir leur moyenne à jour, de faire un effort lorsqu'elle passe en dessous d'un certain seuil, de relâcher cet effort lorsqu'à nouveau ce seuil est atteint, ne travaillant ainsi que pour les notes et en fonction des notes. Actuellement la fraude est un phénomène courant dans nos établissements et même les élèves qui ne fraudent pas se donnent souvent beaucoup de mal pour faire illusion, pour faire croire qu'ils savent alors qu'ils ne savent pas. Leur attitude face à l'évaluation est d'abord une attitude de défense, ils ressentent l'évaluation comme une agression. Parmi les perversions de l'évaluation il faudrait citer aussi celles qui ont pour origine l'enseignant, ainsi la volonté de pouvoir ou de puissance, consciente ou non. Le droit de noter confère un pouvoir, sommes nous bien certains de ne jamais en abuser ? Dans le même ordre d'idées il arrive encore que certaines notes n'aient aucun rapport avec l'objet supposé de l'évaluation, elles sanctionnent le devoir non fait, le livre oublié, le bavardage en classe, etc... toutes choses évidemment répréhensibles, qui peuvent être prise en compte dans l'évaluation du comportement, mais qui ne peuvent en aucun cas être intégrées dans une évaluation disciplinaire.

Un autre obstacle est souvent avancé, c'est celui du temps : plus l'évaluation est diversifiée, plus elle est "mangeuse de temps". On a pu dire aux enseignants qu'ils évaluaient trop, cela est sans doute vrai dans le cas de l'évaluation continue, où toute l'évaluation est utilisée à des fins sommatives et où l'élève à l'impression d'être en examen permanent. Cela n'est plus vrai dans le cas de l'évaluation formative, dès lors que l'élève a pris conscience de son intérêt. Dans ce cas il ne s'agit jamais de temps perdu. D'une part cette évaluation permet le renforcement des savoirs, d'autre part elle donne à l'élève des raisons de poursuivre ses apprentissages.

CONCLUSION

Dans cet exposé, nous avons essayé de faire le tour des concepts relatifs à l'évaluation. En ce qui concerne les modalités pratiques nous renvoyons aux autres parties de ce fascicule ainsi qu'aux fascicules 2 et 3. Pour ce qui est de la théorie, nous avons conscience d'avoir été parfois schématique et souvent incomplet et nous renvoyons à la bibliographie.
L'évaluation est un acte complexe qui ne peut être étudié qu'en se référant à de nombreuses disciplines : psychologie, sociologie, statistiques, didactique, etc..., il n'était pas question ici de clore un débat mais simplement de favoriser la réflexion.
Souhaitons y être parvenu.

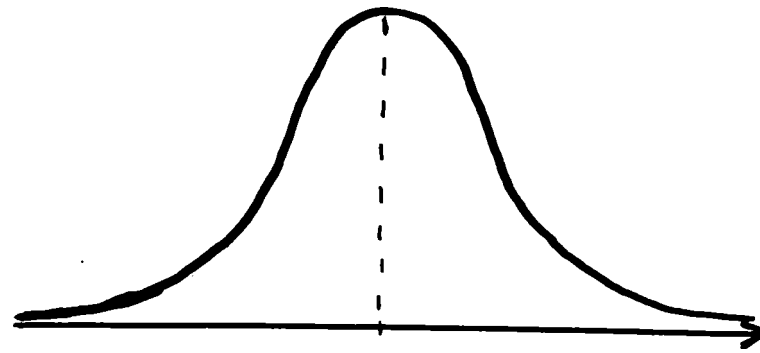
III) LOI NORMALE ET EVALUATION

Dans les pratiques habituelles d'évaluation, la distribution dite "NORMALE" apparait comme le modèle privilégié sinon obligé de toute distribution de notes. De façon délibérée ou inconsciente (par le choix des épreuves par exemple) les correcteurs s'arrangent pour obtenir une distribution qui leur paraisse en accord avec cette conception de la normalité.

Dans ce chapitre, nous voudrions préciser les conditions de validité du modèle normal. Nous supposons que le lecteur est familiarisé avec la notion de distribution d'une variable aléatoire, continue ou non, mais l'ensemble devrait rester compréhensible pour les non initiés.

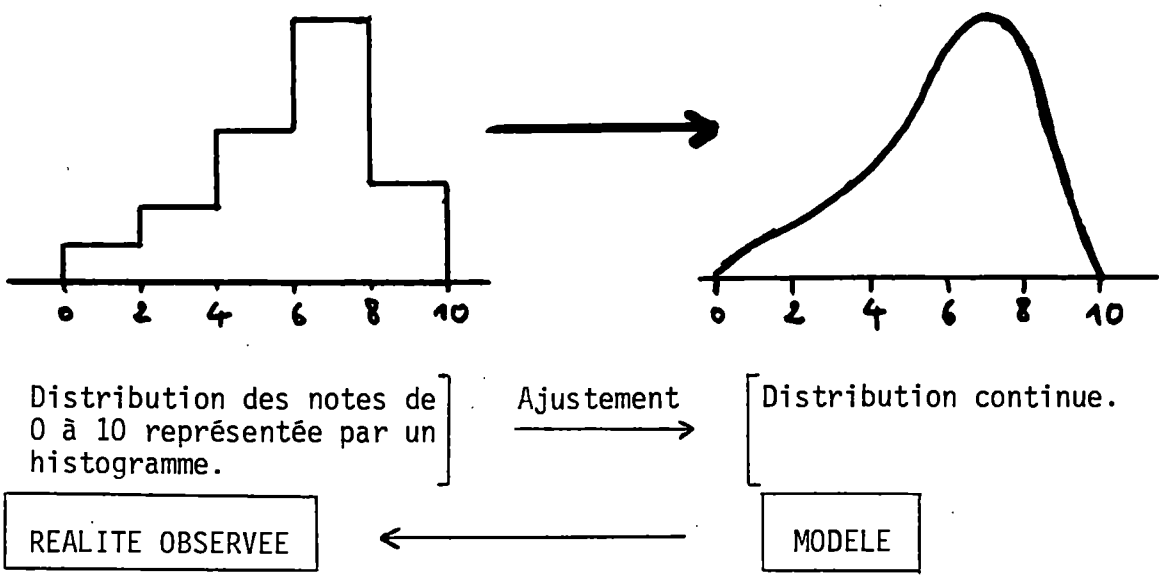
Signalons d'abord que les mots et expressions que l'on trouve dans la littérature sous les vocables suivants :

- Loi normale
- distribution normale
- courbe normale
- courbe en cloche
- courbe de GAUSS
- Loi de LAPLACE-GAUSS



recouvrent le même concept. Dans ce qui suit, nous utiliserons exclusivement les expressions : distribution, loi normale et courbe de GAUSS.

Une évaluation conduit souvent à une note ; l'échelle des notes est en réalité discontinue. La représentation graphique de la distribution est un HISTOGRAMME. Pour le passage au modèle, on est amené à assimiler l'échelle des notes à une échelle continue et à ajuster l'histogramme à une courbe continue, qui est en fait une courbe de densité.



Un exemple : le poids des pommes

Supposons qu'un producteur de pommes pèse une à une chacune des pommes de sa récolte sans en éliminer aucune, et en se limitant à une seule variété : par exemple, des golden. On a une population G (l'ensemble des pommes), un caractère aléatoire X (le poids ou la masse) et une distribution du caractère X. Selon toute vraisemblance (nous expliquerons ce point plus loin), la distribution de X sera correctement ajustée par une distribution normale.

En abscisses : le poids des pommes, en ordonnées le nombre de pommes de poids X. Le nombre de pommes moyennes est alors représenté par l'aire comprise entre la courbe, l'axe des X et les valeurs -a et a.

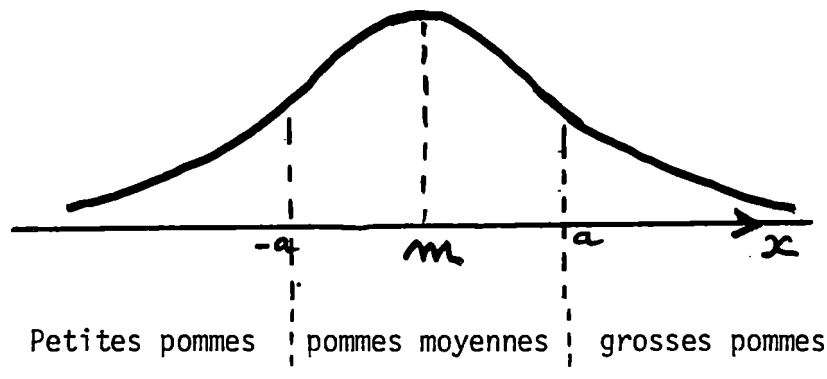


Figure 1.

On obtiendrait le même type de distribution si le caractère X désignait la taille des Français d'âge donné (hommes) ou la production annuelle de lait des vaches laitières de race donnée. Ce ne serait pas du tout le cas si X désignait le revenu annuel des Français.

Supposons maintenant que notre producteur décide de conditionner d'une certaine façon les pommes dont le poids est supérieur à m. Il obtient une sous population G_1 dont la distribution des poids est représentée ci-dessous.

Population G_1

Figure 2.

Si l'on note encore X la variable "poids", il est évident que X ne suit pas une loi normale.

Pour ce qui est de la population G_2 des pommes dont le poids est inférieur à m, on obtient :

Population G_2

Figure 3.

Le producteur peut aussi décider un traitement spécial pour les pommes dont le poids est supérieur à $m + a$ $(a > 0)$, obtenant ainsi une population G_3, ou encore préférer conserver à sa production une certaine homogénéité, ne conditionnant par exemple que les pommes dont le poids serait compris entre $m - a$ et $m + a$, obtenant ainsi une population G_4.

Population G_3

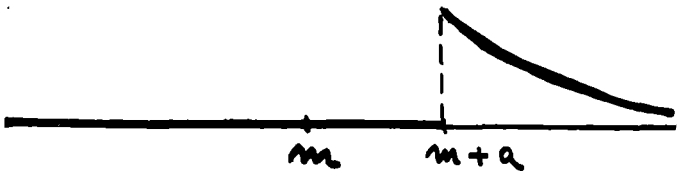


Figure 4.

Population G_4

Figure 4.

Supposons enfin qu'un producteur mélange deux variétés de pommes. Par exemple des golden dont la moyenne des poids serait m_1 et des reinettes dont la moyenne des poids serait m_2. Il obtiendrait alors une population P pour laquelle la distribution de la variable "poids" serait bimodale, du moins dans le cas : $m_1 \neq m_2$.

Population P

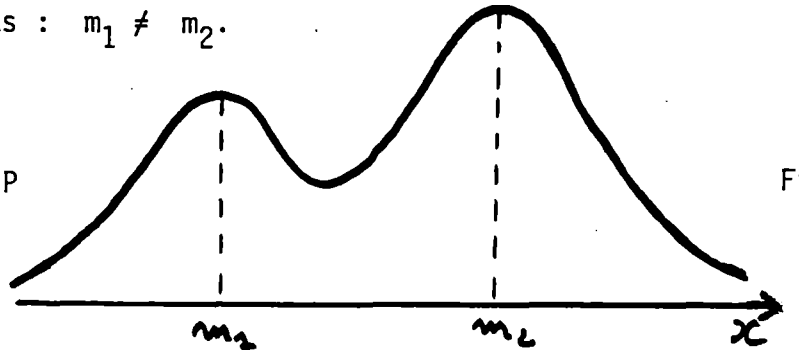


Figure 5.

Pour une pomme donnée, s'il y a doute sur la variété à laquelle elle appartient il paraîtra sans doute naturel de la peser. Son poids ne permettra sans doute pas de déterminer sa variété sans risque d'erreur mais il constituera un indice intéressant.

Dans la pratique, les difficultés inhérentes à toute mesure, les erreurs de tri (devons nous dire de sélection ou d'orientation ?) feront que nous n'observerons pas les discontinuités des figures 2, 3, 4 et 5. Dans le cas de la population G_1 des pommes golden de poids supérieur à m par exemple, on obtiendra plus vraisemblablement la distribution suivante :

Population G_1

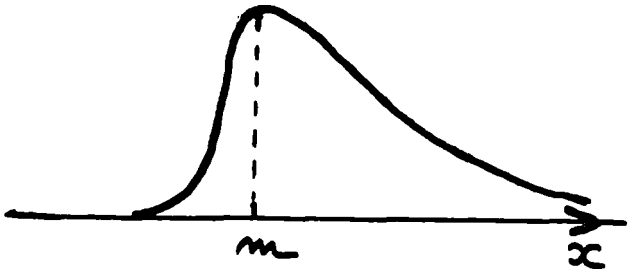


Figure 6.

MESURE ET EVALUATION

Est-il raisonnable de comparer la mesure telle qu'elle se pratique en physique et la "mesure" telle qu'elle intervient en évaluation et plus généralement dans les sciences humaines ?

En fait, le terme de mesure est largement utilisé dans ce domaine. Pour le psychologue J.P. GUILFORD, "MESURER c'est assigner un nombre à un objet ou à un évènement selon une règle logiquement acceptable".

L'évaluation s'appuie souvent sur la mesure telle qu'elle vient d'être définie mais ne s'y réduit pas. "Le terme "évaluation" a un sens beaucoup plus large que le mot mesure. Cette dernière est une description quantitative de comportements, alors que l'évaluation comprend à la fois la description qualitative et la description quantitative des comportements et comporte, en outre, des jugements de valeur concernant leur désidérabilité" (GRONLUND). La mesure n'implique donc pas l'évaluation, mais "l'évaluation n'implique pas davantage la mesure : les informations utiles peuvent rester de type qualitatif..." (J.M. BARBIER).

Dans l'évaluation traditionnelle, la note est de fait considérée comme une mesure (puisque l'on fait des moyennes !) et il semble légitime d'y transposer l'exemple de la distribution du poids des pommes. On peut alors s'étonner de ne retrouver que la distribution de type G et pratiquement jamais des distributions de l'un des types G_1, G_2, G_3, G_4, ou P.

Prenons un exemple : en classe de troisième, les notes trimestrielles de mathématiques sont supposées mesurer un "niveau". Si l'on prend l'ensemble des notes des élèves d'une académie, il ne fait guère de doute qu'elles seront "normalement" distribuées (cf. figure 1). L'orientation ayant joué son rôle, environ 50 % des élèves se retrouvent en seconde, ils correspondent à quelques exceptions près aux élèves dont la note de 3ème en mathématiques était supérieure à la moyenne m ($m \neq 10$). Pour des évaluations faites en début de seconde, on s'attendrait à obtenir une distribution des résultats conformes à la figure 2, ou au moins à la figure 6. Pourtant, l'expérience prouve que dès les premiers jours de l'année c'est le modèle "normal" qui réapparaît. Remarquons que dans le cas de la population G_1 c'est aussi ce qui se passerait si au lieu de continuer à peser les pommes on se mettait à mesurer leur teneur en sucre par exemple, autrement dit si l'on modifiait la variable mesurée. (Voir aussi l'article EVALUATION et LIAISONS COLLEGE-LYCEE). Si nous avons pris cet exemple, ce n'est pas parce qu'il nous semble que les professeurs de Lycée notent d'une façon fondamentalement différente que les professeurs de collège, bien au contraire. Il se trouve simplement qu'au niveau de cette "coupure", l'inclusion d'une population dans l'autre permet d'observer un phénomène qu'il serait plus difficile d'observer au niveau du seul collège. Nous pensons plutôt qu'à tous les niveaux les enseignants ont besoin d'avoir leurs lots d'élèves forts, d'élèves moyens et d'élèves faibles et que quelle que soit la population de départ, il rétablissent d'une façon ou d'une autre la "normalité". Il ne s'agit pas ici de chercher à culpabiliser qui que ce soit. Nous avons tous été formés ainsi et avons sans doute complètement intériorisé ce processus. De plus certaine remarques qui seront faites plus loin pourront être utilisées pour justifier cette pratique, au moins dans certains cas.

L'EVALUATION DES EVALUATEURS

Que dit-on d'un enseignant dont les notes se répartissent suivant une courbe en $\int$?

On dira en général qu'il est trop indulgent ou même qu'il est laxiste.

Pourtant dans une évaluation succédant à un apprentissage réussi, c'est cette distribution qui mériterait le qualificatif de "normal" : la plupart des élèves réussissent, quelques uns échouent. Dans une évaluation portant sur les objectifs, on dira que la plupart des élèves ont atteint la plupart des objectifs.

Courbe en j
ou courbe de la REUSSITE

Figure 7.

Que se passe-t-il maintenant dans le cas d'un enseignant dont les résultats se distribuent selon une courbe en i ?

On dira le plus souvent qu'il est trop sévère. Cela peut être vrai, mais si l'enseignant a évalué par rapport à des objectifs propres au niveau considéré, cela devrait simplement signifier que les objectifs ne sont pas atteints et que la formation devrait être reprise. Cela ne signifie aucunement que l'enseignement ait été de mauvaise qualité, bien d'autres variables interviennent et entre autre la compléxité cognitive de l'objectif ou de la classe d'objectifs évalués (penser à la proportionnalité !).

Courbe en i
ou courbe de L'ECHEC

Figure 8.

Ces deux modèles d'évaluation correspondent à une évaluation objective permettant non seulement de comparer l'élève à un groupe mais aussi et surtout de rapporter ses savoirs et savoir-faire à des objectifs bien spécifiés.

Que dit-on enfin de l'enseignant dont la distribution des notes est approximativement normale ? (figure 1).

On dit qu'il note bien, qu'il est plutôt sévère ou que sa classe est faible si m est inférieur à 10, plutôt indulgent ou que sa classe est forte si m est supérieur à 10.

On s'étonne rarement d'être ramené au cas des pommes sur lesquelles aucune opération de sélection ni de formation n'a été effectuée.

Pour des objectifs de maîtrise, l'évaluation en fin d'apprentissage (réussi !) conduit à une distribution en j, mais pour ces mêmes objectifs, l'évaluation en début d'apprentissage est nécessaire. Par contre une évaluation en cours d'apprentissage conduit "normalement" à une distribution normale. Il faut alors se demander si nous n'effectuons pas trop systématiquement des évaluations en cours d'apprentissage, alors que le savoir de l'élève est en train de se structurer, que les maturations nécessaires n'ont pas encore eu lieu ? Les différences individuelles ainsi mesurées ne sont-elles pas autre chose que des différences dans les rythmes d'acquisition ?

Avant d'aller plus loin, il est temps de définir de façon plus rigoureuse la distribution normale.

LA DISTRIBUTION NORMALE

DEFINITION : Une variable aléatoire X a une distribution normale de moyenne μ et d'écart-type σ si la loi de probabilité de X a pour densité l'application f telle que :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

La courbe représentative de f est appelée COURBE DE GAUSS.

On effectue en général un changement de variable

$$x \longmapsto \frac{x - \mu}{\sigma}$$

pour se ramener à la distribution normale réduite (de moyenne 0 et d'écart-type 1).

La formule de changement de variable permet de passer facilement d'une distribution normale à une autre.

Une étude simple montre que la courbe normale réduite est symétrique par rapport à l'axe des y et qu'elle admet ses points d'inflexions pour points d'abscisses +1 et -1.

On montre d'autre part en notant U la variable réduite et Pr la probabilité :

$Pr\ (-1 \leqslant U \leqslant +1) \simeq 0{,}66$

$Pr\ (-2 \leqslant U \leqslant +2) \simeq 0{,}95$

$Pr\ (-3 \leqslant U \leqslant +3) \simeq 0{,}99$

Pour une population finie (qui ne peut qu'approximativement être normalement distribuée) cela signifie que :

66 % de l'effectif appartient à l'intervalle $[m-\sigma\ ,\ m+\sigma]$

95 % de l'effectif appartient à l'intervalle $[m-2\sigma,\ m+2]$

.... ce qui est résumé ci-dessous.

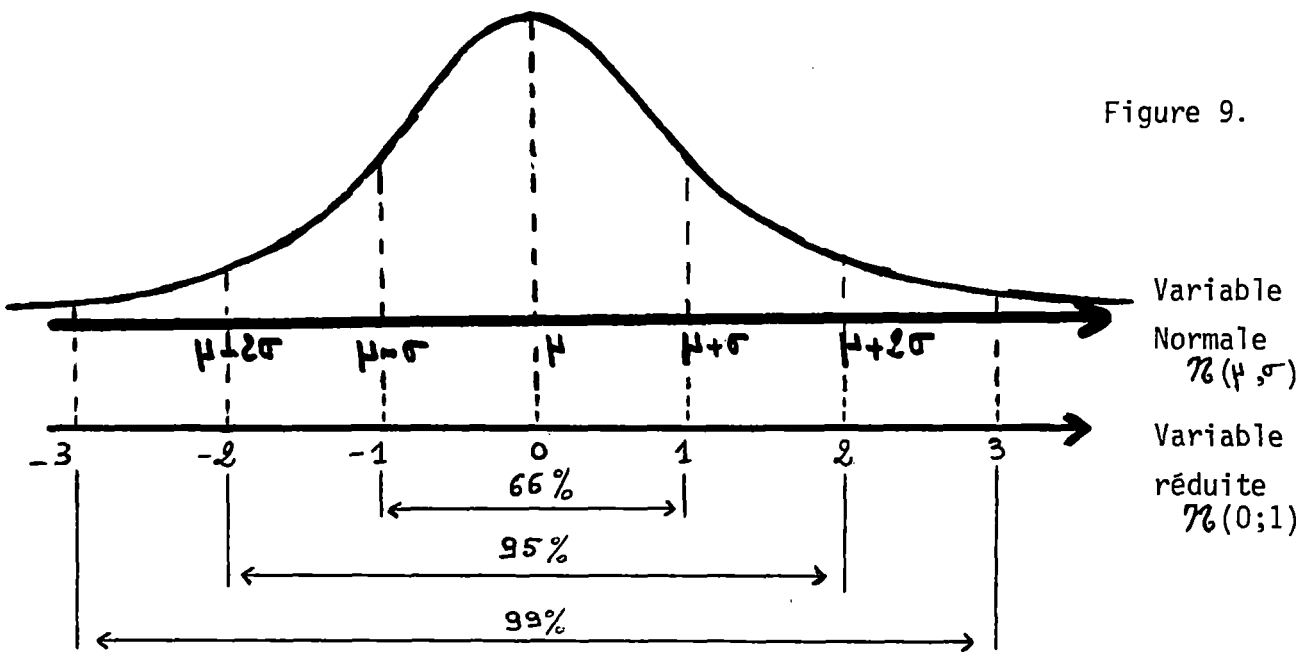
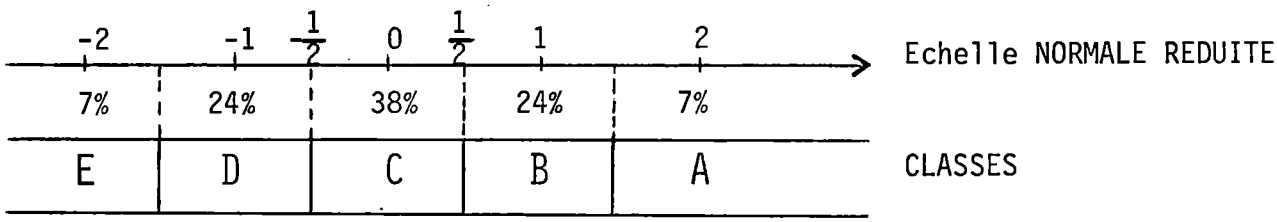


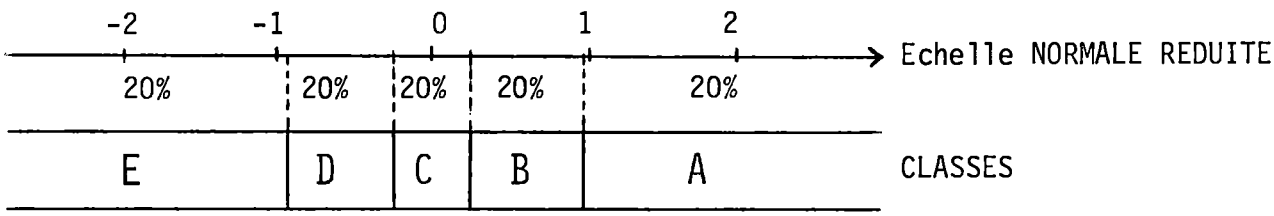
Figure 9.

Epreuves normalisées : Une épreuve est dite normalisée si elle a été étalonnée sur une population de référence. Les notes sont alors rapportées à l'échelle normale réduite et réparties en classes définies par des fractions d'écart-type.

Par exemple, pour une étalonnage en cinq classes :



On utilise parfois la méthode des quantiles qui conduit à des classes de même effectif (ou de même probabilité) par exemple, toujours dans le cas d'un étalonnage en cinq classes :



Ces techniques d'étalonnage peuvent être utilisées même si la distribution des notes ne relève pas du modèle "normal", la signification à accorder aux classes A, B, C, D, E, doit bien entendu être alors revue dans chaque cas.

Domaine de Validité du modèle "NORMAL"

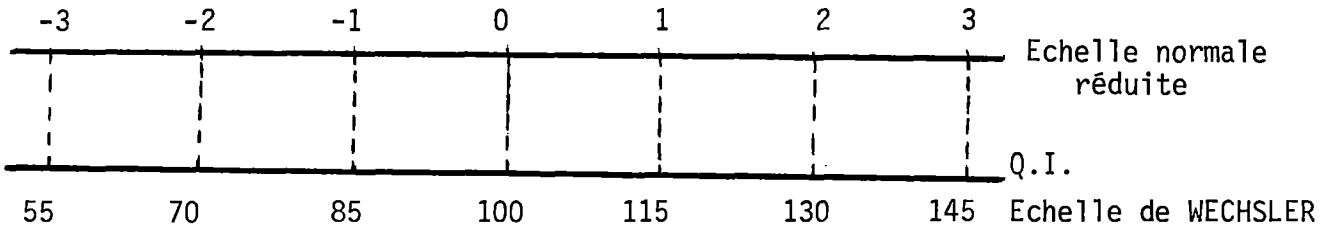
Nous avons remarqué ci-dessus qu'une variable aléatoire n'était pas toujours distribuée normalement. Pourquoi en serait-il ainsi des capacités en mathématiques, du niveau de raisonnement, du niveau de connaissances dans un domaine particulier ? Surtout lorsqu'il s'agit d'une population

d'élèves ayant été préalablement sélectionnés, ou du moins orientés.

En réalité, on a démontré qu'une variable aléatoire qui subit l'influence de nombreuses causes de variations, indépendantes les unes des autres, petites et de même ordre de grandeur, agissant de façon additive, pouvait être approchée par une loi normale.

Pour revenir aux pommes du début, on pouvait effectivement penser que l'ensemble des facteurs intervenant dans la croissance des pommes vérifiait les conditions ci-dessus. Il reste que si par hasard une variété d'insectes attaque électivement les pommes les plus grosses, notre modèle risque fort de ne plus être adapté.

Sans vouloir entrer dans le débat sur le Q.I. et l'intelligence qui agite périodiquement les spécialistes, on peut admettre en première approximation que l'intelligence générale (dont nous ne saurions donner une définition précise), si elle est mesurable (?), se distribue selon la loi normale. Il n'est donc pas surprenant que les épreuves considérées comme valables dans ce domaine conduisent à des résultats distribués normalement. Le Q.I. n'est autre qu'une mesure rapportée, au moins dans l'échelle de WECHSLER, à l'échelle normale.



Si l'on propose une épreuve de mathématiques du baccalauréat C à des professeurs de mathématiques enseignant en TC, on s'attend à obtenir une distribution des résultats selon une courbe en J, le contraire ne manquerait pas d'inquiéter. Si on propose la même épreuve à des professeurs de lettres, on s'attend plutôt à une courbe en L. Si maintenant on mélange les deux

populations on peut s'attendre à ce que les résultats se distribuent selon une courbe en U.

Courbe en U

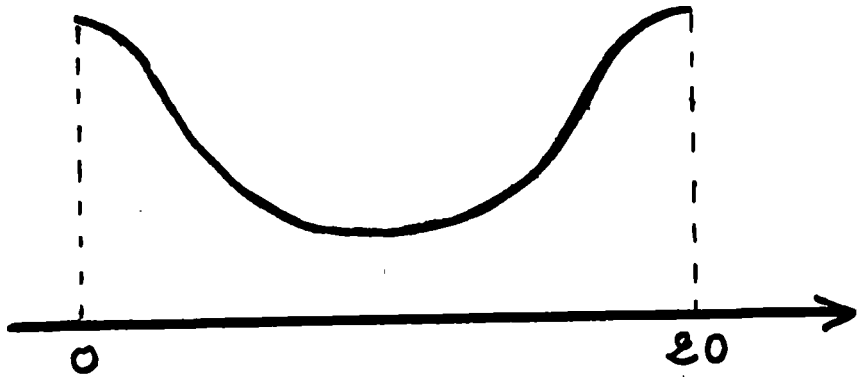


Figure 10.

La distribution des résultats permettra sans doute de distinguer par leurs seules productions les professeurs de mathématiques des professeurs de lettres. Mais qu'a-t-on mesuré ? certainement pas l'intelligence ! On a plutôt mesuré des capacités spécifiques.

Comment se fait-il alors que la même épreuve proposée à des candidats au baccalauréat nous ramène inexorablement à la distribution normale ? Ces candidats ont eu la même formation, ont subi les mêmes apprentissages. En ce qui les concerne, cette épreuve ne constitue-t-elle pas en réalité une épreuve composite prenant en compte pêle-mêle l'intelligence du sujet, ses aptitudes spécifiques pour les mathématiques, ses capacités du moment, ses savoirs et savoir-faire en mathématiques...? On peut se demander si, pour parler comme les psychologues, l'épreuve n'est pas pour eux, "saturée en facteur g" (qui rend compte de l'intelligence générale). On peut trouver quelque justification à cette méthode, mais il convient alors de noter le caractère redondant des évaluations pratiquées. Le facteur g intervient en effet dans toutes les épreuves de toutes les disciplines, il n'est pas ou très peu modifié par les apprentissages. Tout semble se passer comme si en éducation physique, l'évaluation prenait largement en compte la taille du sujet. C'est d'ailleurs ce qui se passe, mais à contrario, plus le sujet est petit, moins il doit sauter loin pour avoir une note donnée.

Dans le cas de l'évaluation à la fois certificatrice et intégratrice que constitue le baccalauréat on peut sans doute défendre certaines pratiques mais il est évident que leur transposition à tous les niveaux de la scolarité et spécialement dans le contexte de l'évaluation formative ne peut manquer de poser des questions. La réflexion actuelle sur l'"échec scolaire" ne pourra pas faire l'économie de ces questions et devra s'appuyer sur une analyse rigoureuse des procédures d'évaluation.

En conclusion

Malgré les habitudes acquises, il faut se garder de considérer le modèle "normal" comme le modèle souhaitable. L'enseignement a pour but de détruire cette normalité, toute capacité spécifique acquise est en fait une anormalité. Dans le cas d'une évaluation de maîtrise ou d'une évaluation portant sur un seul objectif, le modèle approprié serait plutôt le type dichotomique (réussite-échec ; 0 - 1 ; oui-non). Si cette évaluation porte sur une famille d'objectifs, les distributions en $\mathbf{L}$ et en $\mathbf{J}$ sont les plus satisfaisantes.

Enfin si l'évaluation a pour objet de discriminer comme c'est souvent le cas de l'évaluation sommative, mais aussi de l'évaluation diagnostique, le modèle en U de la figure 10 semble le mieux adapté.
Dans tous ces cas, lorsqu'une coupure doit être effectuée, elle peut se faire à un endroit de densité minimum, alors que dans le modèle normal elle se fait souvent à un endroit de densité maximum, ce qui a pour effet évident de maximiser les risques d'erreurs de jugement.

IV) OBJECTIFS ET TAXONOMIES

Cette partie est destinée à servir de document de travail, et à donner aux lecteurs ainsi sensibilisés, l'envie de se reporter aux ouvrages spécialisés. (En particulier : "Définir les objectifs de l'éducation" de G. et V. de Landsheere).

La taxonomie est la science des classifications. La première taxonomie des objectifs pédagogiques est celle de B.S. BLOOM (publiée en 1956). Elle concerne les objectifs cognitifs. C'est surtout un moyen de classer les exercices. Cette taxonomie est maintenant largement connue et utilisée. Elle a donné naissance à d'autres modèles plus ou moins voisins. En ce qui concerne les mathématiques, elle a été adaptée pour fournir le modèle N.L.S.M.A. que nous présentons plus loin. Auparavant, nous voudrions parler du modèle tridimensionnel de J.P. GUILFORD ; il ne s'agit pas d'une taxonomie d'objectifs, mais d'une représentation du fonctionnement de l'intelligence. Ce modèle nous parait particulièrement intéressant pour des pédagogues qui cherchent à mettre en valeur et à développer les divers aspects de l'intelligence de leurs élèves. Même dans le cas de l'évaluation traditionnelle, il peut être utilisé pour repérer les comportements sollicités implicitement.

GUILFORD considère trois dimensions : les OPERATIONS (intellectuelles), les CONTENUS et les PRODUITS. "Les processus mentaux se traduisent par des OPERATIONS aboutissant à des PRODUITS de CONTENUS divers" (L. Vandevelde : peut-on définir les objectifs en éducation). Le schéma proposé plus loin est assez explicite, voici toutefois quelques précisions :

- LA COGNITION désigne ici le connu disponible, le souvenir, ce qui subsiste d'une assimilation antérieure.
- LA PRODUCTION CONVERGENTE est génératrice d'informations uniques, à partir d'un donné, avec acceptation de règles précises, de conventions.
- LA PRODUCTION DIVERGENTE est génératrice d'informations variées à partir d'un donné unique, elle est source d'originalité, d'invention.

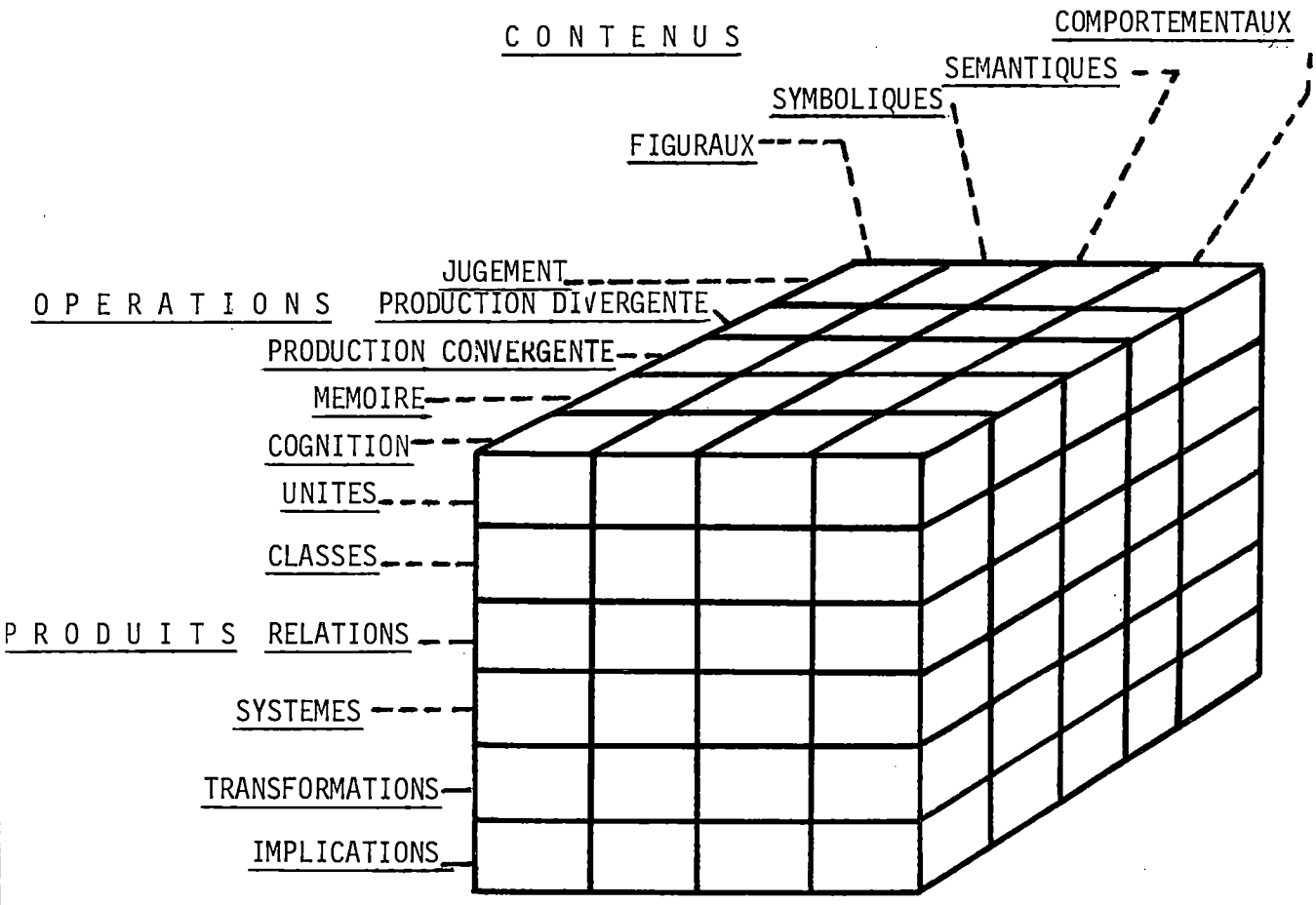
En ce qui concerne les produits, voici la définition qu'en donne de LANDSHEERE

- UNITES : portions d'informations relativement isolées ou circonscrites
- CLASSES : unités groupées en raison de leurs propriétés communes
- RELATIONS : connexions reconnues entre les unités
- SYSTEMES : groupements d'unités organisées ou structurées. Complexe de parties se trouvant en interrelations ou en interaction.

- TRANSFORMATIONS : Changements apportés dans des informations ou dans leur utilisation
- IMPLICATIONS : Extrapolation d'informations : prédiction, conséquence, antécédents

L'analogie avec les concepts qui nous sont familiers : ELEMENTS, ENSEMBLES, RELATIONS, STRUCTURES, APPLICATIONS, Raisonnement INDUCTIF et raisonnement DEDUCTIF, est ici manifeste.

Voici le schéma annoncé :



On obtient ainsi 120 combinaisons définissant autant d'activités intellectuelles élémentaires. Par exemple, le passage d'une représentation graphique à une autre prendra place dans la cellule : Cognition de Transformations Figurales. Ce modèle a parfois été utilisé pour la définition des objectifs et pour classer les exercices. Il faut remarquer que, le plus souvent, une tâche donnée recouvrira plusieurs cellules du cube, ce qui ne semble pas être un obstacle majeur à son utilisation.

D'une façon plus classique, nous avons surtout utilisé le modèle N.L.S.M.A. (National Longitudinal Study of Mathématical Abilities) dont nous donnons plus loin une version extraite de la Thèse de F. PLUVINAGE (cf bibliographie). Nous avons aussi reproduit un tableau, extrait de cette même thèse, qui explique en partie le mode d'emploi de la taxonomie. Pour plus de détail, il conviendrait de se reporter au travail initial.

Nous nous sommes aussi référé à la taxonomie de R. GRAS (cf bibliographie) qui insiste davantage sur les comportements attendus des élèves.

Les taxonomies sont d'abord des instruments d'analyse des objectifs existants, mais comme le souligne G. de LANDSHEERE : "en pointant dans une taxonomie les comportements susceptibles de traduire l'objectif poursuivi, on peut aussi apercevoir des comportements pour lesquels on ne dispose pas d'objectifs. A ce moment, la taxonomie devient une source d'inspiration d'objectifs nouveaux". De fait, ces taxonomies nous ont servi de guide dans notre travail et nous ont amenés à diversifier nos objectifs.

Partant d'une pratique où le questionnement se plaçait pratiquement toujours au niveau des savoirs spécifiques (cf l'article : analyse d'une épreuve de B. E. P. C.) nous avons peu à peu construit des questions qui prennent place dans les niveaux supérieurs des taxonomies.

LA CLASSIFICATION N.L.S.M.A.

Avertissement

Cette classification s'applique aux énoncés d'exercices mathématiques. Elle ne prétend situer qu'une partie de leur difficulté. Celle qui concerne la complexité des connaissances nécessaires à leur résolution. Son application exige de connaître l'apprentissage qui précède un exercice.

Préliminaire

- Définition 1. Les faits scientifiques sont les "connaissances atomiques" : parmi les connaissances, les faits scientifiques sont caractérisés par le fait d'être isolément mémorisés et formulés. Autrement dit, un fait spécifique est exprimé par une phrase, française ou symbolique, simple (c'est-à-dire sans subordonnée).
- Définition 2. Un concept est un ensemble de faits spécifiques.

La classification

Niveau A : La connaissance et l'utilisation des faits scientifiques mémorisés.

- A1 : Connaissance des faits spécifiques
- A2 : Connaissance de la terminologie
- A3 : Aptitude à effectuer des algorithmes

Niveau B : La connaissance et l'utilisation des concepts mémorisés.

- B1 : Connaissance des concepts
- B2 : Connaissance de principes, règles, généralisations
- B3 : Connaissance des structures mathématiques

B4 : Aptitude à traduire un énoncé d'une formulation à une autre

B5 : Aptitude à suivre un raisonnement

B6 : Aptitude à interpréter des données

Niveau C : Les applications

C1 : Aptitude à résoudre des problèmes routiniers

C2 : Aptitude à comparer, ordonner

C3 : Aptitude à analyser des données

C4 : Aptitude à reconnaître des relations (ex : périodicité, symétrie... reconnaissance de formes)

Niveau D : La découverte

D1 : Aptitude à résoudre des problèmes inhabituels

D2 : Aptitude à découvrir des relations

D3 : Aptitude à démontrer

D4 : Aptitude à critiquer la validité d'un raisonnement

D5 : Aptitude à formuler et valider des généralisations

Addendum

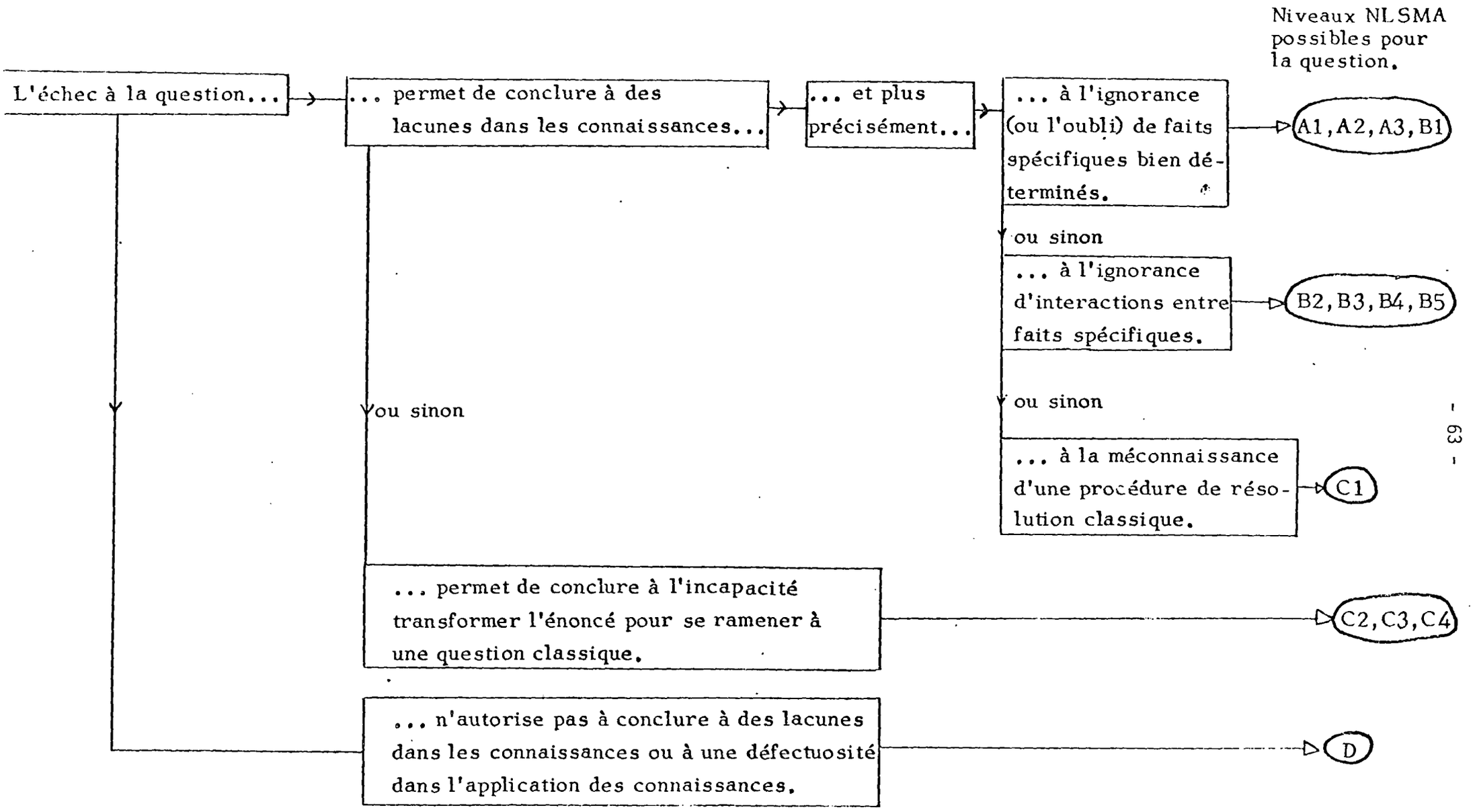
Aux quatre niveaux ci-dessus concernant le domaine cognitif, Wilson ajoute deux niveaux qui concernent le domaine affectif :

Niveau E : Attitudes et intérêts (motivation, appréhensions, goûts personnels)

Niveau F : Appréciation (utilité des mathématiques, valeur de l'enseignement mathématique, élaboration).

o
o o

Signification de l'échec à une question et niveau NLSMA.



(Thèse de F. Pluvinage)

TABLEAU D'OBJECTIFS COGNITIFS (MATHEMATIQUES)

(taxonomie de Rigi Gras)

CLASSES	RUBRIQUES	OBJECTIFS	ACTIVITES ATTENDUES
(A) nnaissance des vils de préhension l'objet et du fait thématiques	A_1	Connaissance de la terminologie et du fait spécifique	associer assembler
	A_2	Capacité d'action intériorisée sur l'évocation d'une forme physique du concept	simuler observer
	A_3	Capacité de lire des cartes, des tableaux, des graphiques, des notices	déchiffrer décrire
	A_4	Effectuation d'algorithmes simples	organiser, calculer
(B) alyse de faits transposition	B_1	Substitution d'une démarche représentative à une manipulation Anticipation graphique	abstraire prolonger induire
	B_2	Reconnaissance et usage d'une relation implicite simple où intervient l'objet mathématique connu	analyser comparer
	B_3	Traduction d'un problème d'un mode dans un autre avec interprétation	schématiser, traduir transposer
(C) préhension des tions et structures	C_1	Compréhension du concept, de ses relations avec les autres objets mathématiques	reconnaître construire
	C_2	Compréhension d'un raisonnement mathématique : justification d'un argument	justifier
	C_3	Choix et ordonnancement d'arguments	déduire
	C_4	Application dans des situations familières	analyser, abstraire appliquer, interpole
(D) nthèse et éativité	D_1	Effectuation et découverte d'algorithmes composites et de nouvelles relations	organiser, calculer optimiser
	D_2	Construction de démonstrations et d'exemples personnels	illustrer, démontrer valider, créer, inventer
	D_3	Découverte de généralisations	généraliser, induire prévoir, extrapoler
	D_4	Reconnaissance du modèle et applications dans des situations non routinières	modéliser, identifier, différencier, classifier, résumer
(E) ritique et Evaluation	E_1	Distinction du nécessaire et du suffisant	formuler des hypothèse, déduire
	E_2	Critique de données et de méthodes ou de modèles résolvants	contrôler, optimiser prévoir, critiquer, questionner, vérifie
	E_3	Critique d'argumentation et construction de contre-exemple	critiquer, tolérer contredire

V) LE RECUEIL DE L'INFORMATION

Une fois les objectifs reconnus et opérationnalisés, il est indispensable de disposer d'un instrument de recueil des résultats qui permette à tout moment de l'année de savoir comment l'élève se situe par rappport à chacun des objectifs.

Le travail de l'I.R.E.M. porte essentiellement sur les objectifs de type cognitif et nous disposons maintenant d'instruments d'évaluation couvrant la plus grande partie de ces objectifs au niveau du collège (voir fascicules 2 et 3). Il ne faudrait pas pour autant négliger l'évaluation des comportements, des attitudes et des aptitudes. L'intérêt manifesté par l'élève, ses qualités d'attention, son aptitude à communiquer, sa capacité à travailler avec d'autres... sont autant d'éléments que l'évaluation doit prendre en compte. Dans ce domaine, une évaluation objective est sans doute possible dans la plupart des cas, mais suppose une explicitation des objectifs qui est loin d'être réalisée. Quand dira-t-on d'un élève qu'il est soigneux ? Qu'il est créatif ? Comment mesurer l'attention, la rigueur ? Pour l'instant, nous devons nous contenter d'une évaluation subjective de ces qualités.

Depuis plusieurs années, nous utilisons la "fiche individuelle d'évaluation et d'observation continue" qui est reproduite ci-après. Cette fiche utilise dans sa partie supérieure le découpage en familles d'objectifs et en micro-objectif de l'I.R.E.M. Elle permet dans sa partie inférieure une évaluation subjective des attitudes et des aptitudes par recueil d'indices et cela par le moyen de l'observation continue des élèves. Par exemple, au cours de l'étude avec la classe, d'une situation mathématique, tel élève montrera un intérêt inhabituel, tel autre aura une idée originale permettant de faire avancer la solution, tel autre parviendra à communiquer clairement ses observations... ces faits méritent d'être notés (pas sur 20 !). Il est difficile de pratiquer l'observation tous azimuths et il est parfois plus efficace, au cours d'une séance, de pratiquer l'"observation dirigée", qui consiste en fait à sélectionner pour un temps le type d'informations recueillies. Dans tous les cas, on en tire une meilleure connaissance de l'élève et le dialogue avec lui, avec la famille et au sein de l'équipe éducative s'en trouve amélioré d'autant.

En particulier, cette fiche a été utilisée dans une expérience d'enseignement par groupes de niveaux. La très grande mobilité des élèves d'un groupe à l'autre, selon les thèmes étudiés, et en fonction des tests de prérequis, en a été grandement facilitée. La fiche-professeur reproduite plus loin est semblable à la fiche-élève mais est cartonnée et une place y est réservée à la photo de l'élève. A chaque changement de groupe il suffisait donc aux professeurs d'échanger leurs fiches pour que la continuité soit assurée. Dans notre expérience, les enseignants changeaient aussi de groupe à chaque changement de thème, si bien que pour chacun d'eux le renouvellement des élèves était quasi total et il est clair qu'un instrument résumant les acquis de l'élève était nécessaire. Cette fiche s'est révélée en outre très utile lors des conseils de classes, n'importe lequel des trois professeurs de l'ensemble pouvant intervenir à propos de n'importe que élève, en se référant à sa fiche individuelle.

Sur la fiche, les micro-objectifs sont notés en REUSSITE-ECHEC (VERT ou ROUGE), le code utilisé montre bien que l'échec est considéré comme temporaire et l'élève sait qu'il sera a nouveau confronté à une tâche semblable (épreuve-bis). La réussite est définitive mais devrait être lue : "a su faire..." les épreuves de validation ayant justement pour but de s'assurer de la permanence des acquisitions ainsi que de la possibilité de les mobiliser dans une épreuve intégrant plusieurs objectifs. Les tests de validation font encore partie de l'évaluation formative, ne serait-ce qu'à cause de l'existence de tests bis et ter. Pour éviter la confusion avec l'évaluation sommative, ils sont notés sur 100. Il s'agit en fait d'un SCORE et non d'une note au sens habituel.

Les seules notes sur 20 sont donc les notes trimestrielles. Pour des raisons de communication et pour tenir compte des habitudes actuelles, il ne nous a pas paru possible d'éviter ces notes, mais il est évident qu'elles jouent pour nous un rôle très secondaire. Compte tenu du système d'évaluation utilisé, elles ne peuvent être des moyennes de notes attribuées. La note de NIVEAU résume la partie objective de l'évaluation tandis que la note d'APPLICATION en résume la partie subjective.

La méthode de calcul de ces notes est communiquée aux élèves, mais ne peut pas toujours être la même pour tous. En effet, l'évaluation étant en partie individualisée, tel élève a pu passer une épreuve renforcée au niveau des savoirs d'approfondissement tandis qu'un autre aura passé un test bis ou ter portant sur les savoirs minima. Les résultats sont difficilement comparables et il y aurait quelque injustice à les intégrer sans précaution à une évaluation sommative.

La méthode de calcul change d'ailleurs en cours d'année et selon les classes. Au 1er trimestre elle peut sans grand inconvénient intégrer les résultats du contrôle des micro-objectifs, de caractère purement formatif, tandis que la note du 3ème trimestre qui a une fonction nettement sommative, prend en compte les résultats aux épreuves de validation en ne retenant s'il y a lieu que les épreuves bis ou ter ainsi que les épreuves renforcées, les tests portant sur les activités et certains devoirs de type traditionnel.

Les élèves, tout au long de l'année tiennent eux-mêmes leur fiche personnelle et il est remarquable que sur les 15 classes qui ont utilisé ce système nous n'avons pratiquement pas rencontré d'élèves qui n'aient eu a coeur de tenir soigneusement leur fiche à jour. Ils ne parlent plus de notes mais d'objectifs atteints ou non, ils sont souvent capable de demander une fiche bis : "je voudrais repasser la fiche D4 que je n'ai pas réussi le mois dernier, maintenant, j'ai compris...".

Les devoirs faits à la maison sont corrigés mais non notés, ils sont l'occasion de relever des indices pour la partie subjective de l'évaluation. Un élève en difficulté qui produit un travail personnel (réellement personnel !) faible mais sérieux sera encouragé et les indices recueillis seront positifs, un élève fort qui produira le même type de travail aura une appréciation sévère et les indices recueillis seront négatifs.

Contrairement à ce que l'on pourrait craindre, la suppression des notes dans ce domaine n'a jamais eu pour conséquence la démission des élèves. Les devoirs sont faits régulièrement, la qualité n'a pas diminuée mais par contre le copiage a en grande partie disparu.

Les parents dans l'ensemble réagissent favorablement. Quelques uns, surtout en début d'année, s'inquiétent de l'absence de notes, mais après explication du POURQUOI autant que du COMMENT, ils acceptent assez facilement cette pratique. Nous avons remarqué que les parents comme les élèves acceptaient d'autant plus facilement les changements que l'on prenait la peine de leur en expliquer les motifs et que leurs avis étaient sollicités. Dans certaines classes, nous avons pris la précaution d'informer les parents par une note explicative. Ce document, que nous publions dans les pages qui suivent, était aussi destiné à l'information des élèves, des collègues et de l'administration.

Enfin, on trouvera à la fin de cet article une fiche d'évaluation simplifiée qui ne prend en compte que les résultats aux tests de validation. Plusieurs collègues l'on utilisée avec profit. Si l'on pense que les changements brusques ne sont pas souhaitables, il peut être intéressant d'utiliser une fiche de ce type qui permet de laisser cohabiter les deux systèmes d'évaluation en ménageant ainsi, pour nous mêmes et pour les élèves, une transition peut-être nécessaire.

Ces documents sont publiés à titre d'exemples et non de modèle, ils sont encore imparfaits et demandent à être améliorés. Il appartient en fait, à chaque professeur, à chaque équipe éducative de construire les instruments d'évaluation qui corresponde à ses besoins, en s'inspirant des documents existants que l'on peut trouver aussi bien dans cette brochure que dans d'autres ouvrages (voir bibliographie).

NOM :
CLASSE :

FICHE INDIVIDUELLE d'EVALUATION et d'OBSERVATION CONTINUE

I.R.E.M. de BESANCON

familles d'OBJECTIFS (les lettres A, B, C,.... ainsi que les nombres 1, 2,... renvoient à des documents annexes)	prérequis	CONTRÔLE DES MICRO-OBJECTIFS															TESTS de Validation					
		savoir minimum										Approfondissement					n° 1		bis		ter	
		1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	t	c	t	c	t	c
A																						
B																						
C																						
D																						
E																						
F																						
G																						

TESTS RENFORCES	t	c
R1		
R2		
R3		
R4		

TESTS activités	t	c
k1		
k2		
k3		
k4		

Observation continue des attitudes et des aptitudes	
attention	
motivation - intérêt	
régularité des apprentissages	
travail personnel	
expression écrite	
expression graphique	
expression orale	
imagination - créativité	
soin	travail en groupe
rigueur	auto-évaluation
rapidité	.....
mémoire	.....

Notes trimestrielles attribuées				
	1er T	2ème T	3ème T	
niveau				sur
application				20

LEGENDE

t désigne le taux de réussite au test (sur 100)

c désigne un code couleur

vert (réussite)..... $t \geq 70$
vert hachuré..... $50 \leq t < 70$
rouge hachuré..... $30 \leq t < 50$
rouge (échec)..... $t < 30$

aptitudes et attitudes

..... très bien
..... bien
..... passable
..... insuffisant
..... très insuffisant

SIGNATURES des PARENTS
1
2
3
4
5
6

- 689 -

fiche - professeur

PHOTO de l'ELEVE

NOM :

CLASSE :

FICHE INDIVIDUELLE d'EVALUATION et d'OBSERVATION CONTINUE

	prérequis	CONTRÔLE DES MICRO-OBJECTIFS															TESTS de Validation					
		savoir minimum										Approfondissement					n° 1		bis		ter	
		1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	t	c	t	c	t	c
A																						
B																						
C																						
D																						
E																						
F																						
G																						

TESTS RENFORCES		t	c
R1			
R2			
R3			
R4			

TESTS activités		t	c
k1			
k2			
k3			
k4			

Observation continue des attitudes et des aptitudes

attention	
motivation - intérêt	
régularité des apprentissages	
travail personnel	
expression écrite	
expression graphique	
expression orale	
imagination - créativité	
soin	travail en groupe
rigueur	auto-évaluation
rapidité	.....
mémoire	.....

Notes trimestrielles attribuées				
	1er T	2ème T	3ème T	
niveau				sur
application				20

FICHE INDIVIDUELLE D'EVALUATION
ET D'OBSERVATION CONTINUE EN MATHEMATIQUES

Note destinée aux PARENTS, aux élèves et aux membres de l'équipe éducative.

L'utilisation de cette fiche s'inscrit dans le cadre d'une "PEDAGOGIE PAR OBJECTIFS" associée à une "PEDAGOGIE DE MAITRISE" et à une "PEDAGOGIE DE LA REUSSITE" qui sont autant de pratiques trouvant naturellement leur place dans un "ENSEIGNEMENT DIFFERENCIE".

Définissons d'abord ces termes :

PEDAGOGIE PAR OBJECTIFS : Les objectifs de l'enseignement sont définis le plus clairement possible, en termes de SAVOIR et de SAVOIR-FAIRE que l'élève doit acquérir, et dont il devra faire la preuve à un moment ou à un autre. Concrètement, il s'agit de répondre à la question : qu'est-ce que l'ELEVE DOIT savoir à la fin de l'année ? plutôt qu'à la question : qu'est-ce que le MAITRE doit avoir dit pendant l'année ? Chacun des objectifs particuliers est défini de telle façon qu'à tout moment on puisse dire si OUI ou NON l'élève a atteint cet objectif. On distingue entre OBJECTIFS MINIMA qui devraient être atteints par tous les élèves d'une classe et OBJECTIFS d'APPROFONDISSEMENT qui concernent les élèves manifestant une certaine aisance.
Les objectifs pris en compte sont, dans la mesure du possible, portés à la connaissance des personnes intéressées : élèves, parents, éducateurs. Une "note" attribuée à un élève prend plus de sens si l'on peut la rapporter à l'objectif où la classe d'objectifs que l'on était censé contrôler, ainsi qu'à la tâche proposée pour vérifier que l'objectif a été atteint.

PEDAGOGIE DE MAITRISE : On cherche à favoriser des apprentissages installés de façon durable et parfaitement dominés. Il ne s'agit pas d'avoir beaucoup de connaissances éparses et mal assurées ; mais des connaissances en nombre peut-être moins grand, sur lesquelles on puisse compter.

PEDAGOGIE DE LA REUSSITE : L'importance de l'échec temporaire est minimisé. L'échec sert de révélateur des difficultés rencontrées par les élèves, il est considéré comme une étape parfois nécessaire vers la réussite, il permet d'aider l'élève à rectifier ses erreurs. Pour un objectif déterminé, l'observation d'une réussite efface pratiquement les échecs antérieurs.

ENSEIGNEMENT DIFFERENCIE : Il s'agit de mieux prendre en compte les différences existant entre les élèves. Les enfants n'ont pas tous les mêmes qualités (ni les mêmes défauts !), ni la même forme d'esprit : tel est plus à l'aise dans le concret, tel sera accessible à un raisonnement abstrait. Dans un enseignement différencié, on cherche à s'appuyer sur ces différences plutôt qu'à les nier, dans le but affirmé d'obtenir le meilleur de chacun des élèves. Il ne s'agit donc, en aucune façon d'une "PEDAGOGIE DE LA FACILITE" ou du "LAISSER FAIRE".

L'EVALUATION.

La pédagogie par objectifs peut difficilement se satisfaire de l'évaluation traditionnelle. Lors de la correction d'un devoir par exemple, il conviendra de rechercher les indices prouvant la présence, ou l'absence, de maîtrise par rapport à un ou plusieurs objectifs particuliers. Si le devoir porte sur un seul objectif, les seules notes véritablement informatives ne peuvent être que 20 ou 0, ce que nous préférons appeler REUSSITE ou ECHEC. Si le travail recouvre plusieurs objectifs, c'est chaque question qui devrait être notée de cette manière. Bien sûr, on peut alors envisager de faire une moyenne pour attribuer une note globale à l'ensemble du devoir, mais dans ce cas, l'élève risque d'oublier les notes partielles et leurs significations pour ne retenir que la note globale. On connait la tendance (naturelle !) des élèves qui consiste à

s'intéresser d'abord et surtout à leurs notes, les considérant trop souvent comme une fin, et non comme un moyen leur permettant de mieux se connaitre, pour les aider à progresser :

- J'ai 10 ! j'ai la moyenne ! ça me suffit.
- J'ai 15 ! c'est une bonne note (et je ne cherche pas à combler mes lacunes).
- J'ai 5 ! de toutes façons, je suis nul...

En réalité, une note sur 20 est toujours un résumé, un condensé de nombreuses informations. Ce résumé peut être souhaité par les familles à divers moments de l'année, il est nécessaire dans certains cas : dossier du brevet, dossier d'orientation...
Une bonne connaissance du niveau et des capacités de l'élève suppose cependant que l'on puisse remonter aux informations qui lui ont donné naissance.

LA FICHE INDIVIDUELLE.

Cette fiche permet de visualiser les acquisitions de l'élève. Le professeur en possède une par élève, qu'il met à jour au fûr et à mesure. L'élève de son côté a sa propre fiche qui est le double de celle tenue par l'enseignant. Cette fiche sert en particulier de fiche de liaison avec les familles.

On remarquera sur la fiche plusieurs tableaux :

- Un premier tableau concerne l'évaluation "objective" des objectifs particuliers. Chaque case correspond à un micro-objectif. Par exemple, en troisième, la case A6 concerne l'objectif : "L'élève sait utiliser une table de carrés pour y rechercher les racines carrées...". Chaque quart de case coloré en rouge témoigne de l'enregistrement d'un échec pour cette question. Lorsque la partie restante de cette case (qui peut être la case entière) est colorée en vert, cela signifie que l'objectif a été atteint.

- La première colonne concerne l'évaluation des PREREQUIS. Il s'agit de vérifier que l'élève a les connaissances nécessaires pour suivre avec profit l'enseignement qui va suivre. Bien sur si tel n'était pas le cas, les développements ultérieurs ne manqueront pas d'en tenir compte.

Les objectifs d'approfondissement sont proposés aux élèves les plus rapides, ils mettent en jeu des savoirs et savoir-faire dépassant le savoir minimum défini pour le "niveau" considéré. Insistons sur le fait que tous ces objectifs sont définis pour un niveau : classe de sixième, classe de cinquième etc... et non pour telle ou telle classe particulière qui peut être plus ou moins forte. En particulier, les objectifs minima sont les mêmes pour tous les élèves du même niveau. Bien entendu, ces objectifs sont conformes aux instructions officielles.

- Un sous tableau concerne les tests de validation. Ces épreuves recouvrent une famille d'objectifs, elles ont pour objet de vérifier que les savoirs acquis le sont durablement et peuvent être mobilisés pour une tâche nécessitant la reconnaissance de questions particulières dans un ensemble assez vaste de questions. Ces tests sont notés sur 100, il s'agit en fait d'un SCORE ou pourcentage de questions que l'élève a été capable de résoudre. On considère qu'une telle épreuve est REUSSIE lorsque l'élève a un taux de réussite supérieur ou égal à 70 %. En cas d'échec, le test est repassé plus tard sous une forme légèrement différente (tests BIS ou TER).

- Un autre tableau concerne les TESTS RENFORCES. Il s'agit d'épreuves se situant au delà des savoir minima et cherchant à mettre en oeuvre les capacités d'adaptation de l'élève à des situations nouvelles nécessitant le TRANSFERT des connaissances et supposant de plus une bonne intégration de celles-ci.

- Le tableau "TESTS ACTIVITES" concerne aussi des épreuves visant à contrôler l'intégration des connaissances, mais ces tests portent moins sur les connaissances elles-mêmes qui se situent en-deçà des savoir minima, que sur la capacité à les mettre en oeuvre dans des situations plus ou moins complèxes.

- Le tableau "OBSERVATION CONTINUE DES ATTITUDES ET DES APTITUDES" concerne une évaluation plus "subjective" du comportement de l'élève et de certaines de ses qualités. Le système de codage utilisé, des flèches plus ou moins inclinées permet de suivre l'évolution de l'élève au cours de l'année.

- Enfin, le tableau "NOTES TRIMESTRIELLES" : On y trouvera les notes figurant sur le bulletin trimestriel. Deux notes sont prévues : une note de niveau résumant la partie objective de l'évaluation, une note dite d'APPLICATION résumant la partie subjective de cette évaluation et prenant plus particulièrement en compte la bonne volonté de l'élève et la qualité de son travail personnel, ceci indépendamment de son niveau réel. Ces deux notes sont à examiner ensemble, chacune d'elles apportant un éclairage particulier à l'autre.

Pour des raisons matérielles, il n'est pas possible, actuellement, de diffuser largement les listes d'objectifs pris en compte pour l'évaluation. Ces listes (ou référentiels) existent et peuvent être communiqués sur demande. D'une façon plus pratique mais moins rigoureuse, on pourra se faire une bonne idée des objectifs poursuivis en consultant les épreuves ayant servi à les contrôler. Ces épreuves sont toujours rendues aux élèves et ceux-ci doivent les conserver soigneusement. Ils peuvent s'en servir utilement pour faire le point de leurs connaissances, pour vérifier leurs progrès et pour préparer le cas échéant les épreuves bis ou ter qu'ils seront amenés à passer.

I. R. E. M. DE BESANCON

NOM :

CLASSE :

FICHE INDIVIDUELLE D'EVALUATION EN MATHEMATIQUES

OBJECTIF CONTRÔLÉ		TEST 1			TEST 2			TEST R			OBSERVATIONS
		A	B	C	A	B	C	A	B	C	
A											
B											
C											
D											
E											
F											
G											
H											
I											
J											

LÉGENDE :

A — DATE À LAQUELLE EST PASSÉ LE TEST .

B — INDICE DE RÉUSSITE I : QUOTIENT DU NOMBRE DE BONNES RÉPONSES PAR LE NOMBRE DE QUESTIONS POSÉES .

C — CODE COULEUR : { VERT SI $I \geqslant 2/3$; VERT HACHURÉ SI $1/2 \leqslant I < 2/3$; ROUGE SI $I < 1/3$; ROUGE HACHURÉ SI $1/3 \leqslant I < 1/2$

N.B. POUR CHAQUE OBJECTIF, LE TEST R EST D'UN NIVEAU PLUS ÉLEVÉ QUE LES TESTS 1 ET 2 QUI SONT ÉQUIVALENTS .

TROISIEME PARTIE

Nous avons rassemblé dans cette partie les articles sur l'évaluation qui ont été publiés ces dernières années dans le bulletin de l'I.R.E.M. de BESANCON.

I) LE COMPORTEMENT DE L'EVALUATEUR

Au cours des stages "Evaluation", le premier travail a consisté à mettre en évidence le peu de fiabilité de l'évaluation traditionnelle, puis à envisager quelques remèdes permettant une certaine modération des divergences observées.

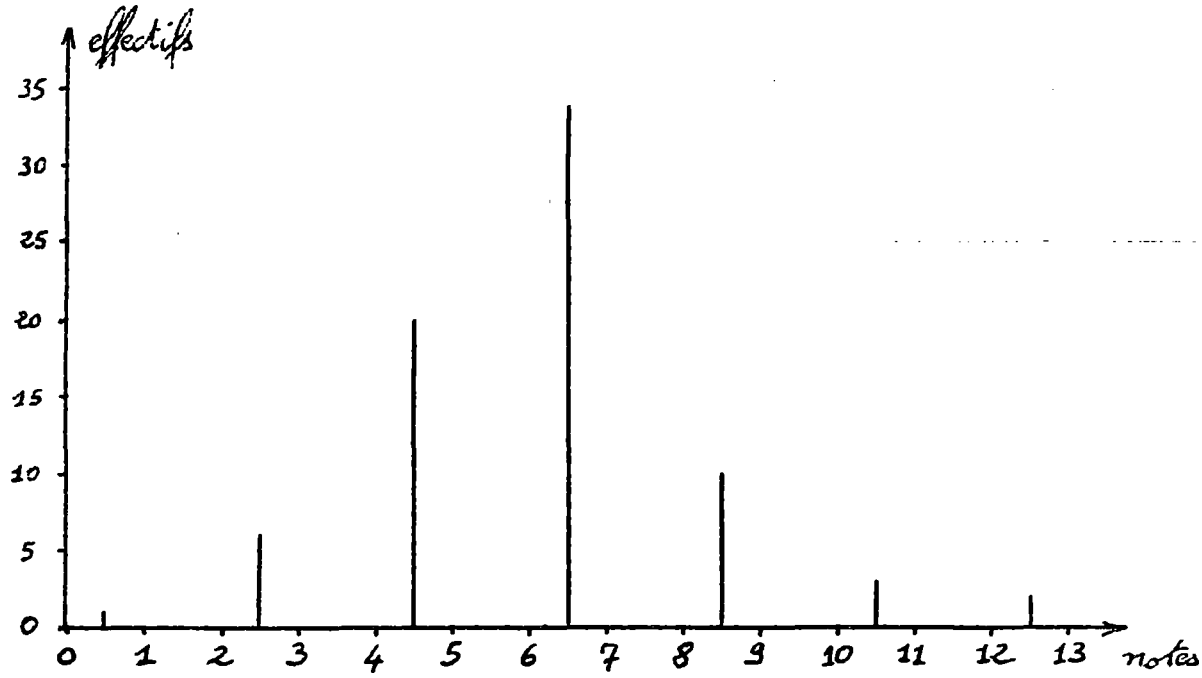
Nous allons à présent tenter d'expliquer ces divergences en étudiant le comportement de l'évaluateur:

Une expérience conduite en 1936 a consisté à faire corriger une copie de Français par 76 correcteurs. La distribution des notes est reproduite ci-dessous:

Enquête de Laugier & Weinberg

note	0-1	2-3	4-5	6-7	8-9	10-11	12-13
nombre d'évaluateurs	1	6	20	34	10	3	2

Histogramme:



Cette distribution en forme de cloche est caractéristique de phénomènes dont les causes de fluctuations sont nombreuses, indépendantes les unes des autres et de faible amplitude. Nous allons mettre en évidence certaines de ces sources qui conduisent à des variations systématiques.

1)Le cadre conceptuel

Pour parvenir à son évaluation, le correcteur extrait de la copie un certain nombre d'indices correspondant à des critères du type, par exemple: richesse des idées, correction du style, s'il s'agit d'une dissertation. Pour cela il compare la production de l'élève à un modèle de référence.

Ce modèle de référence d'une part, est constitué antérieurement à l'évaluation et d'autre part se modifie au fur et à mesure que la tâche d'evaluation se poursuit, si bien que souvent, la fin d'un paquet de copies n'est pas corrigée comme l'a été le début: on constate des décalages si l'on prend soin de recorriger les premières copies après avoir corrigé les dernières.

Si le modèle est susceptible d'évolution chez le même évaluateur, on peut supposer, à plus forte raison, que des évaluateurs différents ne se réfèrent pas au même modèle, ce qui peut être un début d'explication aux divergences observées.

Il faut donc pousser plus loin l'étude du modèle de référence: il est constitué tout d'abord par le produit-norme (dans le cas d'une dictée le texte écrit sans faute d'orthographe). Ce produit-norme est plus ou moins facile à définir selon les disciplines. Mais entrent aussi dans le modèle les produits attendus, dont le correcteur estime l'apparition vraisemblable.

En effet, le même sujet de philosophie peut être donné au baccalauréat ou à l'agrégation, et l'évaluateur ne s'attendra pas à rencontrer les mêmes produits. Il en va de même s'il dispose d'informations sur les auteurs de chaque copie et ses attentes diffèreront suivant que cet auteur est connu comme bon ou mauvais. Enfin si le correcteur ne rencontre que des bonnes copies dans la première partie du lot, la probabilité que de graves erreurs aient été commises va diminuer à ses yeux.

Le troisième élément constitutif du modèle est une échelle de référence qui se traduit par les repères ou barêmes choisis par l'évaluateur.

Pour préciser ces notions, nous allons maintenant relater cinq expériences qui mettent en évidence les différences entre les modèles de référence utilisés par des évaluateurs corrigeant le même lot de copies mais ne disposant pas des mêmes informations:

2) Les effets d'assimilation

En 1975 une expérience a porté sur une version anglaise de niveau bac. Sur chaque copie figurait une note censée correspondre à un devoir antérieur et, suivant les évaluateurs, cette copie était fictivement attribuée autant de fois soit à un élève ayant obtenu une note antérieure forte, soit à un élève ayant obtenu une note antérieure faible. Les résultats ont été très nets: la moyenne des notes données aux mêmes copies quand la note antérieure est forte dépasse de 2 points celle obtenue quand la note antérieure est faible. Cette différence est presque la même pour toutes les copies et elle est largement significative puisqu'elle a moins de 5 chances sur 1000 d'être due au hasard.

Toujours en 1975 une expérience a porté sur des devoirs de Sciences Naturelles du baccalauréat. Sur chaque copie était indiquée une série de notes fictives, jouant le rôle des notes portées sur les livrets scolaires des candidats. Les différences entre les moyennes des notes données par les correcteurs aux mêmes copies varient un peu moins, mais encore dans le même sens qu'au cours de l'expérience précédente, selon les notes fictives portées sur les copies. Mais cette fois-ci le correcteur devait extraire l'information d'une série de notes et l'on a observé qu'il confirmait la progression des performances antérieures de l'élève, et ce d'autant plus que cette progression lui paraissait plus crédible: un élève censé avoir obtenu 12 puis 14 était davantage surévalué qu'un autre qui était passé de 8 à 18.

En 1972 une autre expérience a porté sur des rédactions françaises de niveau moyen censées avoir été produites tantôt par des élèves d'une sixième de type 1, tantôt par des élèves d'une sixième de type 3. En moyenne les copies ont été notées avec un point et demi de plus quand elles étaient attribuées à des élèves de type 1, cette différence allant dans le même sens pour toutes les copies. Notons qu'on aurait pu s'attendre ici à l'effet inverse car l'évaluateur à qui l'on présente des rédactions françaises censées avoir été composées par des élèves d'une 6e de type 3 pourrait considérer que pour ce type d'élèves les copies sont d'un bon niveau et leur attribuer de ce fait des notes supérieures aux notes données lorsqu'elles sont censées provenir d'élèves d'une 6e de type 1, mais visiblement c'est le contraire qui s'est passé.

Le recrutement des élèves selon les différentes filières n'étant pas indépendant de leur origine socio-économique, il convient de décrire d'autres expériences portant sur des caractériatiques autres que proprement scolaires:

En 1975 ont été corrigés des devoirs de Sciences Naturelles de baccalauréat censées provenir tantôt d'un lycée de quartier résidentiel, le lycée Janson-de-Sailly dans le XVIe, tantôt d'un lycée de banlieue ouvrière, le lycée Marcellin-Berthelot à Pantin. Dans l'ensemble les copies sont surévaluées d'un demi-point quand elles sont attribuées aux élèves du lycée Janson-de-Sailly, mais cette fois la tendance n'est pas générale, les évaluateurs s'étant comporté différemment selon leur statut et leur ancienneté.

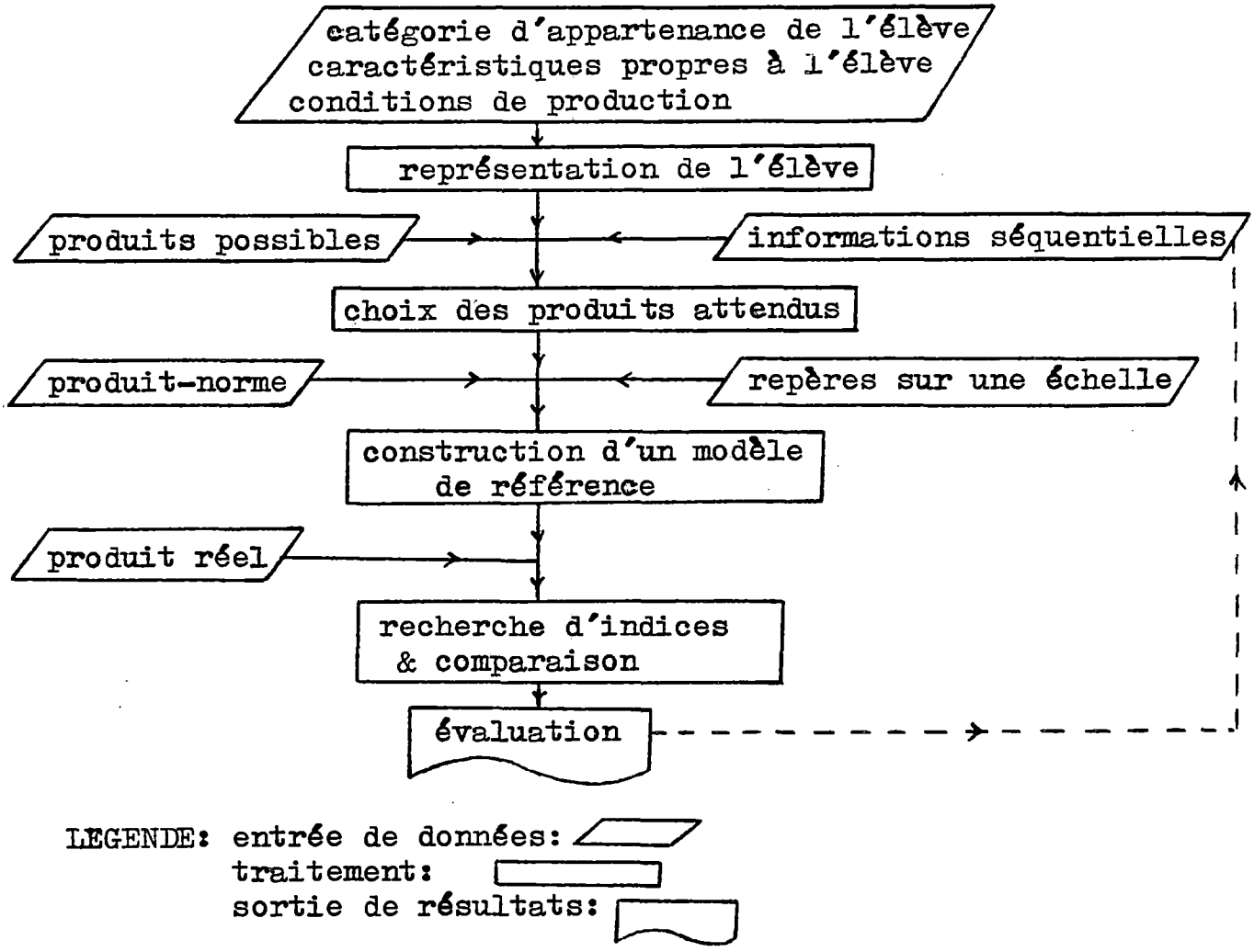
Enfin, en 1975, des rédactions françaises ont été attribuées tantôt à des élèves dont le nom était à consonnance française, tantôt à des élèves dont le nom était à consonnance étrangère. Ici on n'observe globalement aucune différence selon l'origine ethnique fictive des élèves. Par contre les évaluateurs dont le nom est à consonnance étrangère se montrèrent plus indulgents de 1 point que leurs collègues, et surévaluèrent les élèves dont le nom était à consonnance française tandis que les évaluateurs dont le nom était à consonnance française surévaluaient les élèves dont le nom avait une consonnance étrangère. Cette interaction, statistiquement significative, montre que les évaluateurs ont bien été influencés par la pseudo-origine ethnique des élèves, mais que cette influence dépend de leur propre origine. Il s'agit ici d'une question psychologique complexe relative aux problèmes de coopération ou de conflits dans des situations de groupe et nous n'entrerons pas dans les détails.

A la lumière de ces cinq expériences une conclusion s'impose: les mêmes copies sont évaluées différemment lorsqu'on donne aux évaluateurs des informations différentes sur les élèves censés les avoir rédigées.

Comment expliquer ce phénomène ? Voici une hypothèse: les informations à priori induiraient chez l'évaluateur une représentation de l'élève lui-même ou de telle ou telle de ses catégories d'appartenance, ce qui provoquerait un certain nombre d'attentes, c'est-à-dire que le correcteur ne sélectionnerait pas les mêmes produits attendus et ne construirait pas le même modèle de référence. Il ne lirait alors pas la copie de la même façon et ne recueillerait pas les mêmes indices, comme s'il tendait à éviter ceux susceptibles de contredire son attente. De cette attitude pourrait provenir la dépendance observée, qui est étrangère au contenu des copies, et qui se manifeste par une assimilation de l'évaluation suivante au niveau présumé de l'auteur de la copie, comme si le correcteur essayait de réduire la distance entre la note qu'il allait mettre et le niveau scolaire que les informations a priori permettent d'inférer.

En résumé, l'attente suscitée par la représentation du niveau de l'élève se traduirait chez l'évaluateur par une attitude différente dans la recherche des indices qu'il prend dans la copie.

"organigramme" du comportement d'évaluation:



Peut-on, dans ces conditions, défendre l'efficacité du contrôle contimu ? Quel rôle joue, en fait, le livret scolaire lors du baccalauréat ? Voilà des questions auxquelles les expériences précédentes apportent quelques éléments de réponse. Mais elles peuvent aussi rendre compte de la constance de la trajectoire des élèves: si les bons restent bons, et les mauvais restent mauvais, c'est certainement d'abord à cause de la stabilité générale de leurs capacités, mais une part non négligeable de cette constance tient à la dépendance entre évaluations.

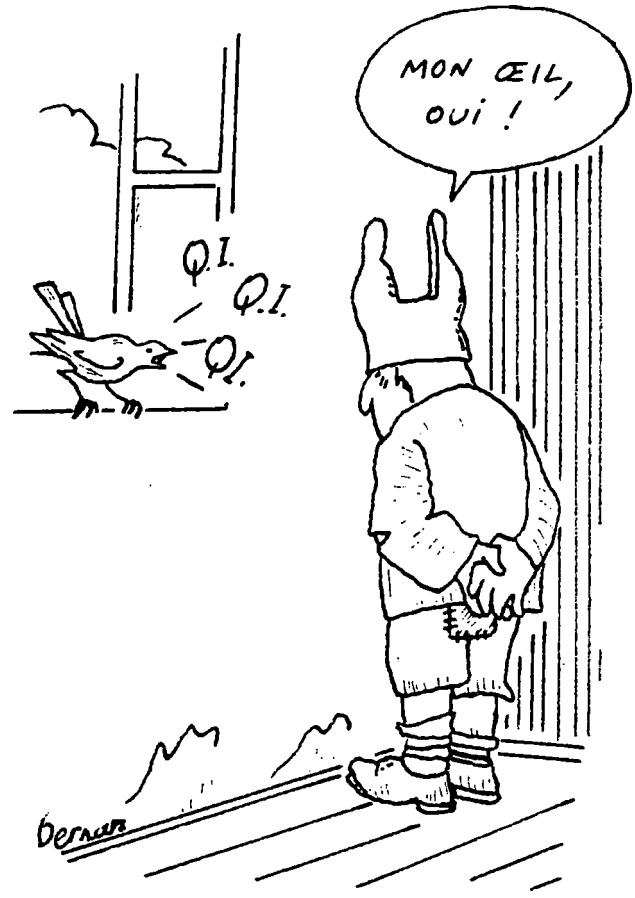
Avant d'aller plus loin, nous ouvrons une parenthèse sur une expérience qui va nous montrer que les représentations et les attentes dont nous venons de parler font sentir leur poids non seulement au niveau de l'évaluation, mais tout au long de la scolarité:

3)L'effet Pygmalion

L'expérience a été construite dans le but de vérifier la validité de l'hypothèse selon laquelle, dans une classe donnée, les enfants dont le maître attend davantage font effectivement des progrès plus grands. L'expérience s'est déroulée dans une école primaire d'un quartier pauvre de San Francisco de 1964 à 1966: on a choisi par tirage au sort 20% des élèves de l'établissement et on a dit à leurs maîtres qu'un test psychologique nouveau permettait de prédire qu'ils étaient à la veille de progrès rapides. Huit mois plus tard, on constatait chez ces enfants-miracle un démarrage scolaire réel que confirmait une augmentation très significative de leur QI par rapport à leurs camarades.

Au cours de l'année expérimentale, les enfants du groupe-témoin gagnèrent 8 points au QI tandis que ceux du groupe expérimental en gagnèrent 12, dont 7 points aux seules épreuves de raisonnement qui consistaient, pour chaque question, à dire lequel parmi cinq dessins diffère des quatre autres. Un tel progrès n'a qu'une probabilité de 0,02 d'être dû au hasard. C'est au CP et au CE1 que les effets des prophéties des maîtres furent les plus extraordinaires, ce qui semble confirmer que plus les enfants sont jeunes, plus ils y sont sensibles. On peut remarquer aussi que les enfants ont davantage bénéficié des préjugés des maîtres dans le domaine intellectuel où ils étaient plutôt légèrement avantagés au départ, en l'occurence les filles dans le domaine du raisonnement et les garçons dans le domaine verbal (où le test consistait, pour chaque question, à relier les illustrations aux descritions verbales données par l'examinateur).

Si l'on se fie aux évaluations des maîtres, on trouve que l'avantage du préjugé en ce qui concerne la lecture est similaire à celui trouvé au QI. Il en est de même en calcul pour les deux plus petites classes, et ce sont les enfants des sections moyennes qui en profitent plus que leurs camarades des sections lentes et rapides. Par rapport à des tests de lecture plus objectifs, il semble que les enfants du groupe expérimental aient été sanctionnés plus sévèrement par les maîtres que ne l'ont été les autres enfants. Il se pourrait même que ce soit précisément ce type de cotation qui soit pour une part responsable des effets des préjugés favorables. Il peut y avoir ici un processus cyclique avantageux: plus un maître obtient, plus il espère et plus il espère, plus il obtient.



L'expérience a duré deux ans , ce qui a permis de préciser que les préjugés favorables commencent à rapporter quelque chose aux enfants dès le 4e mois. Le maximum est atteint à la fin de l'année scolaire. Les élèves changent ensuite de classe et ont un nouvel instituteur qui n'a reçu aucun renseignement spécial, mais les effets observés l'année précédente sont encore, 20 mois après le tirage au sort, supérieurs à ceux constatés au bout d'un semestre. Ce sont les élèves les plus âgés qui maintiennent le plus longtemps, de façon autonome, l'avantage acquis.

Ajoutons que les effets ont été légèrement plus importants pour les élèves mexicains de l'école que pour les élèves américains, ce qui renforce, étant donné que les maîtres étaint américains, ce que nous avions observé dans l'expérience sur les noms à consonnance française ou étrangère. La mention de ces élèves sur une liste de "démarreurs" probables a sans doute surpris leur maître chez lequel l'intérêt a pu succeder à la surprise.

Nous ne savons pas comment l'espoir que fonde le maître sur le développement intellectuel d'un élève est communiqué à ce dernier. Peut-être les maîtres traitèrent-ils leurs élèves de manière plus enthousiaste et il existe de nombreuses études montrant qu'un examinateur plus chaleureux déclenche souvent un comportement plus intelligent de la part du sujet (ROSENTHAL,R., Experimenter effects in behavioral research,Appleton, New York, 1966). Les maîtres ont pu laisser les élèves signalés donner plus de réponses soit en renforçant plus rapidement leurs réponses correctes, soit en leur laissant faire des réponses erronées qui, petit à petit, étaient ramenées sur la bonne voie, ou encore par les deux procédés à la fois.

Mais ce qui est important ici, c'est d'avoir apporté la preuve expérimentale que les préjugés d'une personne sur le comportement d'une autre peuvent devenir des prophéties à réalisation automatique: en bref, les gens évoluent comme on les traite. Et cela a des implications sur la politique éducative elle-même: si les écoles normales se mettaient à enseigner ce principe, une nouvelle espérance pourrait naître, à savoir que les enfants sont susceptibles d'apprendre plus qu'on ne le croit.

Avec une telle espérance, les maîtres ayant affaire à des enfants désavantagés pourront difficilement dire qu'ils ne sont pas éducables.

Voici donc la fin de cette longue parenthèse sur le pouvoir des attentes des maîtres.

Nous allons revenir au comportement d'évaluation et nous attacher maintenant à son déroulement dans le temps, puisque, lorsqu'on corrige des copies, on les corrige les unes après les autres, de même que, quand on a à évaluer des exposés oraux, on les entend les uns après les autres.

4)Les effets de contraste

Au cours d'une expérience réalisée en 1965, des versions anglaises de classe terminale ont été corrigées dans l'ordre 1 à 26 puis dans l'ordre 26 à 1 par des enseignants différents. Afin de rendre leurs résultats comparables, on a opéré une réduction en moyenne et en écart-type des notes de chaque évaluateur. En premier lieu les évaluateurs ont manifesté une tendance à surévaluer de presque 1 point les copies corrigées dans le premier tiers ou la première moitié. On note cependant une exception en ce qui concerne la première copie, due sans doute au fait que le correcteur ne pouvait la comparer qu'au produit-norme: comme elle contenait nécessairement des erreurs, elle s'est trouvée sous-évaluée par contraste. Dès la seconde copie le correcteur pouvait faire référence à la précédente, et ce sont encore des effets de contraste que l'on rencontre: la même copie est surévaluée quand elle vient après une copie faible et sousévaluée quand elle suit une copie forte. On rencontre ainsi de la 1º à la 26º copie 18 effets de contraste dont 11 particulièrement importants.

Une nouvelle expérience conduite en 1971 a consisté à comparer la notation d'un lot de copies d'Anglais et la notation de ce même lot après l'introduction de une à trois copies successives toutes très bonnes ou toutes très mauvaises et qui jouaient ainsi un rôle d'ancre. On appelle "ancre" en psychologie de la perception, les objets ayant une correspondance immédiate avec une réponse, notamment parce qu'ils correspondent à l'une des réponses extrêmes. C'est grâce à de telles ancres que la tâche d'évaluation se transforme en tâche de comparaison. La même copie est notée autour de 9 quand elle suit une ancre basse et autour de 6 quand elle suit une ancre forte. La différence atteint 4 points si l'ancre est constituée de trois copies. Les effets continuent à se manifester, mais de façon atténuée, sur les trois ou quatre devoirs qui suivent, et ils sont d'autant plus forts que les ancres sont situées dans la première moitié du paquet de copies, c'est-à-dire tant que le système d'évaluation n'est pas encore stabilisé.

La contradiction entre les effets d'assimilation provoqués par les informations a priori et les effets de contraste provenant des informations séquentielles n'est qu'apparente: en effet le modèle de référence constitue une mise en correspondance entre les produits attendus et certains repères sur l'échelle de notation, ce qui constitue un système d'ancres permettant d'évaluer par comparaison. Et les productions des élèves, nous l'avons vu, font généralement l'objet d'une assimilation aux attentes de l'évaluateur. Mais ,dans le cours de l'acte d'évaluation apparaîtront des copies extérieures aux autres, nettement meilleures ou moins bonnes que celles notées jusque-là. Ces copies provoqueront un ancrage qui interfèrera avec le système initial. Pour pouvoir évaluer ces copies, le correcteur doit déplacer l'échelle, ce qui entraîne la sur ou sousévaluation des copies suivantes. C'est l'assimilation forcée aux attentes des copies extrêmes qui donne les effets de contraste observés.

Tel est le modèle analogique sur lequel d'autres expériences peuvent se construire. Il apparaît ainsi que l'évaluateur ne recueuille dans les copies que les indices correspondants à une attente qui est elle-même variable. Il reste donc à étudier comment est extraite l'information contenue dans les copies:

5)Le recueil des indices

Tout d'abord, le fait qu'il n'existe aucune corrélation entre les critères que les enseignants déclarent spontanément utiliser et ceux qu'ils choisissent, par ailleurs, quand on leur présente une liste de critères nous incite à nous méfier de ce qu'ils disent sans pour autant mettre en cause leur bonne foi. Egalement, le fait qu'ils aient tendance à surévaluer d'environ un point les élèves dont l'origine sociale est comparable à la leur, en l'abscence de toute indication portée sur les copies montre l'existence de critères implicites, à côté de ceux que les enseignants déclarent utiliser.

Nous devrons donc étudier expérimentalement les processus de recueil des indices. Cela exige la constitution de lots de copies parfaitement descriptibles sur un certain nombre de dimensions telles que correction du style, ou erreurs de calcul. Il devient nécessaire de construire les copies à évaluer pour assurer la présence de toutes les variantes possibles: copie originale, cohérente et correcte, copie originale, cohérente mais incorrecte, etc. Cela est réalisé, en général, à partir de phrases effectivement rédigées par les élèves et qui, de façon isolée, sont considérées par les enseignants comme présentant ou non, de façon significative, telle ou telle incorrection.

Une expérience construite en 1975 a porté sur des rédactions françaises construites selon quatre dimensions: style, cohérence des idées, orthographe, présentation. C'est le style qui provoque l'écart le plus important (4 points), puis la cohérence des idées (2,6), l'orthographe (1,7) et enfin la présentation (0,7). Une autre expérience datant de la même année a concerné des devoirs de mathématiques et ce sont les erreurs de calcul qui entraînent des écarts de 8 points contre 6 pour les erreurs de raisonnement.
Il y a donc une hiérarchie des indices (même si cette hiérarchie peut changer en fonction des devoirs).

Lors d'une expérience réalisée en 1977 sur des copies de Français, parmi les devoirs dont le style était correct, la différence était en moyenne de 0,5 point entre copies cohérentes et copies incohérentes du point de vue de l'ordre des idées. Mais cette différence atteignait 2 points si le style était incorrect.

Inversement, parmi les devoirs pour lesquels l'ordre des idées était cohérent, la différence était en moyenne de 2 points entre les copies correctes et les copies incorrectes du point de vue du style. Mais elle était de 4 points quand l'ordre des idées était incohérent. Par conséquent, contrairement à une opinion répandue, les critères d'évaluation ne sont pas pris en compte indépendamment les uns des autres.

Hiérarchie des indices et dépendance entre les critères résultent peut-être d'un apprentissage des relations les plus fréquentes entre les dimensions des devoirs que les évaluateurs ont l'habitude de corriger.

On peut aussi se demander si, selon l'echelle de notation qu'il utilise, l'évaluateur n'est pas conduit à sélectionner différemment les indices qu'il recueille dans la copie.

Une première expérience a consisté à noter des rédactions de 6º construites sur quatre dimensions tantôt sur 10, tantôt sur 20. Les notes attribuées ont été après coup transformées en notes sur 100. On constate que l'écart dû au style est presque deux fois plus important dans la notation sur 10. De même, l'intervalle moyen pour la notation sur 10 est presque supérieur de moitié à celui de la notation sur 20. Par contre, au cours d'une seconde expérience portant sur des devoirs de mathématiques de 6º construits sur deux dimensions, on ne constate pas les différences précédentes. Il convient de remarquer enfin que les évaluateurs avaient le droit dans les deux expériences d'attribuer des demi-points, et que le nombre de notes différentes est sensiblement le même dans les deux échelles.

Il en résulte que les correcteurs n'ont pas noté en fonction des propriétés formelles de l'échelle de notation, mais plutôt selon une échelle implicite (caractérisable par un nombre donné de notes et un intervalle moyen entre deux notes successives) à laquelle ils ont adapté ensuite l'échelle qui leur était imposée. Il est clair que cette adaptation était insuffisante dans le cas des copies de Français, le type de production et le nombre de critères pris en compte jouant certainement sur l'utilisation de l'outil d'évaluation.

Enfin la lecture d'une copie exige une certaine durée, durée qui permet à des effets de s'additionner ou de se contrebalancer. Il est probable que les premiers indices recueillis provoquent des attentes qui guident le recueil des indices suivants (de ce point de vue, le rôle des informations a priori peut se comprendre comme étant de provoquer une attente préalable au recueil d'indices). Dès lors la question se pose de savoir si le choix des indices, et l'importance qui leur est accordée diffère ou non selon leur place dans la copie.

Une première recherche a porté sur un lot de rédactions françaises construites sur deux dimensions, le style et l'orthographe. Les copies comportaient sur l'un ou l'autre trois ou six incorrections placées toutes soit dans la première moitié du devoir, soit dans la seconde. Les copies reçurent alors des notes moins élevées d'un point en moyenne lorsque les incorrections étaient placées dans la première moitié du devoir, cette différence étant d'autant plus marquée que le nombre de fautes était plus grand. Cet effet de place des indices négatifs, peu conforme à l'équité que l'on attendait, traduit un phénomène similaire à celui mis en évidence par les expériences où étaient manipulées des informations sur les auteurs des devoirs.

Une dernière expérience a donc été réalisée sur le modèle de la précédente, mais dans laquelle les évaluateurs disposaient en plus d'une information quant au niveau scolaire des élèves censés avoir rédigé les copies: ainsi le début de la copie confirme ou non l'attente initiale induite chez l'évaluateur par le niveau scolaire. En comparant les résultats des deux expériences. il apparaît que l'effet de sous-évaluation dû aux incorrections nombreuses placées au début est annulé lorsque l'information indique un niveau faible, mais inversé lorsqu'est annoncé un niveau fort (comme si l'évaluateur assimilait les incorrections au modèle qu'il se fait du bon devoir). Ce privilège accordé au niveau fort rappelle les résultats obtenus à propos de l'origine socio-économique des élèves et semble confirmer l'existence de processus de compensation jouant, ainsi que tous ceux que nous avons observés, par des distorsions au niveau du recueil des indices.

Ces dernières expériences semblent aussi montrer que les évaluateurs procèdent habituellement par retrait de points (l'echelle de notation étant ainsi parcourue dans un seul sens), c'est-à-dire qu'ils recueillent préférentiellement des indices négatifs suivant une stratégie correspondant à un recueil d'information minimal et à une référence faite au pôle positif (le produit-norme). Un tel résultat signifierait que le comportement de l'évaluateur peut être décrit par la théorie plus générale de l'estimation, bien connue de la psychologie cognitive.

6)Conclusion

La conclusion générale est que l'évaluation scolaire relève de déterminants systématiques qui ne réfèrent pas tous à des caractéristiques proprement scolaires. Par conséquent l'évaluation ne constitue pas une capacité qu'une bonne maîtrise de la discipline suffit à assurer, si bien qu'une formation des enseignants limitée à cette maîtrise est nécessairement incomplète. Et comme l'évaluation est un comportement qui répond à des déterminants par ailleurs connus dans le cadre plus général de la psychologie de la perception, ce comportement peut être modifié par une action adéquate, c'est-à-dire qu'il est susceptible d'apprentissage. Par exemple, il n'est pas exclu de tenter d'améliorer la fiabilité des évaluations en apprenant à chaque évaluateur à traiter plus objectivement l'information véhiculée par chaque copie, par exemple en traitant indépendamment les critères utilisés.

Mais plus encore que de livrer un contenu, l'importance d'une réflexion sur l'évaluation des productions scolaires est de favoriser une prise de conscience et de susciter une attitude: il faut mettre en route un programme de recherche fondamentale sur l'évaluation tout en demeurant humblement persuadé du côté humain, et par conséquent relatif et imparfait, de toute note que l'on pourra attribuer.

Exposé le 5 Février 1982 au cours du Colloque "Evaluation"

Condensé de:

NOIZET,G., et CAVERNI,J.-P., Psychologie de l'évaluation scolaire, Paris, Presses Universitaires de France, 1978, 231p., deuxième partie: l'analyse expérimentale du comportement d'évaluation, pp. 61 à 146

ROSENTHAL,R. et Lenore JACOBSON, Pygmalion à l'école, Tournai, Casterman, 1971, 293p.

Jean CESAR

II)

E V A L U A T I O N

A propos d'une expérience docimologique

La plupart des travaux de docimologie que nous connaissons se situent dans un contexte d'examen ou de concours. Ces travaux ont montré de façon convaincante à quel point la note d'un élève pour un examen donné, et dans une matière donnée (y compris les mathématiques) dépendait :

1 - Du sujet de l'examen (qui peut affirmer que tel élève qui a eu 12 à l'épreuve de mathématiques du BEPC 1980 n'aurait pas eu 8 à celle proposée en 1979 ?).

2 - De l'élève, ce qui est heureux ! Mais nous voulons parler ici de ses conditions physiques et affectives au moment de l'examen.

3 - Du correcteur, à un point tel que PIERON a pu écrire dans "examens et docimologie" que : "Pour prédire la note d'un candidat (au baccalauréat) il vaut mieux connaitre son examinateur que lui-même".

Dans le but d'atténuer les conséquences du dernier point cité, la docimologie classique a proposé un certain nombre de "procédures de modération" : sujet national ou académique, barême de notation, multicorrection...

Au niveau du collège en particulier, le contexte a changé, l'évaluation continue y ayant complètement remplacé les traditionnelles compositions et examens, y compris le BEPC puisque le nouveau brevet des collèges sera attribué à partir de l'évaluation continue de l'élève. On peut penser que cette méthode viendra diminuer les effets des points 1 et 2 mentionnés ci-dessus, cela est possible mais il nous semble qu'une étude plus fine s'impose. Quoi qu'il en soit il n'est pas dans notre propos de condamner l'évaluation continue, elle a certainement ses avantages, mais nous voulons mettre en évidence quelques uns de ses défauts. Dans ce qui suit nous ne nous intéresserons qu'au troisième point. Que se passe-t-il en fait ? Le professeur-correcteur est à l'origine et à l'extrémité du processus d'évaluation : maître à la fois du sujet et du barême, des critères pris en compte, du moment de la passation, de la préparation plus ou moins poussée (cela va du collègue un peu anxieux des résultats de ses élèves qui prépare l'épreuve avec eux quelques jours avant à celui qui pour éprouver les capacités d'adaptation de ses élèves leur soumet un sujet tout à fait original).

Parmi les défauts qui nous semblent affecter l'évaluation continue, citons :

- L'absence de consensus concernant la notation. Le professeur, l'élève, les parents, l'administration accordent-ils la même signification aux notes ? Et qu'en est-il des enseignants entre eux ? Or, mettre une note c'est poser un acte social, la note n'appartient pas à la personne qui la met, ni d'ailleurs à celle qui la reçoit.
- Le risque de l'examen continu. Au niveau du collège, compte tenu du nombre de disciplines enseignées, l'élève est sans cesse sollicité, il continue à appeler "colle" la moindre interrogation et se trouve ainsi plusieurs fois par jour devant ce type d'épreuve. Les réactions sont diverses, certains élèves finissent par se désintéresser de cette évaluation d'autres font plusieurs fois par jour leurs moyennes pour y intégrer les dernières notes.
- Les risques de déviations de l'évaluation du niveau des élèves. Consciemment ou non, l'enseignant est tenté d'introduire dans sa notation des éléments qui n'ont pas grand chose à voir avec le niveau des élèves. C'est ce qui se passe lorsque l'on note zéro un devoir non fait ou copié. En sens inverse c'est aussi ce qui se passe lorsque l'on note normalement des travaux faits à la maison. Plus insidieusement c'est ce qui arrive lorsque l'on fait coïncider devoir ou interrogation avec le moment où la classe est le moins capable de fixer son attention sur ce que l'on avait l'intention de faire.
  Bien entendu nous rangeons parmi les déviations, la tendance qu'ont certains élèves à ne jamais produire un résultat qui leur soit personnel. Nous pensons même que cette inclination sévit de façon endémique dans certaines classes, au détriment d'ailleurs du développement des qualités intellectuelles de ces mêmes élèves. Nous devons donc réagir, mais à notre avis, sans utiliser la note de niveau à cette fin. Dans les cas limites une telle note ne pourra pas être attribuée.
- Le manque de fiabilité des notes particulières. Si au niveau des examens, la fiabilité de la note était illusoire, qu'en est-il en ce qui concerne l'évaluation continue ? On peut penser que globalement la situation est améliorée, cela reste à prouver. En ce qui concerne une note particulière à un devoir particulier, il n'y a pas de raison à priori pour que la fiabilité soit accrue.

Dans ce qui précède, nous avons voulu ouvrir le débat, présenter quelques réflexions générales sur le sujet sans chercher à l'épuiser. Dans la suite de cet article, nous présenterons une petite expérience docimologique qui nous semble éloquente.

L'EXPERIENCE

Au cours du stage académique "Evaluation en Mathématiques", six collègues ont été invités à noter douze copies d'une épreuve de géométrie de quatrième. Les correcteurs ont été placés dans les conditions de l'évaluation continue à ceci près qu'ils n'avaient pas choisi le sujet. A quelques réserves près ils ont accepté de considérer que le sujet était valable et qu'ils auraient pu l'utiliser dans leurs classes. D'autre part, ces collègues enseignent tous en quatrième. Ce qui précède implique en particulier qu'il n'y a pas eu de concertation avant ni après la correction, que chacun est resté maître des critères pris en compte et du barême éventuel et avait pour consigne de corriger comme s'il s'était agi de ses propres élèves. Ces copies sont des copies réelles d'élèves réels, le septième correcteur est le correcteur réel.

LE SUJET (en annexe)

Ce sujet n'est pas quelconque, il est le résultat d'une concertation entre les professeurs du collège d'ORNANS. Il s'agissait d'évaluer l'objectif suivant :

- L'élève étant supposé connaitre le théorème concernant la droite et le segment joignant les milieux de deux côtés d'un triangle, celui concernant les parallèles équidistantes, et les propriétés du parallélogramme, il sera capable d'utiliser ces connaissances dans la résolution de problèmes simples, mettant en jeu un seul théorème à la fois.

Il n'est pas dans notre propos d'examiner ici la validité de cet objectif, pas plus que celle du sujet correspondant. Nous devons toutefois signaler qu'à notre avis ce sujet souffre de plusieurs imperfections et qu'il n'est pas question de le proposer comme modèle de ce qu'il conviendrait de faire. Redisons néanmoins qu'il a été considéré comme utilisable tant par ceux qui l'ont élaboré que par ceux qui l'ont corrigé.

LES RESULTATS

Copie / Correcteur	1	2	3	4	5	6	7	8	9	10	11	12	ECART MAXI	MOYENNE
A	13	14	10	8	6	5	10	6	4	3	5	14	11	8,2
B	13	11	13	6	5	6	11	5	5	4	7	11	9	8,1
C	14	12	11	11	6	11	11	8	7	5	11	13	9	10
D	15	14	13	5	5	6	14	4	5	3	5	15	12	8,7
E	13	13	11	6	5	8	10	4	5	6	5	14	10	8,3
F	17	14	9	5	3	17	13	9	3	3	8	11	14	9,3
ECART MAXIMUM	4	3	4*	6*	3	12*	4	5	3	3	6*	4		
Moyenne	14,1	13	11,2	6,8	5	8,8	11,5	6	4,8	4	6,8	13		8,8
CORRECTEUR REEL	15	12	10	7	7	7	9	6	6	7	8	15	9	9,1

Que faut-il penser de telles divergences portant aussi bien sur la note attribuée que sur l'ordre de classement des copies ? Quatre copies sont notées tantôt au dessus, tantôt au dessous de la moyenne. L'étendue de la notation des correcteurs varie de 9 points à 14 points sur le même lot de copies. Certes les moyennes générales "se tiennent", doit-on s'en satisfaire ?

Nous n'avons pas cru utile d'introduire l'arsenal statistique dans cette étude, aucun test de signification ne viendra mettre en défaut le fait que la même copie peut recevoir la note 5 ou la note 17 selon le correcteur.

Cette expérience est évidemment limitée et nous ne prétendons pas apporter des réponses définitives, d'autres expériences seraient nécessaires. Nous espérons simplement avoir contribué à ce que les questions concernant l'évaluation continue soient clairement posées.

Antoine BODIN

o
o o

ANNEXE : Le sujet

Collège d'Ornans — Nom :

Classes de quatrième — Classe :

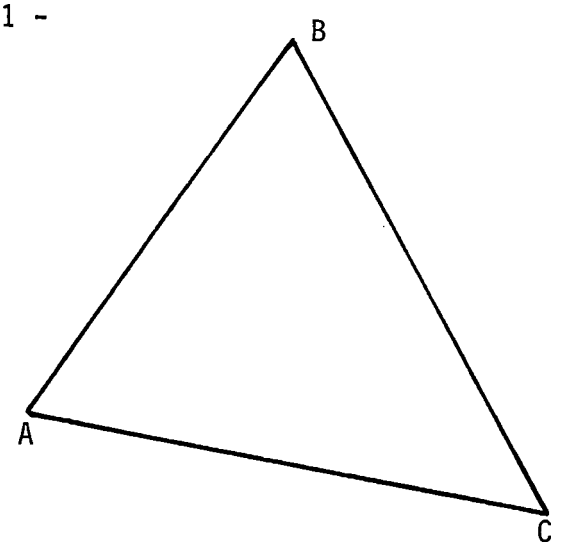
Epreuve n° 1 — Note : /20

Voici quatre exercices, tu devras répondre proprement sur la feuille. Utilise un brouillon pour préparer tes réponses.

S'il t'arrive de manquer de place, continue sur une feuille supplémentaire.

---

1 -



A B C est un triangle quelconque

Soit D le milieu du segment [AC]

Soit E le milieu du segment [AB]

On considère la parallèle à la droite (AC) qui passe par E ; elle coupe [BC] en un point F.

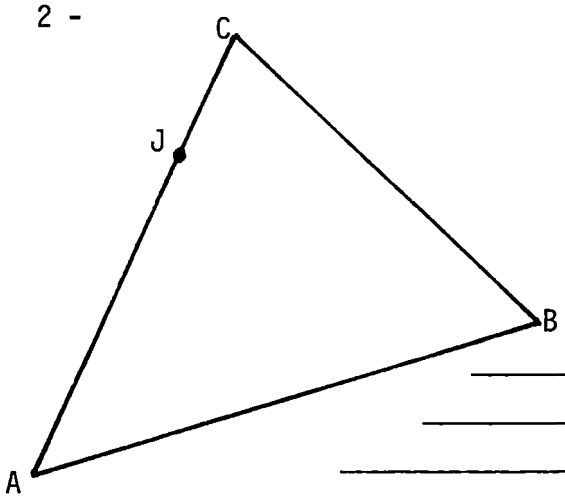
1°) Complète la figure

2°) Démontre que le quadrilatère F C D E est un parallélogramme

3°) Démontre que F est le milieu de [BC] .

---

2 -



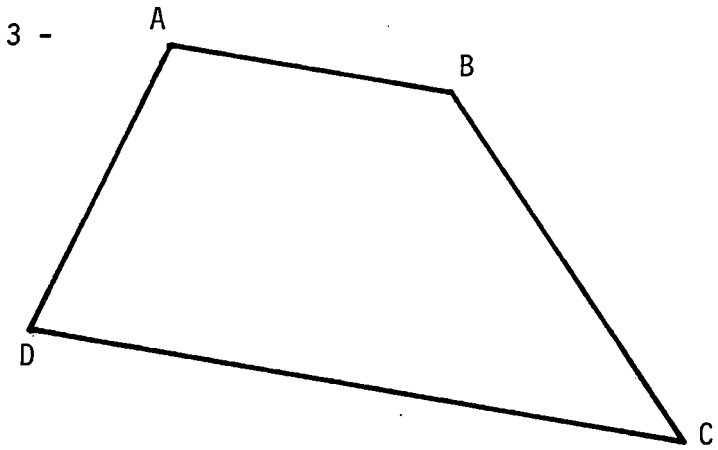
Soit un triangle ABC et sur le côté [AC] un point J.

On appelle I le milieu du segment [AJ] . Les parallèles à AB qui passent par I et J coupent la droite BC respectivement en L et S.

1°) Complète la figure

2°) Démontre que L est le milieu du segment [BS]

3 -



ABCD est un quadrilatère tel que les droites (AB) et (CD) soient parallèles. Soit T le milieu du segment [BD] . La droite (AT) coupe la droite (DC) en un point K.

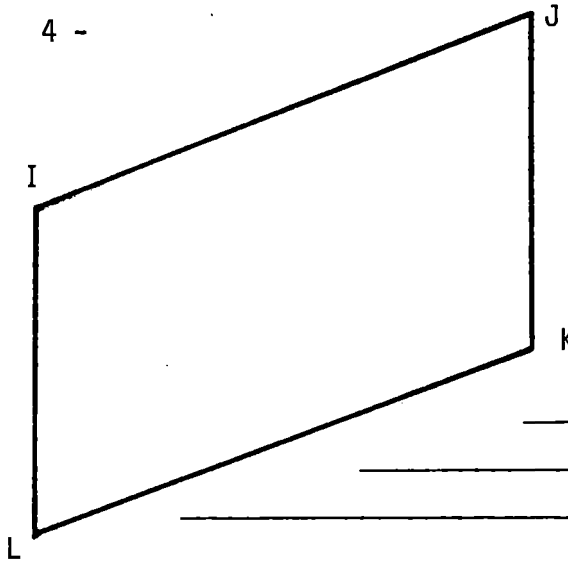
1°) Complète la figure

2°) Démontre que T est le milieu de [AK]

3°) Si tu n'as pas pu le démontrer, tu admettras que T est le milieu de [AK] .
Démontre que le quadrilatère ABKD est un parallèlogramme.

---

4 -



IJKL est un parallélogramme.
Soit M le milieu de [IL] et soit N le milieu de [JK] .

1°) Démontre que le quadrilatère MJNL est un parallélogramme.

2°) Démontre que les segments [IK] , [JL] , et [MN] , se coupent en un même point.

III) LES Q. C. M.

AVANT PROPOS

On trouvera ci-après, pour l'essentiel, le texte de l'exposé qui a été présenté le 5 Mars 1982 dans le cadre du Colloque Académique sur l'"Evaluation" organisé par l'I.R.E.M.

Il s'agit d'un article assez général sur les Q.C.M. dont l'objet est de proposer une certaine information quant à cet aspect particulier de l'évaluation. Il ne prétend donc pas à l'originalité et, d'ailleurs, il reprend dans une très large mesure l'étude de G. NOIZET et J.P. CAVERNI (in "Psychologie de l'évaluation scolaire" - P.U.F. - Paris 1978) dans laquelle on trouvera de nombreux compléments.

Envisagé dans le cadre d'une étude générale des problèmes de l'évaluation, ce texte fait suite à celui de Jean CESAR ("Le comportement de l'évaluateur") publié dans le numéro 17 de ce même bulletin.

Un questionnaire à choix multiple (Q.C.M.) est un ensemble de questions fermées (plus ou moins ; on y reviendra, car c'est là un problème important) pour chacune desquelles est proposé un éventail de réponses entre lesquelles l'évalué doit choisir ou, plus exactement, auxquelles il doit attribuer une valeur de vérité, binaire en général: Vrai (V) ou Faux (F). En fait les choses peuvent être plus compliquées ; par exemple, outre "Vrai" et "Faux", on peut avoir l'éventualité de réponse "Je ne sais pas", ou bien, pour certaines questions, l'éventail des réponses peut-être constitué de "Jamais", "Parfois", "Toujours".

Toutefois, pour simplifier, on se limitera ici à la considération des deux réponses "Vrai" et "Faux" ; on est ainsi amené à définir un Q.C.M. comme "un ensemble de questions décomposées en sous-questions à deux éventualités V et F".

Donnons un premier exemple d'une telle question :

Exemple 1 (Classe de 5ème) Les nombres suivants sont primaires (c'est-à-dire puissance d'un nombre premier) :

1.	144	V	F
2.	81	V	F
3.	91	V	F
4.	61	V	F
5.	35	V	F

C'est essentiellement dans le cadre d'une évaluation-bilan, évaluation sommative, que l'on a imaginé l'utilisation des Q.C.M. en vue d'éliminer les divergences de l'évaluation.

Les Q.C.M. se caractérisent en effet par la stabilité du modèle de référence : d'une part stabilité relativement aux divergences entre évaluateurs et d'autre part stabilité, au niveau d'un évaluateur donné, par suppression des effets d'assimilation et des effets de contraste (1). On obtient ainsi, avec les Q.C.M., un jugement absolu et pas seulement un jugement comparatif.

Outre cet avantage indiscutable, les Q.C.M. présentent un intérêt matériel certain du à leur possibilité de correction automatique, utilisant, par exemple, un ordinateur. Et, à cet égard, il est certain que le développement de la micro-informatique ne peut qu'inciter à l'emploi de cette technique d'évaluation qui garantit une totale fidélité de correction.

Mais si les problèmes de correction sont résolus (et encore, il y a beaucoup à dire à ce sujet ; voir plus loin), les Q.C.M. posent de nombreux problèmes de choix au niveau de leur conception. Par rapport aux procédures plus traditionnelles d'évaluation, les problèmes se situent en amont, sous une forme d'ailleurs différente.

Pour examiner la validité des Q.C.M., il faut donc d'abord se placer au niveau de leur élaboration et analyser un certain nombre de problèmes de choix, notamment :

(1) Se reporter à l'article de Jean CESAR cité dans l'avant-propos.

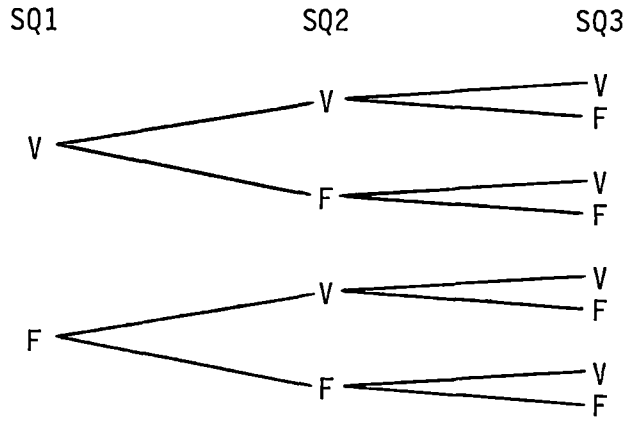
- Choix des questions
- Choix des éventualités de réponse
- Choix du système de pondération des résultats, et, ensuite et surtout, s'interroger sur le fait de savoir si les Q.C.M. sont adaptés ou peuvent être adaptés à une évaluation formative dans laquelle les objectifs pédagogiques sont définis en termes de comportements terminaux.

I - La structure des Q.C.M.

1 - Le patron des sous-réponses

Une question est un ensemble ordonné de sous-question (SQ) qu'il est intéressant de structurer.

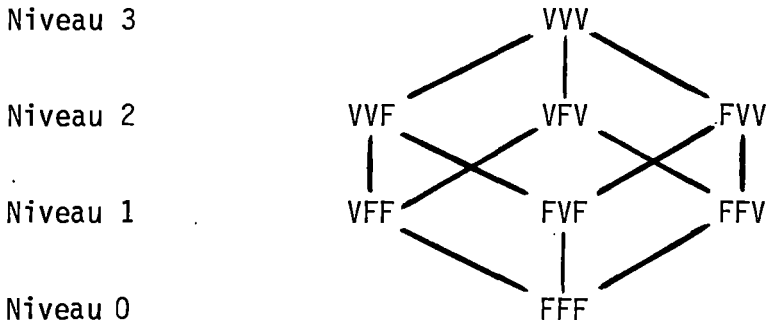
L'éventail des possibilités de réponse définit une arborescence dans laquelle une seule réponse est valable : c'est une suite (ordonnée) de la forme VVF ... FV. Cette arborescence constitue le patron des sous-réponses (SR). Par exemple, si l'on a 3 sous-questions, on aura :



la réponse pouvant être VVF, par exemple.

Bien entendu, si l'on a n sous-questions, il y a 2^n réponses possibles.

Une autre représentation de l'ensemble des sous-réponses est probablement plus intéressante. Elle consiste à représenter l'ensemble des réponses sous la forme d'un treillis décomposé en niveaux, chacun d'eux correspondant à un même nombre de sous-réponses V (donc de sous-réponses F). Ainsi dans le cas de 3 sous-questions on aura :



Dans cette représentation, deux réponses sont liées si et seulement si on passe de l'une à l'autre par modification d'une et d'une seule SR.

On notera que pour n SQ, il y a (n + 1) niveaux (de OV à nV) ayant respectivement 1, ..., C_n^p ... 1 éléments. Dans l'exemple 1, il y a 2^5 = 32 réponses à 6 niveaux ayant respectivement 1, 5, 10, 10, 5, 1 éléments.

2 - Les restrictions de l'ensemble des réponses

Dans la donnée d'une question, il arrive fréquemment que, pour des raisons diverses que nous allons voir, l'ensemble des sous-réponses ne comprenne pas les 2^n possibilités. Les restrictions sont essentiellement de deux types : restrictions de fait et restrictions de contenu.

a) Les restrictions de fait

On a une restriction de fait lorsque l'énoncé de la question impose la recherche dans un sous-ensemble donné. Il s'agit surtout de fixer le niveau (au sens du treillis précédent) dans un souci de simplification de la question.

Dans l'exemple 1, on peut limiter la recherche en prenant comme énoncé : "Parmi les 5 nombres suivants, indiquer les deux d'entre eux qui sont primaires". On impose donc une limitation au niveau 2 où il y a 10 réponses possibles. Naturellement une telle restriction modifie profondément la question et, notamment, a une incidence certaine sur les stratégies de réponse (voir plus loin).

Proposons un autre exemple, intéressant, à notre avis, car bien adapté à une évaluation formative ; cet exemple est extrait d'un test pour la classe de 4ème élaboré par le groupe "Evaluation" de l'IREM.

Exemple 2 (classe de 4ème) On considère l'équation :

$1000\ x + 345 = 0$

Parmi les équations suivantes, indiquer celles d'entre-elles qui ont le même ensemble de solutions que l'équation ci-dessus.

1. $1000\ x = 345$
2. $345\ x = 1000$
3. $1000\ x = -\ 345$
4. $1000\ x + 5\ x = -\ 345 + 5\ x$
5. $1000\ x + 5\ x = -\ 345 + 5$
6. $10000\ x + 3450 = 0$
7. $45\ x(1000\ x) = 45\ x(-\ 345)$

On peut ici encore limiter de la même manière le nombre des réponses en remplaçant la question par "quelles sont les quatre équations qui ont le même ensemble de solutions que l'équation ci-dessus". Cette formulation impose la recherche dans le niveau 4 où il y a encore 35 réponses possibles.

b) Les restrictions de contenu

Une question présente une restriction de contenu lorsque sa nature même impose un nombre déterminé a priori de sous-réponses V, ou, tout au moins, un type déterminé de réponses.

Le cas le plus simple et aussi le plus fréquent de ce type de restrictions est celui des questions dite "de complément simple" où la question est telle que, parmi les SR, une seule est vraie et toutes les autres fausses (la nature de la question impose donc une réponse au niveau 1). C'est le cas de l'exemple suivant :

Exemple 3 On multiplie 1284 par un entier x ; on trouve 444264.

1. $x = 352$ V F
2. $x = 506$ V F
3. $x = 346$ V F
4. $x = 351$ V F
5. $x = 96$ V F

Ici, sous réserve d'avoir un minimum de connaissance sur la multiplication, il est clair qu'une seule réponse est possible (la réponse est FFVFF).

Une autre restriction de contenu se présente lorsque, dans une question, il existe des liens logiques entre les sous-questions qui font qu'une réponse V, par exemple, à une SQ implique une réponse F, par exemple, à une autre.

Considérons, par exemple, la question suivante :

Exemple 4

1. Le système d'équations $\begin{cases} x + y + 1 = 0 \\ x + y + 2 = 0 \end{cases}$ a

   V F

   une solution et une seule

2. Le système d'équations $\begin{cases} x + y + 1 = 0 \\ x + y + 2 = 0 \end{cases}$

   V F

   n'a pas de solution

3. L'inéquation $-1 \leqslant t \leqslant -2$ a

   V F

   au moins deux solutions

4. L'inéquation $-1 \leqslant t \leqslant -2$ n'a

   V F

   pas de solution

Dans cet exemple, la réponse V à la SQ 3 implique la SR F à la SQ 4. En fait les seules réponses logiquement admissibles sont VFVF, VFFF, FVFV, FVFF, FVVF, FFVF ; on a donc ici une restriction de l'ensemble des réponses, mais, contrairement à ce qu'on a vu jusqu'à présent, elle ne consiste pas à se placer à un niveau déterminé du treillis des réponses.

Bien entendu, pour une telle question, la restriction doit d'avantage être considérée comme relative à l'évaluateur plutôt qu'à l'évalué. En effet ce dernier peut fort bien ne pas détecter les liens logiques entre les sous-questions et ceci peut apporter d'utiles informations à l'évaluateur, sous réserve, naturellement, de ne pas se contenter, à la correction, de comparer la réponse donnée avec la réponse exacte (FVFV).

II - Les problèmes techniques de construction des Q.C.M.

1 - La décidabilité des items

Dans toute sous-question d'un Q.C.M., la nature de la réponse, Vrai ou Faux, doit être sans ambiguïté, compte-tenu du corps de connaissances de l'évalué.

Il est donc nécessaire que chaque sous-question soit décidable. Par exemple, si on dit "ABCD est un quadrilatère tel que AB et CD sont parallèles" et si on demande "ABCD est un parallélogramme V . F", c'est une question indécidable (dans l'absolu).

Pour chaque sous-question, il y ainsi une indispensable condition de décidabilité. Elle a, malheureusement, l'inconvénient de conduire souvent à une parcellisation des connaissances, faisant surtout appel à la mémoire, excluant le raisonnement logique et les savoir-faire. De plus la décidabilité entraine une inévitable restriction de l'ensemble des réponses dont une conséquence est de favoriser le développement de stratégie de réponses (notamment dans le cas d'une évaluation sommative).

Dans le cas particulier des mathématiques (c'est probablement vrai dans d'autres disciplines), la parcellisation des connaissances peut se traduire par la mise en avant de l'acquisition du vocabulaire, au détriment d'une véritable assimilation des concepts sous-jacents et, a fortiori, d'une maitrise de certains savoir-faire. Par exemple dans un Q.C.M. du C.N.E.C., à propos des notations usuelles en géométrie, on propose la relation (A, B) $\in \vec{DC}$ (ABCD est un parallélogramme) et on demande, pour chacun des deux termes de cette relation, Bipoint V F, Couple V F, Paire V F, Ensemble de Points V F, Ensemble de Segments V F, Vecteur V F, Axe V F, Droite V F. C'est typiquement une question de vocabulaire dont on peut douter sérieusement de l'intérêt réel.

Mais le nécessaire maintien de la décidabilité des items ne doit pas, toutefois, être considéré comme imposant des contraintes définitives. En effet, grâce à une formulation convenable des sous-questions, on peut restaurer la décidabilité et obtenir des questions fort intéressantes pour l'évaluateur. Dans le cas de l'exemple cité au début de ce paragraphe, on peut le reprendre sous la forme suivante :

Exemple 5 ABCD est un quadrilatère tel que AB et CD sont parallèles

1. On peut affirmer que ABCD est un parallélogramme V F
2. BC est nécessairement parallèle à AD V F
3. Les diagonales de ABCD se coupent en leur milieu V F

La décidabilité a été ainsi rétablie à partir d'un ensemble de données a priori insuffisantes pour conclure. En outre on a obtenu une question où les liens logiques entre les sous-questions peuvent procurer à l'évaluateur d'utiles renseignements par comparaison des sous-réponses. Signalons que le type de la réponse, FFF, gênera certainement les élèves.

A propos de la décidabilité des items, il faut encre remarquer qu'elle doit être une conséquence de l'assimilation de connaissances de la part de l'évalué et ne pas relever uniquement du hasard. On a donc une notion de décidabilité relative, variable avec le niveau de l'évalué.

Par exemple, si l'on pose la question "π est plus grand que 3,1415926536", elle est, bien entendu, décidable dans l'absolu, mais on peut dire qu'en pratique, dans la plupart des cas, elle est indécidable.

2) Le problème des distracteurs

Dans certaines questions d'une Q.C.M., on se borne parfois à proposer des sous-questions qui balayent la totalité des réponses possibles, eu égard au corps de connaissances (optimal) de l'évalué. Considérons, par exemple, la question suivante relative aux "caractères de divisibilité" :

Exemple 6 Le nombre 198 est

1. Premier	V	F
2. Divisible par 2	V	F
3. Divisible par 3	V	F
4. Divisible par 4	V	F
5. Divisible par 5	V	F
6. Divisible par 9	V	F
7. Divisible par 10	V	F

Ici on énumère tous les caractères de divisibilité usuellement enseignés et la liste des sous-questions apparait tout simplement comme un rappel des divers paragraphes du cours.

Cette façon de procéder peut présenter un intérêt matériel dans une évaluation sommative lorsqu'on veut contrôler des acquisitions ponctuelles. En revanche elle limite considérablement l'analyse des résultats de la part de l'évaluateur qui (dans une évaluation formative) a tout intérêt à ce que certaines sous-questions soient des distracteurs, c'est-à-dire des sous-questions invitant à l'erreur (ce qui n'est pas le cas dans l'exemple précédent, puisqu'aucune des sous-questions n'est vraiment une

invitation à l'erreur).

Pour une sous-question, l'erreur peut aussi bien consister à répondre V pou F que F pour V. Dès lors, si, dans le cadre d'une analyse purement objective des Q.C.M., l'on veut conserver la symétrie V - F, il faut se borner à définir un distracteur comme on vient de le faire (sous-question invitant à l'erreur). Toutefois, c'est une définition trop vague qui risque d'introduire des ambiguïtés dans l'étude de cette notion ; de plus, il faut être conscient qu'au niveau psychologique la symétrie V - F n'existe pas, le Vrai étant privilégié par rapport au Faux (voir ci-dessous, partie III) C'est pourquoi, dans un souci de simplification et de clarté, un distracteur désignera dans la suite une sous-question à laquelle il faut répondre Faux.

Un distracteur va d'abord se caractériser par sa vraisemblance. Dans l'exemple 3 (trouver x tel que 1284 x = 444264) si l'on donne la SQ x = 5, ce sera un distracteur peu vraisemblable, on dira un distracteur très faible. Par contre, la SQ x = 351 est un distracteur assez attractif ; on dira donc distracteur fort. Pour une question donnée il convient donc de définir une échelle de vraisemblance (de faible à fort) des distracteurs. C'est parfois très simple (comme pour l'exemple 3), mais, souvent, il sera nécessaire de faire un pré-test, c'est-à-dire, dans le cadre d'une évaluation classique (non QCM), une expérimentation analysant les types de réponses et leurs variations suivant la formulation des questions.

Selon les objectifs poursuivis, on peut avoir intérêt, soit à proposer des distracteurs d'égale vraisemblance, soit au contraire à varier la vraisemblance des distracteurs.

Pour le premier cas, il convient de noter que l'égale vraisemblance des distracteurs peut fort bien être utilisée autrement que comme dans l'exemple 6. Considérons, par exemple, la question suivante (à l'usage d'une classe de terminale C) (2) :

Exemple 7 Le double de l'âge de Patrick est égal au triple de l'âge d'Eric augmenté de l'âge d'Olivier. Le double du cube de l'âge de Patrick est égal au triple du cube de l'âge d'Eric augmenté du cube de l'âge d'Olivier.

Les âges de Patrick, Eric et Olivier sont premiers entre eux dans leur ensemble. La somme de leurs carrées est égale à :

1. 42	V	F
2. 46	V	F
3. 122	V	F
4. 290	V	F
5. 326	V	F

(2) Cette question est extraite de "The contest Problem Book III" (annual High School contests 1966 - 1972), compiled with solutions by C. T. Salkind and J.M. Earl - (The Math. Ass. of America - New Mathematical Library) où tous les problèmes sont rédigés en Q.C.M.

Ici on peut affirmer que les distracteurs sont également vraisemblables. Cette question convient bien, pensons nous, à une évaluation bilan dans la mesure, notamment, où elle évite, ou tout au moins limite, les stratégies de réponse.

Dans le deuxième cas (et cela concerne surtout l'évaluation formative), il ne faudrait plus croire que l'on doive prendre obligatoirement des distracteurs forts, car il peut être intéressant de ménager une progression. C'est ainsi que dans l'exemple 3, avec des distracteurs bien choisis, on peut évaluer l'assimilation de la notion d'ordre de grandeur d'un résultat.

D'autre part, dans le choix d'un distracteur, il peut être intéressant de faire intervenir une espéce de critère structural, permettant de détecter un manque d'information ou un défaut de raisonnement. Par exemple, (à propos des priorités des opérations) considérons les deux SQ :

1. $7 + 4 \times 3 = 19$ V F
2. $7 + 4 \times 3 = 33$ V F

Ici le distracteur est choisi pour donner à l'évaluateur une information en retour qui peut être extrêmement précieuse. Bien entendu proposer de tels distracteurs demande une pratique pédagogique assez importante qui aura mis en lumière les sources d'erreurs les plus fréquentes, ainsi que, souvent, un pré-test.

Enfin, en songeant au développement des stratégies de réponse, il faut tenir compte de la communauté des distracteurs. Supposons qu'à la suite de la question ci-dessus, on propose la question à deux SQ :

1. $8 \times 4 + 5 = 72$ V F
2. $8 \times 4 + 5 = 37$ V F

qui est du même type avec inversion de la sous-question vraie (SQV).

D'une part cette question n'est plus nécessairement significative après la première. D'autre part, elle peut engendrer une stratégie de réponse du type : on a répondu V à la SQ 1 pour la 1ère question, donc on va répondre F à la SQ 1 de la 2ème question, ce qui risque de perturber totalement l'analyse des résultats. Dans cet exemple, il y a clairement une interaction entre les deux questions dont il est indispensable de tenir compte. D'une manière générale, ceci nous amène à dire que la construction d'un Q.C.M. doit s'effectuer dans le cadre d'une prise en compte globale et non pas dans

un processus d'élaboration question par question.

Naturellement l'effet d'interaction entre les questions est d'autant plus sensible que les distracteurs introduits sont faits.

3) La formulation des questions

La formulation des questions peut varier, soit dans la forme (syntaxique), soit dans la source de la question.

a) Variations dans la forme

La question et chaque sous-question peuvent être formulées affirmativement ou négativement.

Considérons, par exemple, la question suivante pour laquelle on se rend facilement compte des possibilités de formulations différentes.

Exemple 8 L'étude des symétries du plan qui laissent invariantes deux droites sécantes montre que :

1. Il y a une différence entre le cas où les droites sont orthogonales et celui où elles ne le sont pas V F
2. Le nombre de ces symétries est toujours au moins égal à 3 V F
3. Parmi ces symétries, il y a toujours au moins une symétrie centrale V F

On est ici en présence d'un schéma QA - SQA (question affirmative - sous-questions affirmatives), mais on aurait fort bien pu formuler la question avec l'un des schémas suivants : (N pour négative et M pour mixte) QN - SQN, QA - SQN, QN - SQA ou encore QN - SQM, QA - SQM.

Ces différents schémas correspondent à des patrons de réponse distincts, mais surtout ils font appel à des capacités cognitives différentes. Par ailleurs, les phénomènes d'attraction des phrases affirmatives (on en reparlera plus loin) créent des différences assez sensibles entre les divers schémas possibles.

A propos de cet exemple, indiquons qu'on aurait pu ajouter la sous-question :

4. Le nombre de ces symétries est plus grand quand les 2 droites sont orthogonales V F

On aurait ainsi créé une restriction de contenu dont il serait intéressant de mesurer l'incidence sur les réponses obtenues.

b) Variation dans la source de la question

Une question repose, en général, sur un ensemble de données et un ou plusieurs énoncés (les données peuvent être incluses dans l'un d'eux) ; en mathématiques, on a le schéma classique Données - Hypothèse $\Longmapsto$ Conclusion.

La question peut alors porter :

- Soit sur la validité d'un ou plusieurs énoncés
- Soit sur un lien entre deux ou plusieurs énoncés (notamment sur leur compatibilité)

par exemple : * $\sqrt{-5} < 1$ V F
porte sur la validité de l'énoncé

* Si $\sqrt{a} > 1$ alors $\sqrt{-a} < 1$ V F
porte sur la validité de la conclusion

* Si $\sqrt{a} > \sqrt{b}$ alors $a > b$ V F
porte sur la nature du lien entre les deux inégalités.

Les différents problèmes liés à la formulation des questions doivent être pris en compte dès l'instant où l'on a des intentions précises quant à la passation des Q.C.M.

En particulier, il ne faut pas oublier que, dans l'enseignement, la formulation est le plus souvent affirmative et qu'un énoncé négatif demande d'avantage de réflexion qu'une restitution pure et simple du savoir.

4) Le problème de la notation

Contrairement à ce qu'on pourrait penser, le problème de la notation n'est pas absolument évident.

Naturellement, on peut songer à donner à chaque question d'un Q.C.M. le même nombre de points.

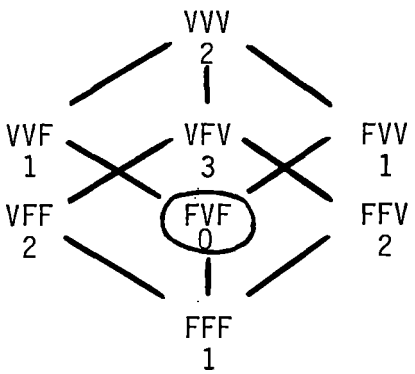
En fait, il est plus raisonnable d'envisager une pondération des questions. Mais alors, doit-elle être déterminée a priori - suivant les objectifs des évaluateurs -

ou a posteriori - d'après une analyse statistique des réponses - ? La première méthode nous parait préférable, voire même obligatoire, dans le cadre d'une évaluation formative ; il va sans dire que cette pondération a priori mérite d'être ajustée par un prétest ou un étalonnage.

a) La notation au niveau d'une question

Là encore, la méthode la plus simple consiste à donner un point à chaque sous-réponse exacte et 0 point à chaque sous-réponse inexacte. Malheureusement ce type de notation a l'inconvénient d'éliminer les liens éventuels entre deux SQ et, notamment dans le cas d'une question de complément simple, de privilégier les réponses inexactes. L'idéal est, bien sûr, de regrouper les sous-questions ayant des liens logiques entre elles.

On peut également introduire la notion de distance à l'erreur. Précisons de quoi il s'agit sur un exemple : soit une question à 3 sous-questions pour laquelle la réponse exacte est FVF et reprenons le treillis des réponses.



La dsitance à l'erreur est le nombre minimal d'arcs à parcourir sur ce graphe pour atteindre la réponse exacte (c'est aussi le nombre de sous-questions auxquelles il a été mal répondu). Ces distances (0, 1, 2, 3) sont inscrites, ci-dessus, en dessous de chaque réponse et l'on pourrait noter : FVF = 3 points, VVF, FFF, FVV = 2 points, VVV, VFF, FFV = 1 point, VFV = 0 point. Une telle notation qui prend en compte la totalité des réponses a l'avantage d'être plus globale.

Si l'on envisage une notation a posteriori, on peut décider de tenir compte de l'ensemble des résultats et, pour simplifier, admettre que le total des notes est 0 (il suffira ensuite de faire une translation). On attribuera un nombre de points positif ou négatif d'autant plus important que la fréquence de réussite est faible.

Pour préciser prenons le cas simple d'une question où l'on considère les seules éventualités de réponse : Exact et Inexact. Si l'on a p réponses exactes et q inexactes, donc une fréquence $f = \frac{p}{p+q}$ de réponses exactes, on attribuera $x > 0$ points (resp. - y points où $y > 0$) pour une réponse exacte (resp. inexacte).

On aura donc $px - qy = 0$, soit $y = \frac{p}{q} x = \frac{f}{1-f} x$ et $x = k(1 - f)$, $y = kf$.

C'est là une méthode de type statistique qui peut être admissible dans une évaluation sommative, mais qui ne l'est certainement pas dans une évaluation formative où il ne s'agit pas de classer les évalués, mais de déterminer si les objectifs pédagogiques sont atteints.

En fait, même dans le cas d'une évaluation sommative, cette méthode est discutable. Dans un ensemble de sous-questions et dans le cadre de la même logique, on peut admettre que le nombre de points à chaque SQ est proportionnel au nombre de réponses inexactes. Ceci entraine des calculs qui font très étroitement dépendre les notes attribuées du score général (3).

b) Le problème des réponses au hasard

Le problème des réponses au hasard peut être résolu par des pondérations probalilistes en estimant la probabilité d'une réponse au hasard. A vrai dire c'est un calcul difficile qui peut reposer sur des hypothèses gratuites, comme celle que l'on peut être amenée à faire en décidant qu'une réponse exacte est donnée au hasard.

En fait, le hasard va dépendre d'un certain nombre de facteurs :

(1) Le nombre de réponses possibles (qu'on a intérêt à prendre suffisamment grand)

(2) La structure du patron des sous-réponses (dans cet ordre d'idées, les questions de complément simple sont sans doute à éviter).

---

(3) *NOIZET* et *CAVERNI* (loc. cit.) proposent un exemple où l'on a choisi $x = f(1 - f)$, $y = f^2$ ce qui est très surprenant. Bien entendu ils ont des résultats contradictoires !

(3) La vraisemblance des distracteurs ; c'est là un facteur fondamental. Naturellement, en vue d'éliminer les réponses au hasard, on pourrait penser à faire en sorte que ses distracteurs aient tous le même degré de vraisemblance, mais, comme on l'a déjà dit, on se priverait ainsi d'une certaine information en retour (dans une évaluation formative). Il nous parait donc nécessaire, répétons le, de disposer d'un classement des distracteurs ; c'est en fait un problème délicat comme on peut s'en rendre compte en réfléchissant à la question suivante :

Question : Pour obtenir $\sqrt{62410}$ il faut multiplier $\sqrt{6241}$ par :

1. $\sqrt{2}$ V F - 2. 10 V F - 3. 0,5 V F - 4. $\sqrt{10}$ V F - 5. 6241 V F

Un tel classement suppose une grande pratique pédagogique qui, de toute manière, ne permettra pas d'éviter les variations d'appréhension de la part des évalués.

III - Les stratégies de réponse

Mis en présence d'un Q.C.M., l'évalué à tendance à développer une "stratégie de réponse" (consciemment ou non) qui va entacher d'erreur la fidélité de la correction.

1 - Discrimination et identification

Dans un Q.C.M., l'évaluateur attend que, pour chaque SQ, l'évalué se livre à une identification du Vrai ou du Faux. Cela suppose que les SQ soient traitées indépendamment les unes des autres et donc que, chez l'évalué, il y ait distinction nette entre savoir et ignorance.

Cette attitude est favorisée dans les questions où il n'y a pas de restriction de fait de l'ensemble des réponses ; il s'agit là d'une tâche de jugement multiple.

Dans la réalité, l'évalué est conduit à établir une dépendance entre ses réponses et à se livrer à une tâche de discrimination. En face d'une question, si ses connaissances ne sont pas suffisantes pour lui permettre de décider de la réponse, il tente de déterminer une probabilité de vraisemblance.

Il faut être conscient d'une telle attitude et, dans l'élaboration d'un Q.C.M., il faut s'efforcer d'en limiter le développement ; pour cela une condition nécessaire est de proposer des questions ayant des patrons de sous-réponses distincts et des

restrictions variables de l'une à l'autre.

A propos des restrictions, il faut noter que l'indication du nombre de SRV est sans influence pour celui qui procède à une identification, mais qu'elle peut avoir une grande importance dans le cas d'une discrimination (4).

Dans le cas d'un Q.C.M. mathématique portant sur des problèmes d'ordre technique, la discrimination peut se confondre avec une vérification du résultat : parmi les réponses proposées, l'évalué essaye de vérifier quelles sont celles d'entre elles qui correspondent à la ou aux bonnes réponses. Cette façon de faire n'est pas forcément condamnable ; en tout cas elle entraine dans la plupart des cas, une perte de temps pour l'évalué qui est ainsi sanctionné. Considérons, par exemple, la question suivante :

Exemple 9 On donne $x = 1 + 2^p$ et $y = 1 + 2^{-p}$; alors y s'exprime en fonction de x par :

1. $y = \frac{x + 1}{x - 1}$ V F
2. $y = \frac{x + 2}{x - 1}$ V F
3. $y = \frac{x}{x - 1}$ V F
4. $y = 2 - x$ V F
5. $y = \frac{x - 1}{x}$ V F

Une attitude d'identification va conduire à utiliser les données pour calculer y en fonction de x. Mais, ici, l'évalué peut aussi, pour chacune des sous-questions, calculer l'expression proposée. C'est aussi une attitude admissible, mais elle demande d'avantage de temps (en songeant à cette façon de procéder, on peut se demander : où est-il préférable de placer la bonne réponse ?).

2 - Repérage du Vrai et repérage du Faux

a) Les erreurs sont de deux types :

SQF ⇝ SRV : détection erronée d'un signal
SQV ⇝ SRF : non détection d'un signal

(4) Pour une étude précise fondée sur des expérimentations, se reporter à *NOIZET-CAVERNI* : Loc. Cit., pp 171 - 173.

Les erreurs du premier type sont, en général, sensiblement plus fréquentes. D'une part, les candidats surestiment le nombre de réponses vraies, et, d'autre part, ils estiment moins grave de déclarer vrai un énoncé faux que de déclarer faux un énoncé vrai.

b) Les bonnes réponses sont également de deux types :

L'expérience montre que les performances sont meilleures dans le premier cas et cela tient sans doute à ce que l'enseignement est formulé sous la forme d'un corps d'énoncés vrais.

Par ailleurs, si l'on oriente la recherche vers celle du vrai, les résultats sont meilleurs que dans une orientation vers la recherche du faux.

Enfin, les énoncés "vrais-affirmatifs" ou "faux-négatifs" sont plus facilement détectés que les énoncés "vrais-négatifs" ou "faux-affirmatifs".

3 - Confiance dans les réponses

Si l'on veut affiner les résultats, il peut être intéressant d'ajouter aux réponses V, F, la réponse "Je ne sais pas". En tout cas cela nous parait souvent utile dans une évaluation formative.

On peut également demander une modulation de la réponse, en l'assortissant d'un coefficient de confiance (1, 2, 3 par exemple). Cela a indiscutablement un effet désinhibiteur et, en pratique, limite l'abstention.

Signalons également que certains ont proposé une espèce de prise de risque de la part de l'évalué à qui on demande de choisir le nombre de points qui seront attribués à chaque question. Nous pensons que cette méthode est à éviter, car elle fait intervenir des facteurs qui n'ont rien à voir avec les capacités réelles de l'évalué : attitude de prudence, goût du risque, etc...

Pour en terminer avec ce paragraphe, il est certain que les stratégies de réponse sont bien réelles et présentent un inconvénient sérieux quant à l'emploi des Q.C.M. Mais il ne faut peut être pas en exagérer l'importance, car en limitant le

temps dont dispose l'évalué pour répondre, on peut très sérieusement réduire le développement de ces stratégies.

IV - Les Q.C.M. et les mathématiques

Dans tout apprentissage, on peut distinguer :

Les capacités de restructuration et de créativité face aux incitations pédagogiques
et Le degré de compréhension et la disponibilité des connaissances acquises.

Il semble à peu près évident que les Q.C.M. privilégient le deuxième aspect, en ignorant, dans la plupart des cas, le premier. C'est pourquoi, ils semblent relativement mal adaptés à un emploi en mathématiques. On peut considérer, pour schématiser, que l'apprentissage de cette discipline consiste :

(1) A assimiler un certain nombre de techniques
(2) A se construire une logique de raisonnement
(3) A faire preuve d'une capacité de mise en oeuvre de ses acquisitions

ces tâches n'étant évidemment pas indépendantes.

Pour ce qui concerne les problèmes techniques, on peut craindre que les Q.C.M. ne se limitent qu'à des épreuves de contrôle alors qu'il est indispensable que l'élève sache effectuer complètement un calcul et produire un résultat. Il convient donc de ne pas réduire une question d'un Q.C.M. à une simple vérification, ce dont nous avons déjà parlé.

Pour cela, on peut utiliser une espèce de "technique de masquage" des résultats qui interdise une simple vérification, en demandant à l'évalué un calcul complémentaire, très simple pour lui. L'exemple 7 illustre cette technique, puisque les sous-questions ne proposent pas le triplet des âges cherchés, mais la somme de leurs carrés : pour cet exemple, on voit mal comment répondre correctement sans effectuer complètement le calcul (nous en prenons le lecteur à témoin : c'est pourquoi nous n'avons pas indiqué la bonne réponse !).

On peut également songer à décomposer une méthode technique (construire un graphe de calcul) et proposer les opérations successives dans un Q.C.M. Cela peut avoir un effet heureux dans la mesure où certains élèves ont du mal à créer leur

propre démarche en l'absence d'un plan pré-établi proposé par les enseignants.

Pour ce qui concerne les aptitudes au raisonnement, on a déjà souligné l'intérêt de liens logiques entre les sous-questions et, bien sûr, de leur prise en compte par l'évaluateur. On peut aussi tout simplement, proposer des questions de type purement logique dont un exemple rudimentaire pourrait être :

Etant donné un quadrilatère, on considère les 2 énoncés :

A = un couple de côtés opposés est parallèle
B = les diagonales se coupent en leur milieu

1. Si A est vrai, alors B est vrai V F
2. Si A est faux, alors B est faux V F
3. Si A est vrai, alors B est faux V F

.....

8. Si B est faux, alors A est vrai V F

Quant à l'activité mathématique (mise en oeuvre des acquisitions), on peut estimer que les Q.C.M. sont très inadaptés. Nous pensons que ce n'est probablement pas tout à fait exact et que, sous réserve d'étudier très soigneusement la question, on peut, par des Q.C.M., amener les élèves à faire preuve d'imagination et de créativité en suscitant chez eux quelques démarches possibles. Mais, à ce sujet, il est probable que tout reste à faire et nous ne conclurons pas, nous bornant à proposer un dernier exemple qui montre , peut-être, comment les Q.C.M. peuvent déboucher sur l'activité mathématique (cet exemple est inspiré d'un exercice posé fréquemment dans les manuels pour le premier cycle).

Exemple 10 Je peux passer le contrat suivant avec ma banque : Je place une certaine somme S et, à la fin de chaque mois, la banque double mon avoir, mais retire de mon compte une somme fixe de 600 F.

1. S ne peut pas être inférieur à 600 F
2. Si S est inférieur à 600 F, je vais me ruiner
3. Si S est inférieur à 600 F, dans 3 mois je serai ruiné
4. Si S est supérieur à 600 F, je suis certain de m'enrichir
5. Si S = 1200 F, mon avoir sera multiplié par 3 en 3 mois
6. Si je veux doubler mon avoir en deux mois, je dois placer S = 900 F
7. Pour multiplier mon avoir par 10 en 1 an je dois placer S = 1000 F
8. Si S = 750 F, mon avoir aura quadruplé en 4 mois

etc.....

J.Claude FONTAINE

0

0 0

IV) A PROPOS DE L'AGE DU CAPITAINE

L'équipe "élémentaire" de l' I. R. E. M. de Grenoble a rendu compte dans le bulletin de l' A. P. M. E. P. n° 323 du comportement d'élèves de C E et C M lorsque des problèmes "absurdes" - par exemple : sur un bateau il y a 26 moutons et 10 chèvres ; quel est l'âge du capitaine ? - leur étaient proposés.

Commentant les résultats obtenus, l'équipe précisait : " On peut remarquer un écart important entre les réponses des élèves de C E et celles de C M : les trois quarts des enfants de C E environ trouvent "l'âge du capitaine" , tandis qu'il n'y en a plus qu'environ un tiers en C M ".

Cette remarque nous a incités à "emprunter" à l' IREM de Grenoble ses énoncés "absurdes" et à les proposer à 23 élèves d'une classe de sixième dans l'ordre suivant :

PROBLEMES

1 J'ai 4 sucettes dans ma poche droite et 9 caramels dans ma poche gauche.
Quel est l'âge de mon papa ?

2 Dans une bergerie il y a 125 moutons et 5 chiens.
Quel est l'âge du berger ?

3 Un berger a 360 moutons et 10 chiens.
Quel est l'âge du berger ?

4 Dans une classe, il y a 12 filles et 13 garçons.
Quel est l'âge de la maîtresse ?

5 Dans un bateau il y a 36 moutons, 10 tombent dans l'eau.
Quel est l'âge du capitaine ?

6 Il y a 7 rangées de 4 tables dans la classe.
Quel est l'âge de la maîtresse ?

Les enfants ont été placés dans les mêmes conditions que pour n'importe quelle autre interrogation écrite.

Ils ont néanmoins été avertis que des problèmes leur seraient proposés pour lesquels ils devraient soit EN DONNER LA SOLUTION soit PRECISER QU'ILS NE POUVAIENT PAS LA TROUVER.

Les énoncés ont été projetés sur un écran, à l'aide d'un rétro-projecteur, les uns après les autres en laissant le temps à chaque élève, entre deux projections, d'indiquer par écrit la réponse choisie.

Précisons enfin que cette épreuve n'avait pas pour but de contrôler la maîtrise d'un objectif, l'enseignant n'ayant eu en l'occurrence qu'une seule intention : se rendre compte.

REPONSES

Parmi les 23 élèves

- 6 disent ne jamais pouvoir répondre
- 3 répondent à certaines questions et disent ne pas pouvoir répondre aux autres
- 2 pensaient ne pas pouvoir répondre puis ont répondu aux six questions
- 12 répondent à chacune des six questions.

PROBLEME 1

L'âge de mon papa est :

9 X 4 = 36	13 réponses
(9 + 4) X 2 = 26	1 réponse
13	1 réponse

PROBLEME 2

L'âge du berger est :

125 X 5 = 625	1 réponse
125 - 5 = 12	1 réponse
120	1 réponse
125 + 5 = 130	2 réponses
125 : 5 = 25	12 réponses

PROBLEME 3

L'âge du berger est :

360 - 10 = 350	1 réponse
360 - 10 = 35	1 réponse
360 + 10 = 370	2 réponses
360 : 10 = 36	11 réponses

PROBLEME 4

L'âge de la maîtresse est :

12 + 13 = 25	15 réponses
12	1 réponse

PROBLEME 5

L'âge du capitaine est :

36 : 10 = 36	1 réponse
36 + 10 = 46	1 réponse
27	1 réponse
36 - 10 = 26	13 réponses

PROBLEME 6

L'âge de la maîtresse est :

7 X 4 = 28	16 réponses

Certes, même en cours d'année, les épreuves de contrôle de connaissances et d'aptitudes sont abordées par les enfants avec appréhension voire anxiété.

Mais la présence de cette tension suffit-elle à expliquer que les deux tiers environ des élèves de cette classe de sixième trouvent "l'âge du capitaine" ?
Que deux d'entre eux pensaient ne pas pouvoir répondre mais ont néanmoins fourni des résultats ?

L' incohérence entre les données et la question posée est-elle toujours remarquée ?

N'est-il pas probable que trop souvent un énoncé de problème représente, pour les élèves, une énigme à résoudre dont le libellé contient les données à utiliser ?

Est-il si certain que les élèves se soucient peu de la pertinence de leurs réponses ?

Est-il enfin judicieux de poser des problèmes "absurdes" pour apprécier la façon dont un énoncé de problème est perçu par les élèves ?

Chantal et Jean Paul GOVIN

0

0 0

V) EVALUATION ET LIAISON COLLEGE - LYCEE

Le professeur de Collège est parfois surpris d'apprendre que tel de ses anciens élèves qui avait manifesté de bonnes aptitudes en troisième est en échec quasi-total en seconde, ou au contraire que tel élève dont les résultats étaient très faibles au collège obtient de bons résultats en seconde. De son côté, le professeur de Lycée peut s'étonner d'apprendre que tel élève vraiment dépassé par les évènements en seconde avait des résultats tout à fait honorables en troisième.

Il nous a semblé que ces remarques méritaient une étude quantitative plus précise. Depuis plusieurs années, les lycées envoient dans les collèges, les notes obtenues par leurs anciens élèves. En réalité, plusieurs lycées n'envoient que les résultats du premier trimestre de seconde, mais cela permettait déjà d'observer la façon dont les élèves sont évalués en début de seconde et de comparer aux résultats obtenus en troisième.

Dans l'étude qui suit, nous nous sommes limités à l'observation du devenir à court terme (un trimestre), en mathématiques (et en français, mais nous n'en ferons que peu état ici) des élèves d'un seul collège : le COLLEGE D'ORNANS, et ceci pour les promotions qui étaient en troisième pendant les années scolaires 79-80, 80-81 et 81-82. Ces promotions regroupent exactement 113 élèves, mais nous n'avons pu obtenir les données nécessaires que pour 100 élèves.

L'échantillon étudié est ainsi composé :

78-79 : 31 élèves
79-80 : 34 élèves
80-81 : 35 élèves

A chacun de ces élèves nous avons associé un couple (x ; y).

- x : Moyenne des notes de mathématiques obtenues au collège au cours de l'année n, en troisième
- y : note de mathématiques du 1er trimestre de l'année n + 1, en seconde.

Nous nous sommes demandés dans quelle mesure la connaissance de x permettait de prédire y, c'est-à-dire si l'évaluation faite en troisième pouvait être considérée comme étant prédictive de celle qui sera faite en seconde, ou encore : Les variables x et y sont-elles corrélées ?

Le plus simple était de présenter une représentation graphique du graphe obtenu. C'est ce qui est fait figure 1, où chaque point correspond à un élève, l'abscisse de ce point étant la valeur de x correspondante, l'ordonnée étant la valeur de y.

Ce tableau parle de lui-même, par exemple un élève qui a obtenu une moyenne de 12 ou 13 en troisième peut s'attendre à avoir en seconde une note comprise entre 5 et 13, et cela de façon à peu près équiprobable. De même un élève qui obtient 7 en seconde avait en troisième une moyenne comprise entre 7 et 15.

Si l'on considère le rectangle :

$$\begin{cases} 8 < x < 15 \\ 4 < y < 12 \end{cases}$$ (rectangle entouré figure 1)

On constate qu'il contient 74 élèves (74 % de l'effectif), et que les points intérieurs y sont à peu près également répartis. Un élève dont le niveau en troisième a été évalué entre 8 et 15 aura donc un niveau de seconde évalué entre 4 et 12, et cela de façon apparemment aléatoire. (distribution de densité uniforme ?). Le moins que l'on puisse dire est que la valeur prédictive des résultats de troisième est fortement mise en doute.

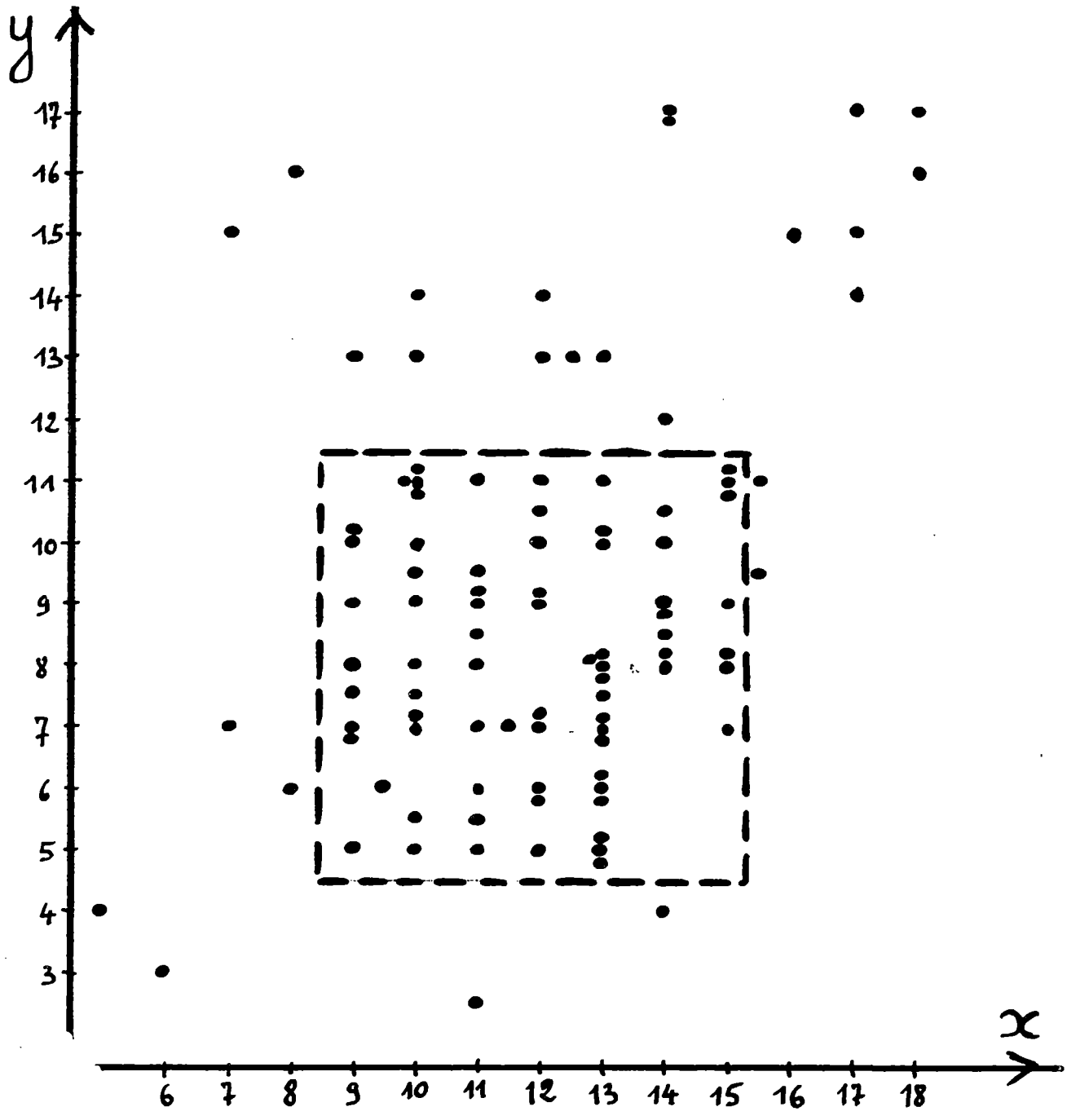


Figure 1

On peut avoir envie de calculer un coefficient de corrélation entre les variables x et y et de calculer l'équation de la droite de régression, bien que la disposition du nuage de points n'incite guère à faire ce calcul.

On trouve en effet :

Coefficient de corrélation : $f \simeq 0{,}24$. Comme on pouvait s'y attendre, les variables x et y sont donc très faiblement corrélées.

Droite de régression : elle a pour équation $y = 6{,}95 + 0{,}17\ x$

Le coefficient de x montre le peu d'influence de x sur y.

Ces résultats étant ainsi livrés, un certain nombre de questions ou d'objections viennent aussitôt à l'esprit. Sans prétendre épuiser le sujet, nous allons essayer d'en aborder quelques unes.

1ère question : Ces résultats ne s'expliquent-ils pas par une grande diversité dans les méthodes d'évaluation des professeurs de collège et des critères qu'ils prennent en compte ?

Il ne sera sans doute pas possible de répondre de façon satisfaisante à cette question, mais il faut remarquer que les 100 élèves de notre échantillon ont été notés par 5 professeurs différents qui n'avaient pas particulièrement l'habitude de se concerter sur leurs méthodes d'enseignement et d'évaluation. Dans tous les cas, les notes maxima et minima attribuées par ces professeurs sont comparables.

En première analyse, ce que nous remarquons, c'est que le critère "professeur" semble moins intervenir que le critère "groupe - classe de 3ème". Telle classe réputée "bonne" au collège verra ses notes en seconde chuter d'un point en moyenne, alors que telle autre, réputée "faible" au collège subira une chute de 5 points en moyenne, et cela pour un même professeur de troisième. Ce qui interviendrait serait donc la tendance que nous avons à ajuster nos critères et nos exigences au niveau réel du groupe que nous avons en charge.

La pratique, qui s'instaure timidement, de l'utilisation concertée d'épreuves étalonnées tels les tests de l'I.R.E.M., devrait permettre d'atténuer l'influence du critère classe.

2ème question : Ces résultats peuvent-ils s'expliquer par une grande diversité dans les méthodes d'évaluation des professeurs de lycée et des critères qu'ils prennent en compte ?

N'ayant pas relevé la variable "professeur de lycée", nous n'essayerons pas de répondre directement à cette question. Cependant, nous pouvons noter que la chute des notes des élèves d'ORNANS qui est en moyenne de 2,71 points, peut varier de façon significative suivant les établissements d'accueil.

Chute moyenne :
- Lycée Pergaud : 3,03 points
- Lycée Victor Hugo : 3,20 points
- Lycée Pasteur : 2,04 points
- Lycée Jules Haag : 3,34 points

Tout se passe donc comme si le Lycée Pasteur était moins sélectif que les autres lycées. Il est remarquable qu'après avoir repris ces calculs année par année, ils se révèlent assez stables dans le temps, y compris après la mise en place des secondes de détermination.

3ème question : Nous avons mélangé dans notre échantillon des élèves de trois années scolaires successives. Or une partie d'entre eux seulement ont connus la seconde de détermination. La mise en place de ces secondes ayant pour objectif affirmé de réduire le rôle sélectif des mathématiques et de faciliter la transition collège-lycée, n'est-il pas probable que les résultats soient très différents selon que l'on se place avant ou après l'instauration de ces nouvelles secondes ?

Nous avons donc repris notre étude en distinguant ces deux périodes.

1ère période : Avant la seconde de détermination

Il convenait dans ce cas de distinguer parmi les élèves ceux qui avaient été orientés en seconde littéraire de ceux qui avaient été orientés en seconde scientifique. Nous avons donc fait des représentations graphiques séparées pour ces deux populations (figures 2 et 3). Ces tableaux sont éloquents et l'on peut faire à leur propos les mêmes observations que pour le tableau général.

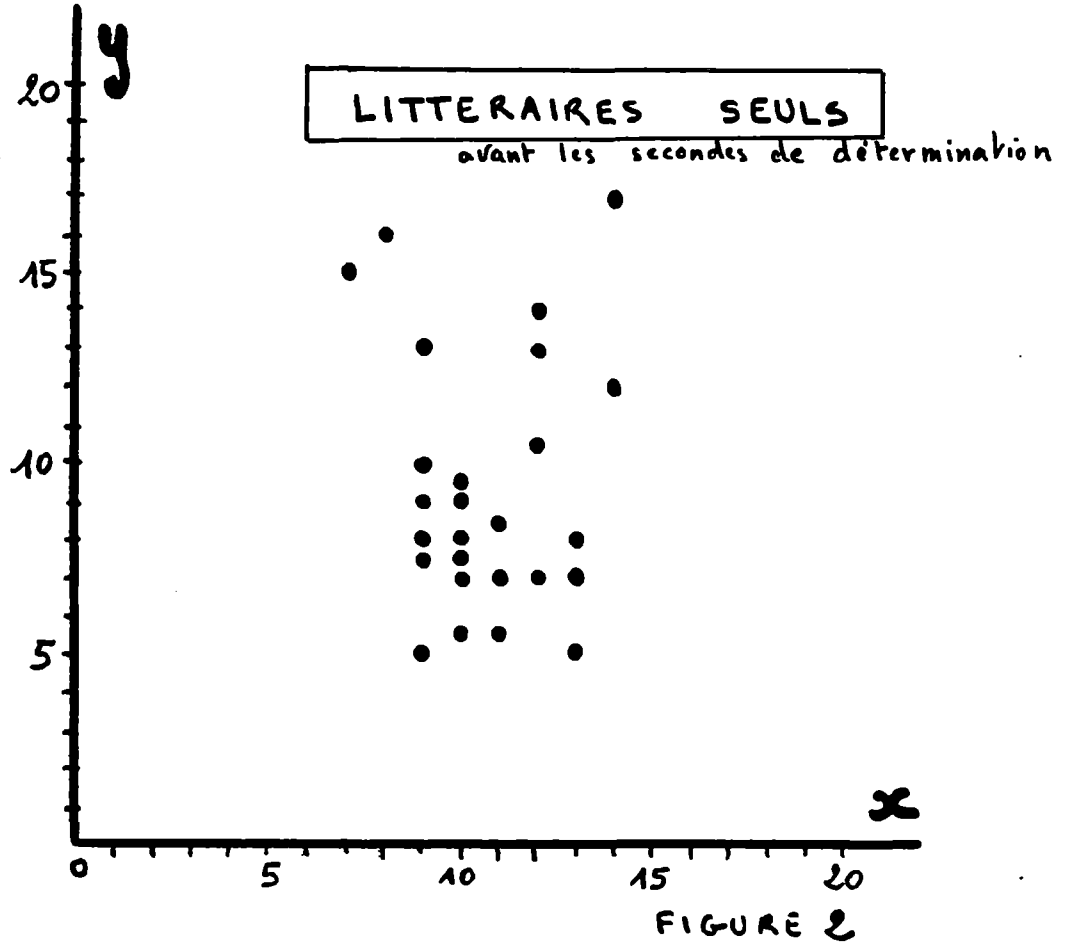


FIGURE 2

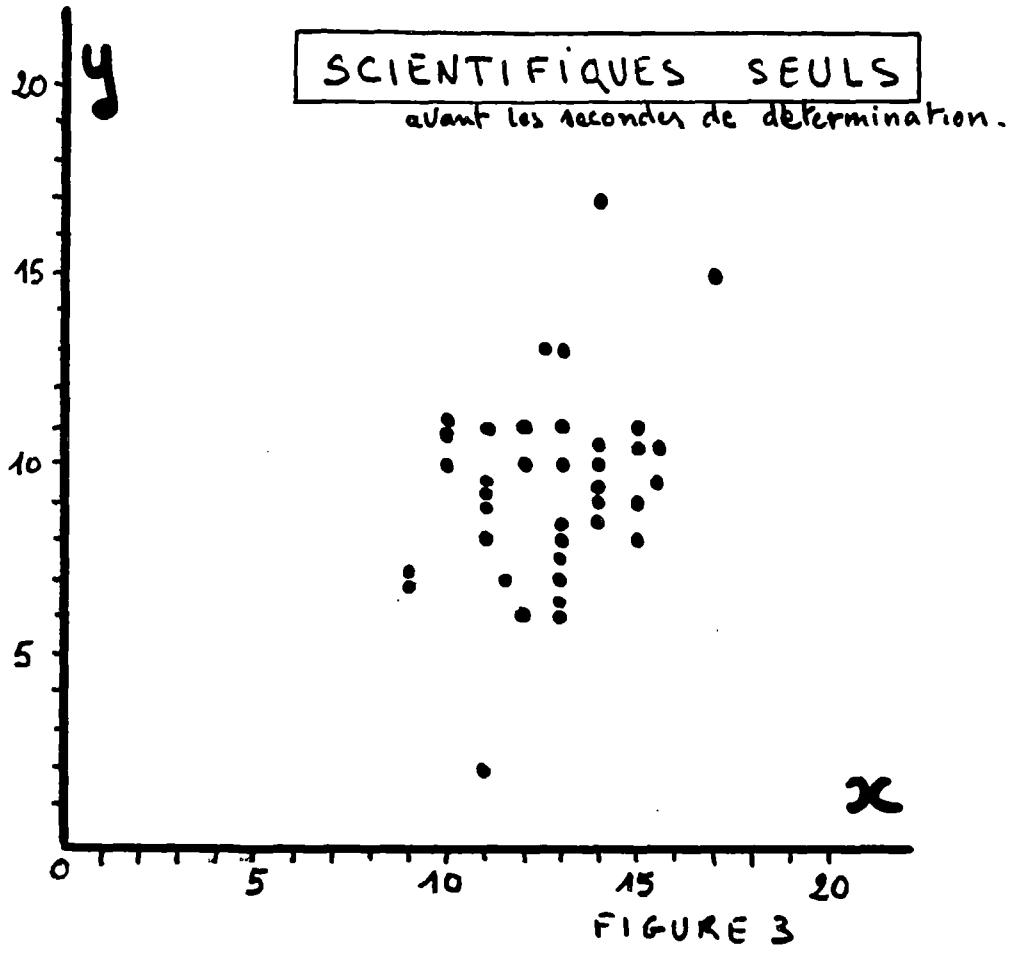


FIGURE 3

2ème période : Seconde de détermination

Cela ne concerne que les élèves qui étaient au collège en 80-81. Les résultats sont regroupés figure 4.

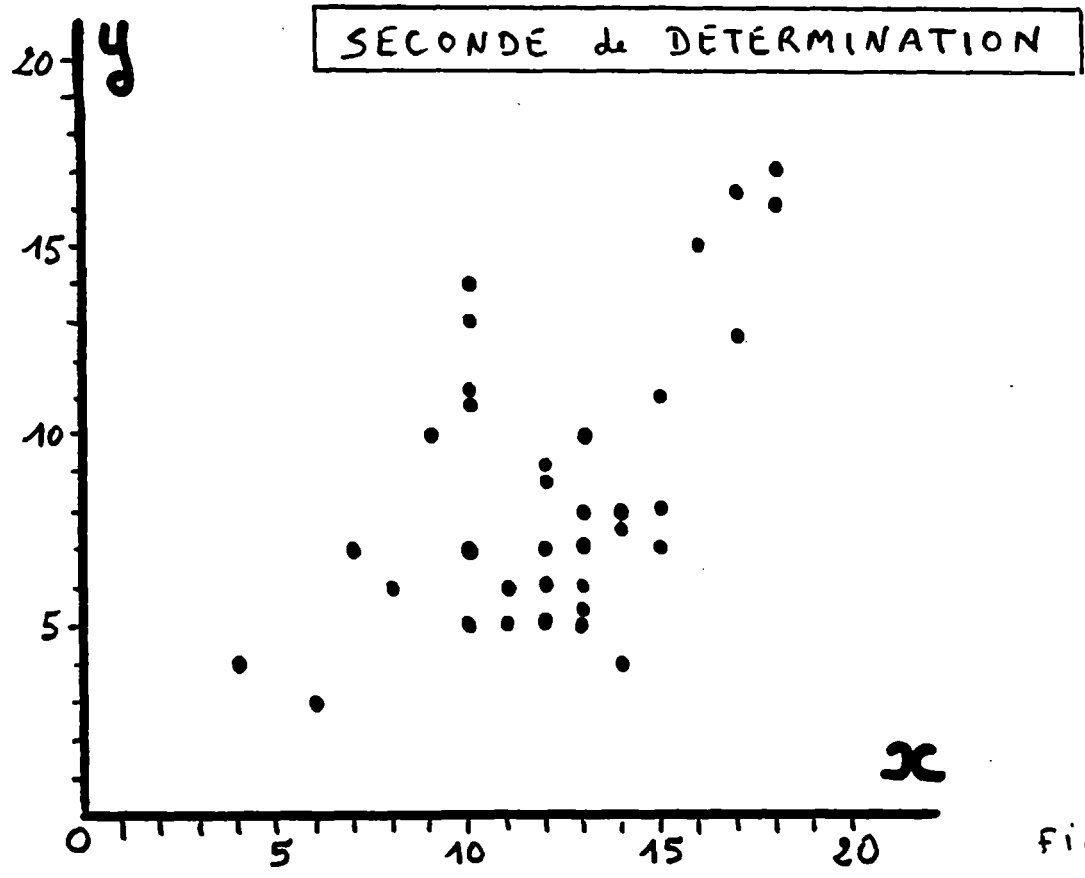


FIGURE 4

Le coefficient de corrélation est dans ce cas de 0,54. Les variables x et y sont donc mieux corrélées que les années précédentes, mais cette corrélation reste faibles. La droite de régression a pour équation :

$$y = 2,75 + 0,51\ x$$

ce qui indique assez bien la meilleure prise en compte des résultats de troisième.

L'effectif de ce sous-échantillon est cependant trop faible pour que l'on puisse se risquer à des conclusions définitives. Il est d'ailleurs possible, que plus que l'introduction des secondes de détermination, ce soit l'utilisation concertée des épreuves étalonnées de l'IREM qui soit en partie responsable de cette amélioration (meilleure harmonisation des critères).

4ème question : Les divergences constatées en mathématiques se retrouvent-elles dans d'autres disciplines ?

Nous avons fait en parallèle la même étude à propos des notes de FRANCAIS, nous ne pouvons en donner ici le détail mais dans l'ensemble les résultats sont tout à fait comparables à ce que nous obtenons en mathématiques : grande dispersion des écarts et chûte moyenne sensible (2 points environ).

5ème question : Les divergences constatées ne tiendraient-elles pas à des difficultés temporaires d'adaptation à un nouveau cycle d'enseignement ? Ces divergences étant par la suite appelées à s'estomper ?

Pour avoir une petite idée sur cette question, nous avons cherché à comparer les variables x et z :

z étant la note de mathématiques du troisième trimestre de seconde.

Nous ne pouvions malheureusement disposer de z que pour 50 élèves, que l'on peut en première approximation considérer comme représentatifs de l'ensemble. On obtient alors la représentation graphique de la figure 5. Notre hypothèse optimiste n'est certainement pas à retenir. Il semble plutôt que les critères d'appréciation soient radicalement différents au collège et au lycée.

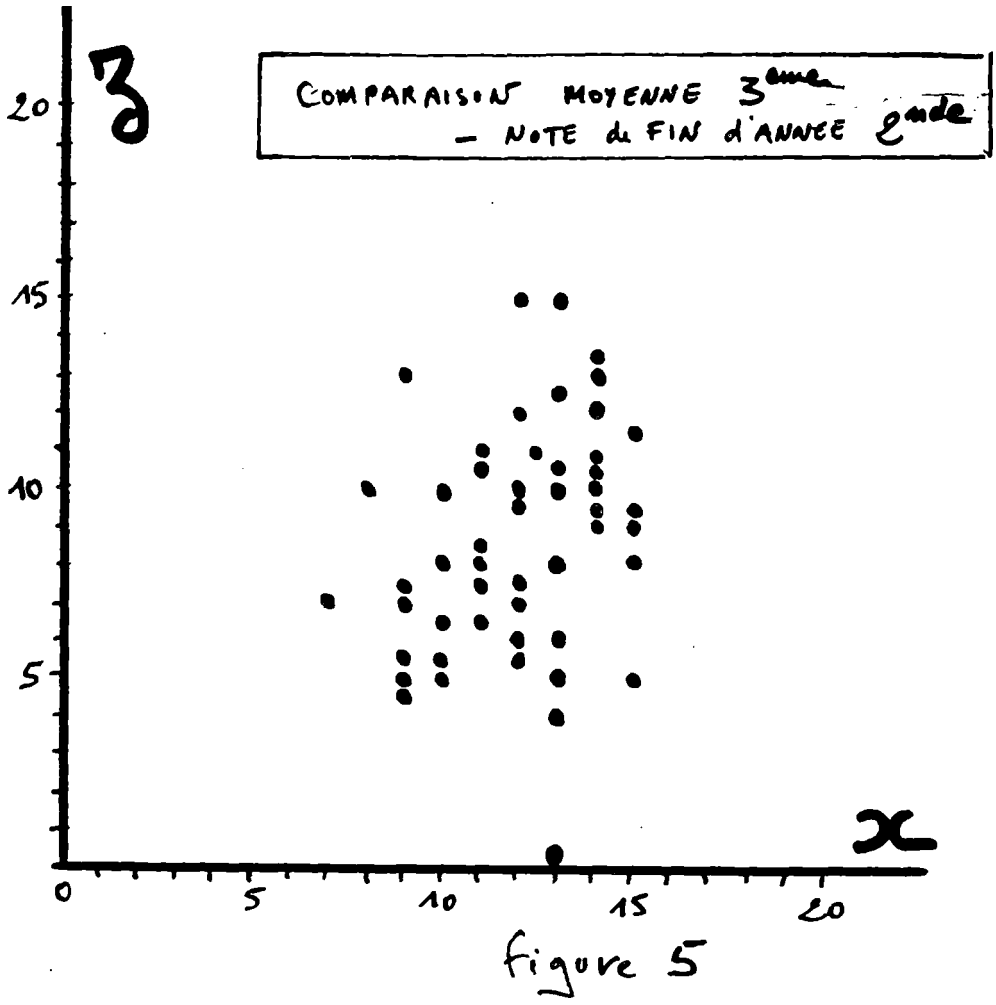


figure 5

6ème question : Les divergences enregistrées ont-elles vraiment un rapport avec le changement de cycle d'enseignement ? Ne pourrait-on pas observer le même phénomène entre la 4ème et la 3ème ? Ou entre le début et la fin d'une année scolaire (même professeur) ?

En ce qui concerne le collège, on observe une forte corrélation entre les notes du 1er trimestre de troisième et celles du troisième trimestre. Pour ce qui est de la transition 4ème-3ème nous n'avons pas fait d'étude précise, mais les divergences importantes semblent assez rares. Il semble plutôt qu'au niveau du collège il y ait un consensus implicite pour l'évaluation du niveau mathématique des élèves.

Ne possédant pas les résultats obtenus en classe de première, nous avons eu l'idée de comparer les notes du 1er trimestre de seconde (y) avec celles du 3ème trimestre de seconde (z). Nous avons eu la surprise de constater que les variables y et z sont aussi peu corrélées que le sont les variables x et y. Nous ne chercherons pas à expliquer ce résultat mais il nous a semblé qu'il valait la peine d'être signalé.

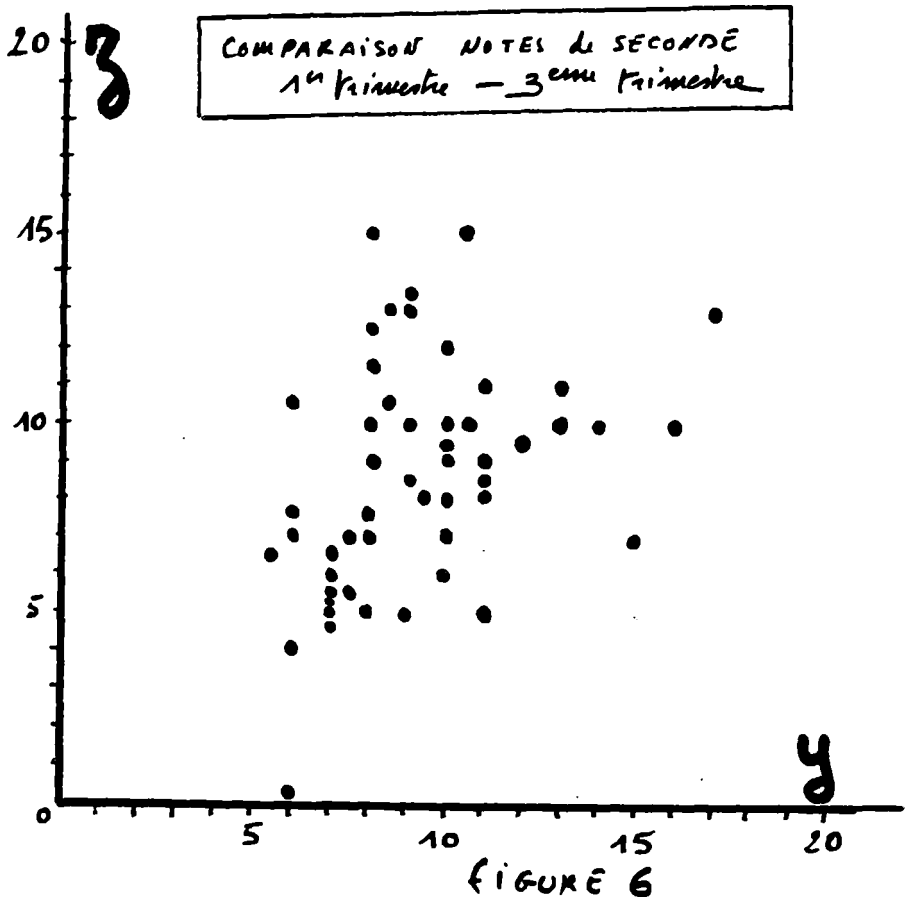


FIGURE 6

7ème question : Dans cet article, les mots NIVEAU, APTITUDES, CAPACITES ont été utilisés. Il y est donc sous-entendu que les notes trimestrielles attribuées aux élèves prétendaient constituer des indicateurs de NIVEAU, d'APTITUDE, de CAPACITES. Et si ce n'était pas le cas ? Si par exemple les notes ne servaient qu'à évaluer les performances des élèves par rapport à certaines tâches particulières, dont la nature serait nécessairement variable d'un professeur à l'autre, d'un trimestre à l'autre Où si les notes évaluaient surtout la qualité de participation des élèves, leur bonne volonté, pendant une période déterminée ? D'ailleurs les appréciations ne sont-elles pas là pour apporter tous les correctifs nécessaires ?

Force est de constater que dans bien des cas, en particulier pour l'orientation et l'affectation, les notes sont effectivement considérées comme des indicateurs de niveau, voire d'aptitudes. D'autre part, il est bien rare que les appréciations apportent une information contredisant l' idée que l'on peut se faire sur le niveau ou les capacités d'un élève à la seule lecture de sa note.

8ème question : Le collège a ses objectifs propres qui ne sont pas ceux du lycée. Pourquoi, dans ces conditions, s'étonner des divergences constatées ?

Il est certain que le collège poursuit des objectifs qui lui sont propres. La mathématique au collège doit-elle pour autant se constituer en discipline distincte de la mathématique au lycée ? Les objectifs des uns et des autres étant le plus souvent implicites, un effort de clarification s'impose. En particulier, l'analyse des prérequis à l'enseignement en seconde d'une part, l'analyse des acquis des élèves admis en seconde d'autre part, devraient conduire à un certain nombre d'ajustements.

Pour illustrer ce propos signalons que nous avons constaté que nombre d'élèves étaient mis en échec en sixième sur des calculs de quotients de nombres décimaux, à 10^{-2} près, ou sur l'utilisation du rapporteur. Ces notions étaient considérées comme acquises par les professeurs de collège alors qu'elles ne figurent plus dans les objectifs du cycle élémentaire.

En conclusion : Malgré l'abondance des questions posées et des réponses ébauchées, cet article avait pour but de présenter une situation. D'autres questions peuvent être posées, d'autres réponses peuvent être proposées. Nous pensons simplement que le problème existe et qu'il est important.

Dans son ouvrage "EXAMENS et DOCIMOLOGIE", H. PIERRON pouvait écrire après une étude sur le baccalauréat :

"Pour prédire la note d'un candidat, il vaut mieux connaitre son examinateur que lui-même".

Faudra-t-il dire maintenant que pour pouvoir prédire les chances d'adaptation d'un élève en seconde, il vaudrait mieux connaitre le couple de professeurs (professeur de troisième - professeur de seconde) que les notes qu'il a pu obtenir au collège ?

Cette étude ayant été menée à partir d'un seul collège, il est certain qu'il conviendrait de la reprendre à partir d'autres collèges. Il faudrait toutefois se garder des jugements hatifs : tous les collèges ne sont pas également sélectifs et il ne faudrait pas conclure que les "bons collèges" sont ceux qui envoient peu d'élèves en seconde.

Antoine BODIN

0

0 0

B I B L I O G R A P H I E

Il nous a semblé utile de donner quelques références bibliographiques, les ouvrages indiqués se trouvent à la bibliothèque de l'I.R.E.M. ou au C.R.D.P.

- PIERRON H., EXAMENS ET DOCIMOLOGIE - P.U.F.

- DE LANDSHEERE G., EVALUATION CONTINUE ET EXAMENS PRECIS DE DOCIMOLOGIE - NATHAN

- DE LANDSHEERE G. et V., DEFINIR LES OBJECTIFS DE L'EDUCATION - P.U.F.

- MAGER R.F., Comment définir des objectifs pédagogiques - BORDAS

- VANDEVELDE L. et VANDERELST P., Peut-on préciser les objectifs en éducation ? - NATHAN

- NOIZET G. et CAVERNI J.P., Psychologie de l'évaluation scolaire - P.U.F.

- IREM de PICARDIE : EVALUATION, docimologie, orientation, taxonomie

- IREM d'ORLEANS : Pédagogie par objectifs ou objectifs en pédagogie Actes du colloque d'Orléans (1977)

- BIGARD A., Mathématiques, échec et sélection - CEDIC

- GRAS R., Vers un programme éducatif par objectifs en mathématiques - IREM DE RENNES

- GRAS R., Thèse de doctorat - UNIVERSITE DE RENNES 1

- PLUVINAGE F., Thèse de doctorat - UNIVERSITE DE STRASBOURG

- PERETTI (DE) et Coll. - Recueil d'instruments et de processus d'évaluation formative. PARIS, CNDP, 1980, 2 tomes, 1028p.

- REUCHLIN (M.) - Problèmes d'évaluation. In DEBESSE (M.), MIALARET (G.) - Traité des sciences pédagogiques, tome 4, Psychologie de l'éducation. PARIS, PUF, 1974, p. 207-236.

- CAVERNI (J.P.), NOIZET (G.) - Les comportements des évaluateurs dans l'évaluation scolaire continue. L'Orientation scolaire et professionnelle, 1978, 7, N°2, p. 175-195.

- NOIZET (G.), CAVERNI (J.P.) - Les procédures d'évaluation ont-elles leur part de responsabilité dans l'échec scolaire ? Revue Française de Pédagogie, 1983, N°62 (Janvier-Février), p. 7-14.

- DE KETELE J.M. - Observer pour éduquer - Collection exploration recherches en sciences de l'éducation. Peter LANG - 1980.

- HAMELINE Daniel - Les objectifs pédagogiques en formation initiale et en formation continue. Editions EJF/Entreprise moderne d'édition - 1980.

- ALLAL L., CARDINET J., PERRENOUD P. - L'évaluation formative dans un enseignement différencié. Collection exploration cours et contribution pour les sciences de l'éducation. Peter LANG - 1979.

- KAUFMANN J. - Etude de la validité d'une batterie de tests. C.R.D.P. - BESANCON.

- C.I.E.A.E.M. 1977 : Evaluation et enseignement mathématique.

- INRP : Enquête sur l'enseignement des mathématiques à l'école élémentaire.

0

0 0

TABLE DES MATIERES

0

0 0

I.R.E.M. de Franche-Comté

UFR des Sciences et Techniques

16, route de Gray, La Bouloie

F-25030 BESANÇON cedex

Tél. : 03.81.66.61.92 - Fax : 03.81.66.61.99

Courrier électronique : iremfc@math.univ-fcomte.fr

http://pegase.univ-fcomte.fr/CTU/IREM/lieux.htm#Besançon

Fascicule 1

MAI 1983

IREM DE BESANÇON

Dépot Légal: 30/83

2ème Trimestre 1983

Réédition 1995