EVALUATION CONTINUE ET EXAMENS

PRÉCIS DE DOCIMOLOGIE

Gilbert DE LANDSHEERE

EDUCATION 2000

LABOR

EVALUATION CONTINUE ET EXAMENS

Précis de docimologie

DU MÊME AUTEUR

Les tests de connaissances, Bruxelles, Editest, 1965. Ouvrage traduit en espagnol.

Rendement de l'enseignement des mathématiques dans douze pays (en collaboration avec T.N. Postlethwaite), Paris, Institut Pédagogique National, 1969.

H. BENJAMIN, *La pédagogie paléolithique ou Préhistoire de la contestation,* Préface et adaptation française, Paris, Nathan; Bruxelles, Labor, 1970. Ouvrage traduit en italien.

Le test de closure, mesure de la lisibilité et de la compréhension, Paris, Nathan; Bruxelles, Labor, 1973, 2e éd.

Recherches sur les handicaps socioculturels de 0 à 7-8 ans, Editeur de cet ouvrage collectif, Bruxelles, Ministère de l'Education nationale, 1973.

Towards a Science of Teaching (en collaboration avec G. NUTHALL *et al.*), Londres, NFER, 1973

Comment les maîtres enseignent. Analyse des interactions verbales en classe (en collaboration avec E. BAYER), Bruxelles, Ministère de l'Education nationale, 1981, 4e éd. Ouvrage traduit en espagnol et en italien.

La formation des enseignants demain (avec la collaboration de S. DE COSTER, W. DE COSTER, F. HOTYAT), Paris, Casterman, 1976. Ouvrage traduit en espagnol et en italien.

Construire des échelles d'évaluation descriptives (en collaboration avec R. DE BAL et J. BECKERS), Bruxelles, Ministère de l'Education nationale, 1976. Ouvrage traduit en italien.

Dictionnaire de l'évaluation et de la recherche en éducation, Paris, Presses Universitaires de France, Coll. «Les grands dictionnaires», 1992, 2e éd. Ouvrage traduit en espagnol.

Les comportements non verbaux de l'enseignant (en collaboration avec A. DELCHAMBRE), Paris, Nathan; Bruxelles, Labor, 1979. Ouvrage traduit en italien.

Evaluation continue et examens. Précis de docimologie, Paris, Nathan; Bruxelles, Labor, 1980, 5e éd. Ouvrage traduit en italien, espagnol, roumain, portugais, grec.

Introduction à la recherche en éducation, Paris, A. Colin-Bourrelier; Liège, Dessain, 1982, 5e éd. Ouvrage traduit en allemand, espagnol, italien, néerlandais.

Définir les objectifs de l'éducation (en collaboration avec V. DE LANDSHEERE), Paris, Presses Universitaires de France; Liège, Dessain, 1992, 7e éd. Ouvrage traduit en espagnol, italien, portugais, anglais, roumain.

La recherche expérimentale en éducation, Paris, Unesco; Lausanne, Delachaux et Niestlé, 1982. Ouvrage traduit en italien, espagnol, portugais, anglais, chinois.

La recherche en éducation dans le monde, Paris, Presses Universitaires de France, 1986. Ouvrage traduit en italien.

Gilbert DE LANDSHEERE
Professeur à l'Université de Liège

EVALUATION CONTINUE ET EXAMENS

PRECIS DE DOCIMOLOGIE

SIXIÈME EDITION REVUE ET AUGMENTÉE

© 1992 by Editions LABOR, ch. de Haecht 156, 1030 Bruxelles
Photocomposition: RS - LIEGE (Wandre)

D/1992/258/112

ISBN 2-8040-0813-4

PREFACE DE LA 6e EDITION

Depuis la première édition de ce Précis, en 1971, et parallèlement au bond en avant que la recherche en éducation a fait pendant cette période, la théorie et la pratique de l'évaluation ont remarquablement progressé. L'ordinateur n'a pas peu contribué à ce développement.

On a de mieux en mieux perçu combien les évaluations normatives - celles qui informent sur la place qu'un élève occupe par rapport aux autres - sont de peu d'utilité pour le pilotage des apprentissages. Pour aider l'élève à se développer, lui dire que d'autres font mieux ou moins bien que lui ne l'avance pas beaucoup. L'important est d'entretenir chez lui le désir d'apprendre et de lui indiquer ou, mieux, de lui faire découvrir où il en est dans son parcours vers le savoir, le savoir-faire ou le savoir-être (ce que font les épreuves critérielles), et en quoi consistent les écueils auxquels il se heurte. L'évaluation formative répond à ce besoin.

Les épreuves critérielles et l'évaluation des compétences minimales sont donc appelées à jouer un rôle croissant dans le processus éducatif.

On perçoit aussi de mieux en mieux le caractère superficiel de bien des questions posées. En effet, même s'il s'agit d'un problème, elles ne concernent trop souvent qu'un cas isolé, présenté sous une seule forme, dans une seule condition, et à un seul moment. La théorie de la généralisabilité s'efforce de lever cette hypothèque.

Par ailleurs, une évaluation dépourvue de validité prédictive est sans grand intérêt. Savoir que l'étudiant connaît les principales règles de la grammaire d'une langue étrangère et réussit les exercices de contrôle à leur propos n'est utile que dans la mesure où l'on établit que l'acquis peut effectivement servir à la prise d'informations ou à la communication.

A côté de la science de l'éducation, la psychologie, en particulier la psychologie cognitive, a fait des progrès considérables au cours de ces dernières décennies. Ils ont des implications importantes pour l'évaluation.

Esquissant une nouvelle théorie des examens, R.J. Mislevy[1] écrit: «Les élèves n'accroissent pas leurs compétences en accumulant simplement des habiletés et des faits nouveaux, mais en donnant une forme nouvelle à leurs structures de connaissances, en automatisant les procédures, en comprimant l'information pour réduire la charge de la mémoire, et en développant des stratégies et des modèles qui leur disent quand et comment les faits sont appropriés. Les types d'observation et les configurations de données qui reflètent les façons dont l'étudiant pense, réalise et apprend ne sont pas couverts par les examens traditionnels.»

Une telle constatation n'implique pas que toutes les méthodes de construction des examens ou tests que nous connaissons sont inadéquates, mais bien qu'il faut les appliquer à des modèles cognitifs bien choisis.

A côté des améliorations qui viennent d'être évoquées, bien d'autres restent à apporter, à commencer par une meilleure diffusion des progrès théoriques de la docimologie dans la pratique scolaire. La productivité du système éducatif global et le trajet scolaire de beaucoup d'élèves, en particulier, souffrent considérablement d'erreurs fondamentales que l'on continue à commettre dans le domaine de l'évaluation. Cette insuffisance est due à plusieurs facteurs dont les effets se conjuguent: carences dans la formation initiale des enseignants, faiblesse de leur formation continuée, manque d'information des décideurs, insuffisance notoire des investissements dans les activités de développement d'instruments d'évaluation, manque d'évaluation des rendements scolaires, esquive de l'obligation de rendre des comptes aux membres de la communauté éducative.

La forme de civilisation dans laquelle nous entrons ne sera apogée historique que dans la mesure où le droit de tous à profiter pleinement de ses apports sera respecté - ce qui exige de nouvelles formes de solidarité-, et où tous seront bien préparés à faire usage de ce droit - ce qui exige une éducation réussie. Une évaluation constructive est l'un des principaux moyens d'y parvenir.

1 R.J. MISLEVY, Foundations for a new test theory. *In* N. FREDERIKSEN ET I. BEJAR, *Test theory for a new generation of tests*, Lawrence Erlbaum et ass., 1990.

Selon J. Stewart et ses associés[1], les principes généraux guidant l'évaluation doivent être les suivants:

1. L'intérêt des élèves doit primer. L'évaluation doit être conçue et exécutée de façon à maximiser les bénéfices pour les élèves et à en minimiser les effets négatifs.

2. La première raison d'être de l'évaluation doit être de fournir une information permettant d'identifier les points forts et de guider les améliorations. En d'autres termes, l'évaluation doit suggérer des actions susceptibles d'améliorer le développement éducationnel des élèves et la qualité des programmes éducatifs.

3. Les résultats de l'évaluation ne doivent pas être utilisés pour porter des jugements ou prendre des décisions politiques susceptibles de nuire aux élèves ou à l'efficacité des enseignants.

4. Tout doit être mis en œuvre pour assurer à tous une évaluation correcte.

5. La participation de la communauté éducative est essentielle pour la crédibilité et l'effet des procédures d'évaluation. Toutes les parties intéressées devraient avoir l'occasion de se faire entendre. L'auto-évaluation doit être le point de départ.

6. Il importe d'envisager soigneusement les effets que les pratiques d'évaluation peuvent exercer sur la motivation. Si ces pratiques ont un effet négatif (pour les élèves, les professeurs, les écoles ou le système éducatif), elles sont presque certainement indésirables.

7. Une évaluation correcte exige que l'on tienne compte de beaucoup de facteurs.

8. Dans l'évaluation des apprentissages cognitifs, il importe d'accorder beaucoup d'attention aux habiletés complexes, telles que la compréhension des principes, l'application des connaissances et des habiletés à des tâches nouvelles (transfert), à l'examen, l'analyse, la discussion de questions et de problèmes complexes.

9. Il faut accorder plus d'importance au développement de l'individu et aux progrès accomplis, qu'à la façon dont il se situe par rapport aux autres.

1 J. STEWART et al., *Assessment for better learning*, Wellington, Ministère de l'Education, 1989, p. 19.

10. Le choix des informations relatives à l'évaluation qui seront diffusées dans le public détermine le bénéfice ou la nuisance qui peut en résulter. C'est pourquoi les raisons, la sélection, la présentation et la diffusion des informations doivent respecter les principes ci-dessus.

INTRODUCTION

De nouveaux modes d'évaluation, dont la raison d'être est plus d'aider que de juger, entrent dans nos mœurs pédagogiques. Les résultats de l'*observation continue* prennent progressivement le pas sur l'examen de fin d'année.

Ces innovations sont louables. Elles ne traduisent nullement la volonté de supprimer la *mesure* à l'école, mais, au contraire, de la rendre plus utile et moralement plus juste et scientifiquement plus exacte.

Ce changement déborde largement le domaine de la notation et coïncide avec des transformations profondes de l'éducation. L'événement n'est pas fortuit. La civilisation contemporaine, l'économie de notre société réclament un homme pourvu de qualités et d'habiletés nouvelles: en pareil cas, les innovations pédagogiques s'imposent irrésistiblement.

La contestation des examens traditionnels a toutefois créé un malentendu grave. Il ressemble de façon frappante à celui qui a surgi, entre 1920 et 1940 surtout, lorsque l'adoption des idées, mal digérées, de la pédagogie fonctionnelle des Claparède, Dewey, et autres «progressistes» conduisit au culte de l'improvisation, au mépris de la discipline rigoureuse, au centrage sur des intérêts n'existant que dans l'esprit des théoriciens...

Avec l'avènement de J. Dewey, l'école aurait pu prendre comme devise: «L'effort est mort, vive l'effort!» De même, serions-nous tentés de proposer aujourd'hui: «Les examens meurent, les examens sont morts, vivent les examens!»

Assurément, les vieux examens qui empoisonnaient l'atmosphère et la matière de cycles entiers d'études et décidaient de la carrière scolaire ou professionnelle en quelques heures - voire en quelques minutes! - doivent disparaître. Assurément aussi, la majeure partie des travaux docimologiques publiés jusqu'ici n'ont été que mises au point

de systèmes perfectionnés pour continuer à mal faire les choses. Car ils ne s'appuyaient pas sur une remise en cause de tout l'enseignement.

Mais si, par contre, on ménage, dans chaque activité d'enseignement de base, la place qui revient à l'évaluation, au *feed-back,* les examens existent alors, de façon *quasi* permanente dans la vie scolaire. Tantôt ils consistent en brèves évaluations, tantôt ils prennent la forme d'épreuves plus longues, couvrant parfois des quantités considérables de matière.

L'essentiel est qu'ils ne s'insèrent pas en corps étrangers dans le processus d'éducation, mais qu'ils en fassent partie intégrante. Pour cette même raison d'ailleurs, ils ne se cantonnent donc pas étroitement au domaine de la connaissance, mais visent à saisir l'ensemble de la personnalité.

Beaucoup d'enseignants se méfient encore des tests et des autres instruments de mesure utilisés dans le domaine de l'éducation. Non sans raison d'ailleurs, car ces outils sont encore loin d'être parfaits. Et comme leurs utilisateurs n'ont pas toujours su compenser l'imperfection des techniques par une grande modération de jugement et une intelligence profonde des situations d'ensemble, bien des erreurs ont été commises.

Mais si la prudence reste nécessaire, si l'esprit critique et le sens clinique ne perdront jamais leurs droits, des progrès considérables ont néanmoins été accomplis: on peut maintenant mesurer bien des comportements humains de façon satisfaisante.

D'aucuns rejettent la mesure objective, non plus parce que toute validité lui fait défaut, mais parce qu'ils n'en comprennent pas l'économie. Dénier toute valeur à ce qu'on ignore est un réflexe de défense bien connu. L'aspect mathématique des méthodes à adopter et de la littérature expérimentale, souvent indigeste, vient encore compliquer la situation.

Le temps de l'opposition entre évaluateurs et praticiens de l'éducation est révolu. Les premiers doivent humaniser leurs chiffres; les seconds, introduire plus de rigueur dans leurs procédures; tous deux doivent unir et harmoniser leurs efforts pour le plus grand profit de l'étudiant et de la communauté.

Que ceux qui, en feuilletant ce livre, ont aperçu quelques chiffres se rassurent: les quatre opérations arithmétiques seront un bagage suffisant!

Comme d'habitude, les procédés dont il va être question sont plus difficiles à expliquer qu'à appliquer.

Certes, tous les secrets de la mesure et de la statistique ne seront pas livrés, mais les notions simples que nous allons rencontrer ont une valeur pratique éprouvée. Elles détiennent, en outre, un indéniable pouvoir de démythification et faciliteront ultérieurement la lecture de travaux docimologiques plus spécialisés.

★
★ ★

PREMIERE PARTIE

DEFINITIONS

I. DOCIMOLOGIE, DOCIMASTIQUE ET PSYCHOLOGIE DE L'EVALUATION

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.

La *docimastique* est la technique des examens.

Au début, la docimologie a revêtu un caractère négatif en critiquant les modes de notation et en montrant expérimentalement le manque de fidélité et de validité des examens.

Par la suite, elle est entrée dans une phase constructive en essayant de proposer des méthodes et des techniques de mesure plus objectives ou, au moins, plus rigoureuses, et en mettant au point les moyens de rendre les notes comparables, de façon à assurer plus de justice scolaire.

Plus récemment encore, on a perçu l'intérêt considérable qu'offre l'étude du comportement de l'évaluateur utilisant les méthodes ou techniques proposées, et de l'élève ou plus généralement du sujet soumis à l'évaluation. On arrive ainsi à une *psychologie de l'évaluation*[1].

J. Guillaumin[2] lui assigne, notamment, les objectifs suivants:

- étude des effets inhibiteurs ou stimulants des différentes formes d'examens;
- étude des réactions émotionnelles des élèves et, de là, des réactions intellectuelles aux jugements du maître;
- étude des effets de l'opinion du maître concernant les élèves sur son enseignement et sur l'apprentissage scolaire;

1 NOIZET et J.-P. CAVERNI, *Psychologie de l'évaluation scolaire*, Paris, P.U.F., 1978.
2 J. GUILLAUMIN, L'aspect interpersonnel de la notation: de la docimologie à la doxologie pédagogique, in *Bulletin de la Société A. Binet et T. Simon*, 86, 1968, 250-275.

- étude des processus mis en jeu et des effets obtenus par l'automatisation, par l'internotation, par la notation d'équipe, par l'absence de notation.

G. Noizet et J.-J. Bonniol introduisent une dernière nuance de terminologie. Ils écrivent: «Si l'investigation docimologique dépasse le plan du constat, si elle permet un perfectionnement du système d'estimation par la connaissance expérimentale des mécanismes en jeu et des causes de distorsion dans leur fonctionnement, alors la docimologie devient en fait une docimonomie[1].»

Ces auteurs forgent ce néologisme par analogie avec l'ergonomie, ainsi appelée parce qu'elle ne se limite pas à l'étude des systèmes hommes-machines (cas où elle serait «ergologie»), mais recherche aussi leur perfectionnement.

On constate toutefois que toutes ces nuances de vocabulaire restent l'apanage de petits groupes de spécialistes. Le mot docimologie a conquis droit de cité; il est pratiquement le seul à être couramment utilisé[2].

II. EXAMENS ET CONCOURS OBSERVATION ET EVALUATION CONTINUES

L'examen et le concours sont difficilement dissociables de l'idée d'épreuve, devenue d'ailleurs leur synonyme après avoir désigné plus généralement la souffrance, le malheur, le danger qui révèlent le courage et la résistance.

Dans l'examen, l'admission est déterminée par une note (fixée a priori ou a posteriori) que le candidat doit atteindre ou dépasser, tandis que dans le concours, le nombre de places proposées, et donc de réussites, est fixé d'avance. La présence de la *menace*, du *danger*, celui du refus, de l'échec, est, dans les deux cas, indéniable. Pour peu que la procédure soit entachée d'imperfections graves, nous ne sommes pas très éloignés de l'ordalie.

Dans la notion de concours et d'examen, nous percevons une charge agressive totalement absente des concepts d'observation et

1 G. NOIZET et J.-J. BONNIOL, Pour une docimologie expérimentale, in *Bulletin de Psychologie*, 1968-1969, pp. 782-787.

2 On observe aussi un usage abusif du mot «docimologie». Certains ne déclarent-ils pas pratiquer la «docimologie», voire la «nouvelle docimologie», pour dire qu'ils ont adopté un nouveau système d'évaluation.

d'évaluation formative continues. La sérénité, la bienveillance, l'indulgence aussi, imprègnent le maître qui suit, avec une sympathie d'où toute sévérité n'est pas exclue, le long cheminement de ses élèves vers l'équilibre du moment et la pleine accession à l'état adulte, dans l'avenir.

Toute épreuve finit par un constat de réussite ou d'échec. La *compétence minimale* marque la limite entre les deux. La définir objectivement n'est pas une mince affaire.

Le mouvement actuel à propos des compétences minimales prend cependant une signification particulière. Malgré une prolongation importante de la scolarité dans les pays industrialisés, on constate que maints élèves quittent l'école sans maîtriser des habiletés de base comme la lecture, l'expression écrite et l'arithmétique élémentaire. C'est pourquoi apparaissent dans de nombreux pays des dispositions réglementaires subordonnant, par exemple, l'attribution d'un diplôme d'enseignement secondaire à la réussite d'une épreuve préliminaire établissant que l'élève maîtrise de façon suffisante les habiletés jugées essentielles[1].

Les examens marquent les fins d'étapes; les concours ouvrent les portes aux élus: ce sont des points dans la durée, des événements dans le processus de l'éducation.

L'apprentissage qui précède ces événements consiste fondamentalement en une succession continuelle de comportements et d'informations en retour éclairant sur leur validité, leur pertinence. Il ne nous appartient pas d'engager ici une longue discussion des modalités et des effets du renforcement des conduites. L'important, c'est que, sans lui, l'apprentissage ne semble pas pouvoir se produire.

L'*évaluation*, au sens restreint que nous lui donnons dans le présent ouvrage, *mérite donc une place importante dans l'enseignement, dont elle fait partie intégrante. Elle a toujours, directement ou indirectement, rapport avec le progrès, en extension ou en qualité, de l'apprentissage.*

L'évaluation joue trois rôles:

1° Un rôle pronostique: l'élève est-il pourvu des qualités cognitives et affectives, et des connaissances nécessaires pour aborder une matière nouvelle ou un cycle d'études supérieures? Est-il là où il doit se trouver? Répondre à ces questions équivaut à prédire le succès dans l'étape qui va commencer.

1 Pour un traitement approfondi des problèmes généraux et techniques que pose l'évaluation de la compétence minimale, voir V. De Landsheere, *Faire réussir - Faire échouer. L'évaluation de la compétence minimale*, Paris, P.U.F., 1988.

2° Un rôle de jaugeage:

a) contrôle des acquisitions;

b) évaluation du progrès, cas où l'on compare l'élève à lui-même (évaluation formative);

c) situation comparative de l'élève à un moment donné (évaluation normative):

- dans sa classe ou son groupe de travail?
- dans l'ensemble des classes parallèles d'une même école?
- dans des ensembles plus vastes: ville, canton, province, pays?

Il ne s'agit pas nécessairement de procéder à un examen ou à un concours, mais de faire le point, de déterminer la position relative.

3° Un rôle diagnostique:

Pourquoi un apprentissage parfait ne s'est-il pas produit?

Quelles matières ou techniques l'étudiant domine-t-il insuffisamment, quels sont les processus mentaux en cause?

Les instruments nécessaires à l'évaluation, à l'examen ou au concours ne diffèrent pas toujours, mais bien la façon de les utiliser. Aussi, pour éviter des précisions verbales fastidieuses emploierons-nous, dans les pages qui suivent, le mot examen dans deux sens différents, que le contexte éclairera toujours: examen proprement dit et, plus généralement, toute procédure pédagogique ayant une mesure ou une évaluation d'apprentissages ou de connaissances pour objet.

Examens internes et examens externes.

Au sens strict, l'*examen interne* dans une branche est construit et noté par le maître qui l'a enseignée, et subi par les élèves qui ont reçu cet enseignement, dans le cadre de la classe ou de l'école.

Au sens plus large, on qualifie d'internes les examens organisés indépendamment dans chaque école, qu'il existe ou non une coordination ou une unification par branche et par niveaux et sections.

Par *examens externes*, on désigne les épreuves organisées et notées par des jurys indépendants des écoles, à l'échelon local, régional ou national. Les plus connues de ces épreuves sont celles du baccalauréat français; citons encore, en Belgique, les examens cantonaux que l'on organisait en fin d'études primaires.

III. MESURE ET EVALUATION[1]

Selon J.-P. Guilford, mesurer c'est «assigner un nombre à un objet ou à un événement selon une règle logiquement acceptable»[2].

La mesure exige donc:

1° que les objets ou, plus exactement, les propriétés de ces objets soient clairement définis - dans toute la mesure du possible - par des comportements ou des caractéristiques observables (définitions opérationnelles);

2° qu'une règle indique comment faire correspondre un nombre à chaque objet.

En rigueur de termes, une mesure se traduit donc nécessairement en chiffres, alors que c'est loin d'être le cas pour l'évaluation.

Opposant mesure et évaluation, H. Taba écrit[3]:

«Le processus de mesure est fondamentalement descriptif, car il indique quantitativement à quel degré un trait est possédé. La mesure en éducation est généralement concentrée sur des caractéristiques spécifiques, étroites et bien définies. L'évaluation dépend de la mesure, mais elle porte sur un profil plus large de caractéristiques et de performances.»

Cette distinction, utile pour préciser les concepts, n'est pas acceptable sans une réserve importante. H. Taba suppose implicitement que toute évaluation passe par une quantification rigoureuse. Or, c'est loin d'être le cas, en particulier dans le domaine des attitudes et pour toutes les productions de grande complexité. Le visiteur d'une exposition de peinture classe les œuvres de l'artiste, et donc les évalue, selon son goût, ses critères personnels, sans devoir (ni pouvoir réellement) se référer d'abord à une ou plusieurs mesures. Bien des comportements et nombre d'œuvres humaines s'évaluent de pareille façon.

Comme le souligne J. Cardinet[4], les enseignants doivent avant tout s'intéresser à une information qualitative, du type diagnostique, car elle suggère des hypothèses de travail sur les démarches correctives à

1 Nous ne donnons ici que les définitions les plus générales. Pour une définition des différents types d'évaluation, voir G. DE LANDSHEERE, *Dictionnaire de l'évaluation*, Paris, P.U.F., 1992, 2e éd.

2 J.-P. GUILFORD et B. FRUCHTER, *Fundamental statistics in psychology and education*, New York, Mc Graw Hill, 1973, 5e éd., p. 19.

3 H. TABA, *Curriculum development*, New York, Harcourt, Brace and World, 1962.

4 J. CARDINET, L'évaluation en classe: mesure ou dialogue. *European Journal of Psychology of Education*, 1987, II, 133-144.

adopter. Ainsi conçue, l'évaluation en classe est prioritairement un processus de communication entre maître et élèves, aboutissant, à la suite d'un nombre imprévisible d'interactions, à l'atteinte des objectifs pédagogiques.

IV. LES TESTS

Un test est une épreuve standardisée dont la fidélité et la validité ont été établies et qui a été étalonnée de façon à permettre une interprétation objective des résultats observés. Classiquement, on distingue les tests d'intelligence, de connaissances et de personnalité. Dans le présent ouvrage, c'est exclusivement des épreuves de connaissances qu'il est question.

Un test peut être conçu à des fins pronostiques, descriptives (inventaires de connaissances) ou diagnostiques.

Un test est dit normatif si ses résultats sont interprétés en fonction de la distribution des résultats d'un groupe de référence. Par exemple, on dira qu'un élève ayant répondu correctement à une partie définie de l'examen se classerait dixième dans un groupe de cent condisciples de son âge fréquentant la même année d'études dans une école similaire à la sienne. Aspect capital : la norme est constituée par ce que font les autres; elle permet de classer, mais n'informe pas effectivement sur les acquis. Savoir que tel élève est sorti premier ou dernier de la dernière année d'études primaires n'informe pas sur ce que cet élève a appris.

Par opposition, un test est dit critériel si la performance individuelle n'est plus estimée en fonction de celles des autres, mais bien en fonction de la distance qui la sépare d'un objectif d'apprentissage dont une définition opérationnelle (ou opératoire) précise les critères qui permettent de dire dans quelle mesure il est atteint.

On reviendra par la suite sur ces notions essentielles.

Le testing adaptatif

Les test dits adaptatifs permettent d'individualiser l'examen.

Un tel test est constitué de questions ordonnées par ordre de difficulté croissante. Soit au hasard, soit en fonction de sa compétence estimée (éventuellement par des examens précédents), l'élève répond à une première question ou à un premier ensemble de questions. Le choix des questions qui suivront est commandé par la qualité des

réponses initialement données, et ainsi de suite. L'épreuve s'arrête quand la stabilité des réponses à un niveau donné permet de penser qu'il est bien celui que l'élève a atteint effectivement.

Grâce à l'ordinateur, les tests adaptatifs semblent appelés à jouer un rôle important dans un avenir proche.

La construction artisanale d'une telle épreuve se fait par tâtonnements et jugements. En s'appuyant sur son expérience et sur des résultats antérieurement observés, l'éducateur détermine empiriquement une hiérarchie de difficulté (qui réside à la fois dans le contenu objectif de la question et dans sa forme). Il existe aussi des techniques statistiques sophistiquées pour construire ce type d'instrument. Les plus connues procèdent de deux modèles d'interactions entre la compétence du sujet et la difficulté des questions : le modèle d'analyse hiérarchique de Guttman[1] et le modèle de Rasch[2].

V. NOTES ET SCORES

La distinction entre notes et scores rendrait, pensons-nous, service en éducation.

Dans une dictée, l'élève peut commettre un certain nombre de fautes dont il ne nous appartient guère d'évaluer l'existence : elles sont ou ne sont pas. Toutefois, le nombre de fautes relevées n'a, en soi, aucune signification éducative : faire cinq fautes dans la dictée de Mérimée ou de Pivot témoigne d'une extraordinaire connaissance des arcanes de l'orthographe; en d'autres conditions, le même résultat annonce une faiblesse grave. Une information relativiste est donc aussi nécessaire.

Par *score*, on désigne les résultats objectifs obtenus à un test ou à toute autre forme d'évaluation par *compte ou décompte de points, selon des règles fixes* : nombre de fautes en dictées, résultats à un test standardisé.

Par *note*, on entend une appréciation synthétique traduisant l'évaluation d'une performance dans le domaine de l'éducation.

1 B. MATALON, *L'analyse hiérarchique*, Paris, Gauthier-Villars, 1975.
2. D. LECLERCQ, Computerized tailored testing, *European Journal of Education*, 1980, 15, 3.

La note peut être objective ou subjective, mais elle est toujours relative. Attribuer la note A à un élève dont la performance se situe à tel niveau dans un étalonnage national relève de la première catégorie ; marquer sa composition d'un *bien* relève de la seconde.

Plus généralement, la *note vraie* (ou le *score vrai*) qui refléterait de façon parfaite, indiscutable, la valeur d'une compétence, d'une production ou de son auteur est une vue de l'esprit. Qui est capable d'exprimer par une note exacte, indiscutable, la connaissance qu'un individu possède de sa langue maternelle ou d'une habileté telle que jouer du piano ? On sait que cette connaissance existe, mais - hormis les cas extrêmes d'ignorance absolue - elle est inaccessible. Car on n'évalue jamais que par comparaison ou par une estimation en tout ou rien de l'adéquation à une fonction (par exemple : dans les conditions d'usage courant, ce bouchon ferme la bouteille de façon étanche). Pour s'en convaincre, il suffit d'observer l'embarras de celui à qui l'on demande d'évaluer un seul produit sans donner la moindre information complémentaire. Toutefois, une fois attribuée, la note tend à s'affranchir du contexte qui lui a donné sa signification et à être, par conséquent, interprétée comme « vraie ».

Le verbe *noter*, défini par Robert comme « apprécier par une observation, une note chiffrée », appartient naturellement à l'usage.

Chez plusieurs docimologistes français[1], *noter en positif* indique qu'un score est attribué par comptage de points ; *noter en négatif* indique une déduction de points. Par exemple, une dictée ou une version sont généralement notées en négatif ; par contre, pour des travaux de sciences ou de mathématiques, souvent le correcteur additionne les points attribués à mesure que certaines exigences ou certains critères sont satisfaits.

Enfin, le mot *cote* reste largement utilisé, notamment en Belgique, pour désigner une note chiffrée ou un score.

1 Voir notamment G. NOIZET et J.-J. BONNIOL, *o.c.*

DEUXIEME PARTIE

L'ACCUSATION ET LA DEFENSE

CHAPITRE 1

CRITIQUE DES EXAMENS

Une critique détaillée des examens traditionnels a été faite par H. Piéron dans son ouvrage fondateur *Examens et docimologie*[1]. Tout enseignant doit l'avoir lu.

Le présent chapitre contient deux types de remarques. Les unes sont de brefs rappels d'imperfections bien connues, inlassablement mises en lumière par les docimologistes; les autres concernent des phénomènes moins bien étudiés: stéréotypie, effet de halo, effet Pygmalion, ... Tantôt, seul l'examen est en cause ; tantôt, c'est toute la pédagogie dans laquelle il s'insère.

1. Corps étrangers dans l'éducation, au service d'une pédagogie dépassée.

Qu'il s'agisse d'interrogations périodiques ou d'examens trimestriels ou annuels, l'évaluation se réduit souvent à un contrôle de rétention de connaissances laissant inexplorés, non seulement les aspects les plus importants de l'intelligence et de la connaissance, mais aussi à peu près tous les traits de personnalité qu'une éducation bien comprise doit cultiver.

Nous nous trouvons, en fait, devant les séquelles d'un système pédagogique où, comme T. Brameld aimait à le dire, les leçons servent de bandes transporteuses de connaissances et de valeurs, sélectionnées en fonction d'un rôle prédéterminé à jouer dans une société non démocratique.

Dans ce cadre, l'examen constitue une sorte de contrôle de fabrication, de vérification de la conformité au moule, au gabarit, bref aux spécifications arbitrairement imposées par l'autorité. Le développement

1 Paris, P.U.F., 1963.

l'épanouissement de la personne humaine n'occupent guère de place dans ces préoccupations, car l'éducation n'est pas conçue pour elle, mais pour le service d'un régime.

Les conséquences malheureuses de cette situation ont été analysées mille fois. Au lieu de servir l'élève, de l' informer fonctionnellement de la valeur de ses comportements en cours d'apprentissage, de faire naître une adaptation et une motivation meilleures, d'apporter donc une évaluation acceptée au point que l'étudiant y participe sincèrement et spontanément, l'examen est mal accueilli.

En outre, le rejet profond de la procédure, combinant son effet avec la pauvreté intellectuelle de certaines questions posées, conduit à un résultat aisément imaginable: dans les quinze mois qui suivent un examen, un oubli de 80% des matières factuelles n'est pas exceptionnel.

Ainsi, si l'on excepte la fonction sociale jouée par l'examen, le fiasco est complet: on n'a ni éduqué, ni instruit.

2. **Anxiété et stress.**

Dans une civilisation où la réussite scolaire conditionne la réussite matérielle et sociale, au point que M. Young a cru devoir dénoncer les dangers de la «méritocratie», l'examen qui décide du passage de classe ou de l'obtention du diplôme est redouté par l'enfant et sa famille. Même en cours d'année, les épreuves de contrôle de connaissances et d'aptitudes sont abordées avec tension et appréhension, ce qui n'est certes pas la condition idéale, et dénature profondément le rôle éducatif de la mesure des apprentissages.

Circonstance aggravante, la tradition veut que, chez nous, les «grands» examens portant sur toutes les branches se déroulent consécutivement en quelques jours. Lors d'une session universitaire, j'eus, une année, le triste privilège d'être interrogé sur quatorze cours en deux jours de suite... Dans l'enseignement secondaire, le même phénomène de concentration se produit souvent aussi en cours d'année, lorsque l'approche de la remise aux parents du bulletin périodique suscite une floraison d'interrogations.

Dans les deux cas, l'accumulation des épreuves et leur préparation, souvent contrariées par l'apport de matières nouvelles jusqu'au dernier moment (des professeurs ne respectent pas toujours les périodes de révision ou ne les exploitent pas de la façon la plus heureuse), fatiguent considérablement le corps et l'esprit.

D'aucuns prétendent que ces conditions difficiles revêtent elles-mêmes une valeur éducative: la vie moderne ne nous épargne guère les *tensions* et il est bon d'y être préparé. D'accord, à condition que les choses se passent dans la clarté!

Avant d'organiser un examen, il importe de définir clairement son but: s'agit-il de contrôler l'acquisition des connaissances? La résistance à la *tension*? Ou la capacité de restituer et d'user des connaissances en situation de *stress*? Ces trois objectifs diffèrent et appellent des épreuves différentes.

3. **Inégalité - injustice.**

Dans notre système scolaire, les professeurs rédigent chacun les questions d'examen destinées à leurs élèves. Le principe est excellent. Toutefois, la liberté, à peu près totale, laissée à nos maîtres conduit à des situations injustes.

Les écoles de notre pays constituent des mondes relativement isolés dont les populations présentent des caractéristiques parfois fort différentes. Tel petit établissement d'enseignement secondaire ne compte que quelques élèves, en majorité défavorisés par leurs origines socio-économiques. A l'opposé, tel autre établissement situé dans une grande ville voit ses sections «fortes» peuplées d'une majorité favorisée socialement et intellectuellement, les adolescents en difficulté étant orientés vers une école voisine, qui a la réputation d'être plus compréhensive.

Avec le temps, et la concurrence aidant, des traditions de sévérité ou de «générosité» se créent dans les établissements. Marion Coulon observait que, quand des écoles de même type sont successivement créées dans une même région, la sévérité des examens est inversement proportionnelle à l'ancienneté de l'établissement. Sans vouloir généraliser, qui oserait affirmer que la proposition est dénuée de tout fondement? Beaucoup de parents l'ont compris, qui ambitionnent pour leur enfant un diplôme déterminé, sans trop se soucier de la valeur intrinsèque des études: en cas d'échec, il est plus aisé de changer d'école que d'installer un traitement ou de modifier l'orientation.

Quoi qu'il en soit, les maîtres adaptent leur enseignement et les examens au niveau de leur classe (ce qui est louable), avec la conséquence qu'un élève bien noté dans une population faible serait parmi les très faibles ou échouerait dans un groupe fort.

Il faut y insister, l'adaptation du maître au niveau de ses élèves n'est pas, en soi, critiquable; elle est, au contraire, caractéristique du bon éducateur. L'injustice apparaît lorsque des notes purement relatives sont utilisées en situation de concurrence extérieure, ou sont l'unique critère pour l'obtention d'un diplôme dont peut dépendre l'avenir des élèves.

De toute façon, on constate des variations considérables dans la *quantité des matières couvertes* et dans la *qualité des réponses exigées*. Ici, la restitution de mémoire d'une partie d'un court syllabus suffit, alors que là, les questions exigent analyse, synthèse et jugement personnel à propos d'une matière abondante.

Tel maître ne pose qu'une ou deux questions, plus choisies pour la facilité de correction qu'elles offrent que pour leur importance réelle, tel autre s'efforce de parcourir tout le cours, au risque de transformer l'examen en un marathon. En sport, on sait que cette longue course doit être réservée à des individus exceptionnels; pourquoi, dans le domaine de l'instruction, les trente élèves d'une classe seraient-ils tous des coureurs de fond?

Trois résultats de recherches aideront encore à illustrer les phénomènes dont il vient d'être question.

B.S. Bloom[1] a comparé les performances scolaires d'étudiants terminant l'enseignement secondaire supérieur dans 48 Etats différents des Etats-Unis. La performance moyenne entre l'Etat le mieux classé et l'Etat le moins bien classé est d'environ un écart type dans la distribution nationale des scores. Cet écart correspond en gros à quatre années d'études.

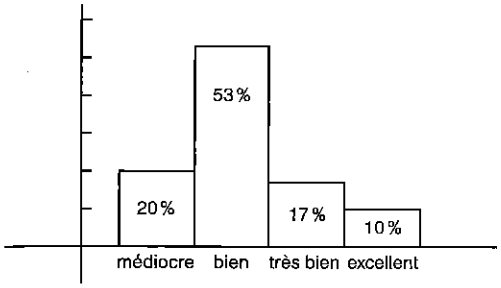
L'I.E.A.[2] a démontré dès 1973 que, parmi les nations industrialisées qui ont participé à ses recherches, la différence entre les scores moyens des pays aux performances les plus élevées et ceux des pays aux performances les plus basses est du même ordre (mathématique, science, littérature, lecture, anglais ou français comme langue étrangère). Avec le pays en voie de développement, participant à la recherche, et classé le plus bas, la différence s'élève à deux écarts types, soit environ l'équivalent de six années d'études.

1 B.S. BLOOM, The 1955 Normative Study of the Test of General Education Development, In *School Review*, 64, 1956, 110-124.
B.S. BLOOM and C. STATLER, Changes in the States of the Tests of General Education Development from 1943 to 1955, In *School Review*, 65, 1957, 204-221.
2 Association internationale pour l'évaluation du rendement scolaire.

Et Bloom de conclure: «... selon le lieu de naissance, un étudiant doit donc consacrer un an et demi à deux ans pour atteindre un niveau qu'en d'autres lieux, il aurait acquis en un an[1].»

R. Gjorgjevski[2] a montré par une expérience simple comment, par relativité, des professeurs évaluent différemment une même performance.

Cinq correcteurs, professeurs de la même branche, ont noté indépendamment cent épreuves écrites provenant d'une école secondaire. On a ensuite extrait quinze copies qui toutes avaient obtenu la note «bien». Elles ont été confiées, pour nouvelle correction, à quatre autres professeurs. Ceux-ci ont spontanément adopté des exigences nouvelles: seulement dans la moitié des cas environ, la note est restée «bien». Voici la répartition nouvelle[3]:



Poursuivant son expérience, Gjorgjevski a extrait, de la même série de cent copies, un groupe de douze jugées très bonnes, et un groupe de douze jugées médiocres.

Dans chaque groupe, il a glissé trois copies jugées bonnes. Deux fois cinq professeurs ont évalué. La moyenne des «bonnes» copies glissées dans les «très bonnes» est descendue de 3 sur 5 à 2,40; dans le groupe opposé, la moyenne est passée de 3 à 3,87.

On doit, par ailleurs, à M. Reuchlin[4] une étude portant sur 4.860 élèves répartis dans 397 écoles primaires de France. Les instituteurs ont été invités à classer leurs élèves en quatre catégories: très bons, bons, moyens, médiocres.

1 B.S. BLOOM, *Time and Learning*, Communication au 81e Congrès de l'American Psychological Association, 1973.
2 Voir N. ROT et Z. BUTAS, Les distributions des notes scolaires comparées aux distributions des résultats obtenus aux tests de connaissances, in *Le travail humain*, XXII 1-2, 1959.
3 Les pourcentages fournis par l'auteur sont calculés sur un nombre trop peu élevé de copies pour avoir une signification statistique.
4 Voir *Le Travail humain*, XXII, 1-2, 1959, pp. 12 sqq.

On a ensuite fait passer un même test de français et de calcul à tous les élèves. Le dépouillement des résultats a montré que le même degré de connaissances était le fait d'élèves qui, selon l'école, étaient jugés très bons, bons, moyens ou médiocres (voir p. 33).

Comme le souligne pertinemment M. Reuchlin: «On peut dire (...) que l'instituteur, certainement, connaît mieux que personne les points du programme qui sont acquis ou non par chacun de ses élèves. Ce qu'il ignore, c'est la gravité qui s'attache à chaque faiblesse, à chaque lacune, lorsqu'on la considère non plus au sein d'une classe qui peut être «forte» ou «faible», mais par rapport à l'ensemble du pays. De là, les divergences d'appréciation mises en lumière par l'enquête.»

Voici encore deux exemples donnés par F. Bacher[1]:

«Dans une enquête française portant sur un échantillon représentatif d'élèves de troisième (fin du premier cycle secondaire), on a fait passer une épreuve de connaissances littéraires et une épreuve de connaissances mathématiques aux élèves de 406 classes. On a constaté que les moyennes de classe s'étalaient de 23 à 60 en français (pour une épreuve notée sur 80) et de 7 à 38 en mathématiques (pour une épreuve notée sur 44) (Reuchlin, Bacher 1968). (...) Aux Etats-Unis, Flanagan (1964) signale que des élèves classés dans le quart inférieur de leur école seraient classés dans le quart supérieur s'ils fréquentaient d'autres écoles de la même région.»

1 F. BACHER, La normalisation de la notation, in *Docimologie et Education*, numéro spécial de la revue *Les Sciences de l'Education*, 2-3, 1969, pp. 51-52.

Synthèse des résultats de M. Reuchlin (1959)

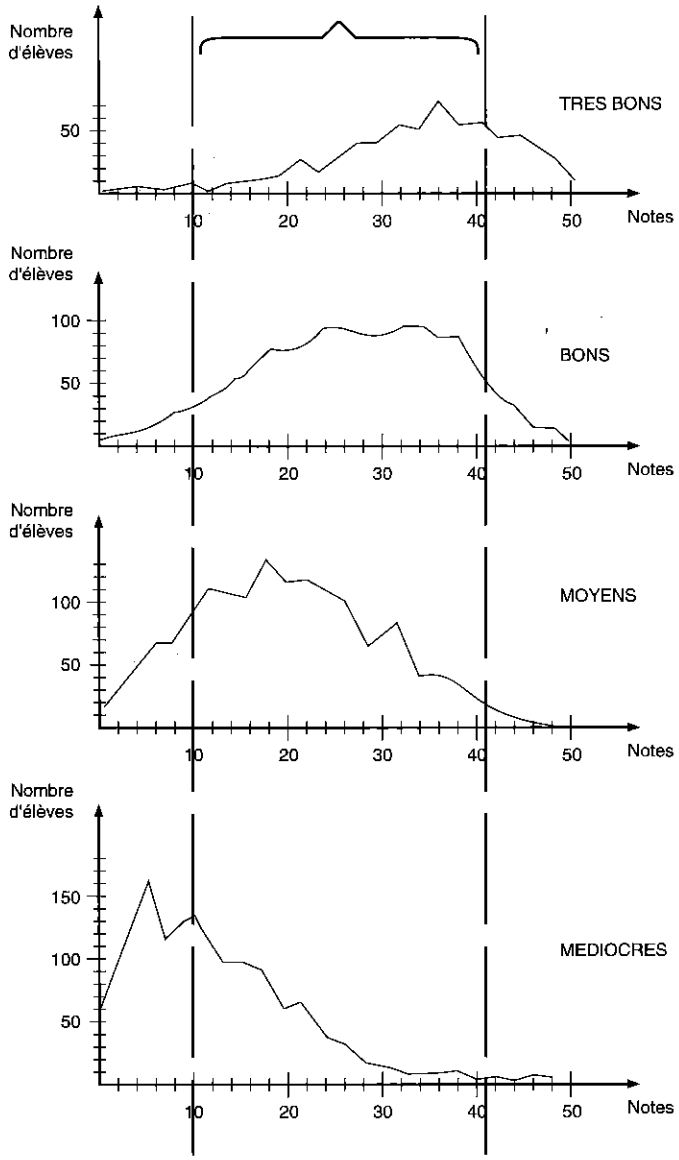


Fig. 1 - Les quatre graphiques permettent de comparer la distribution des notes obtenues, dans l'épreuve de calcul, par 654 élèves jugés «très bons» par leurs maîtres, par 1.303 élèves «bons», 1.551 «moyens» et 1.300 élèves jugés «médiocres». On voit que les distributions de notes se recouvrent largement: dans la zone de notes qui va de 10 à 40, le même niveau de connaissances peut être qualifié, selon le cas, de très bon, moyen ou médiocre.

4. L'échec, générateur d'échecs.

On peut objecter que les études supérieures ou la vie se chargeront de rétablir la justice que les notes scolaires n'auraient pas respectée. Les choses ne sont malheureusement pas aussi simples. D'abord parce que les élèves écartés ou découragés injustement de l'enseignement secondaire général, par exemple, n'auront peut-être plus l'occasion de défendre leur chance à l'université. Ensuite, parce que les favorisés risquent de gagner sur tous les tableaux.

Les recherches actuelles confirment un vieil adage pédagogique: le succès engendre le succès et l'échec d'aujourd'hui prépare un échec demain. Portés par la réussite, certains élèves valorisent un capital intellectuel limité. Le passage de classe «à bon marché» a parfois la double conséquence salutaire de ne pas pousser les parents à retirer leur enfant de l'école où il se trouve et de permettre à l'étudiant de retrouver une sorte de nouveau souffle. Combien de fois les maîtres du secondaire ne voient-ils pas le rendement scolaire d'un élève remonter spectaculairement dès que la crise physique et psychologique de l'adolescence perd de son acuité?

La pratique du redoublement de classe varie d'ailleurs considérablement de pays à pays. En 1991, la Communauté française de Belgique avait le triste privilège d'être le champion des échecs scolaires au sein de la Communauté européenne, et cela dès le début de la scolarité primaire. Or il n'a jamais été démontré que le redoublement de classe est plus bénéfique pour les apprentissages scolaires que le passage automatique. C'est pourquoi certains pays ne font que très peu redoubler. C'est, par exemple, le cas de la Suède où aucun redoublement n'est autorisé avant l'âge de 15 ans. Or les résultats scolaires y sont plutôt meilleurs qu'en Belgique.

R. Rosenthal et L. Jacobson[1] ont publié une étude d'ensemble sur l'aspect de l'*effet œdipien de la prédiction*, ainsi qualifié parce que si, à la naissance d'Œdipe, l'oracle n'avait pas prédit qu'il tuerait son père, il n'aurait pas été éloigné de sa famille. Connaissant son père, il ne l'aurait pas tué. En d'autres mots, la tragédie se produisit *parce qu'elle avait été prédite.*

Il semble que, dans une mesure certes difficile à apprécier, un élève se comporte en fonction du jugement que le professeur porte sur lui. Même les animaux en dressage n'échappent pas à cette règle.

1 *Pygmalion à l'école*, Paris, Casterman, 1971.

Rosenthal rappelle l'expérience suivante: on constitue deux groupes de souris blanches génétiquement identiques. Au moment où on les remet aux étudiants chargés de les dresser, une remarque indique que le premier groupe est composé d'animaux particulièrement bien doués, alors que le second est de pauvre qualité. Les résultats du dressage confirmeront ce pronostic... fantaisiste.

L'expérience suivante, conduite par Rosenthal lui-même, s'inscrit dans le même contexte.

On a annoncé à des instituteurs des six années primaires de l'Oak School que d'éminents chercheurs venaient de mettre au point un test d'«épanouissement» (en réalité, un simple test d'intelligence[1], peu connu et sans vertu particulière, a été utilisé).

Tous les élèves de l'école y ont été soumis, et on a signalé aux enseignants ceux qui étaient sur le point de s'épanouir intellectuellement, pronostic fantaisiste.

Le quotient intellectuel des élèves ainsi désignés s'est élevé d'une façon significative dans les trois années inférieures; en outre, des progrès supérieurs à la moyenne ont été observés en lecture et en arithmétique[2].

5. Rupture entre enseignement et examen.

L'examen doit être le reflet de l'enseignement donné. Imaginons qu'au lieu de conduire ses élèves à la piscine, un professeur d'éducation physique ait passé son année à disserter sur la natation. Il serait évidemment inadmissible que les notes d'examen soient attribuées d'après les performances natatoires... car seuls réussiraient ceux qui ont appris à nager en dehors du cours.

Pareil agissement paraît impensable. Pourtant, combien de fois n'assiste-t-on pas à une aberration similaire. Le professeur fait un cours où il impose les informations, les opinions et les jugements, puis, à l'examen, il pose des questions dites d'intelligence, entendons qui mettent en œuvre des capacités qui n'ont précisément pas été installées pendant les leçons. Pourquoi les étudiants seraient-ils soudain capables de découvrir seuls, à l'examen, la solution d'applications géométriques originales, s'ils n'ont pas été formés en ce sens pendant l'année?

1 Mesurée à l'aide du Toga de Flanagan (1960).

2 Les chiffres avancés par Rosenthal et Jacobsen sont contestés par R.L. Thorndike. Par contre, la tendance à l'augmentation des QI et des scores semble reconnue par tous les spécialistes. C'est évidemment ce qui importe. Pour plus de détails, voir *American Educational Research Journal*, 5, 4, 1968, 708.

Ainsi s'expliquent bien des pourcentages d'échecs anormaux. Une enquête suscitée à cause de nombreuses notes insuffisantes chez un professeur de chimie (4 élèves sur 22 avaient obtenu plus de la moitié des points à l'examen de fin d'année) fit rapidement apparaître que le professeur ne faisait que deux ou trois interrogations écrites par an. Sur les trois questions d'examen, deux impliquaient interprétation et transfert, démarches qui, au niveau de difficulté où le maître se plaçait, nécessitaient une compréhension profonde des phénomènes et un entraînement, long et finement contrôlé, à la solution de problèmes originaux.

Une discussion amicale avec le jeune professeur en question révéla qu'il se croyait obligé de traiter en détail de tous les points du programme, qu'il n'avait pas conscience de la nécessité de l'entraînement systématique et au moins semi-individualisé à la solution de problèmes originaux et que, de toute façon, il ne possédait pas les notions psychopédagogiques de base nécessaires à la conduite d'un tel entraînement.

6. **Désaccord entre correcteurs.**

Tous les docimologistes citent des exemples de grandes divergences entre maîtres appelés à évaluer un même travail d'élève ou un même ensemble de travaux, ce qui permet pourtant une relativité plus sûre. On sait combien le niveau d'exigence varie selon les examinateurs. Les correcteurs ont-ils au moins une conscience objective de leur degré de sévérité ou d'indulgence ? Loin s'en faut. R. Duquenne[1] a réparti des correcteurs en deux groupes : ceux qui s'affirmaient sévères et ceux qui se disaient indulgents. Quatre devoirs ont été corrigés par tous. La moyenne des correcteurs se croyant sévères est de 12,4 ; pour les « indulgents », elle est de 11,6. Les fluctuations des notes d'un même professeur jugeant un même travail à quelque intervalle de temps (fidélité) peuvent aussi être importantes.

Comme tous les ouvrages de docimologie fourmillent d'exemples, nous n'en proposons que quelques-uns, pour rappel.

a) *Composition française.*

Elle est une accusée de prédilection.

- Piéron rapporte qu'une même composition, jugée par 76 correcteurs, tous professeurs de langue maternelle, conduisit aux résultats suivants[2] :

1 R. DUQUENNE, In *L'Education Nationale*, 1967, 840, 19-20.
2 PIERON, *o.c.*, p. 123.

Note sur 20	*Nombre de correcteurs qui l'ont attribuée*
de 0 à 1	1
2 - 3	6
4 - 5	20
6 - 7	34
8 - 9	10
10 - 11	3
12 - 13	2

- Le Service de la Recherche pédagogique du C.R.D.P. de Lyon a invité, en 1967[1], 150 professeurs de français de l'Académie de Lyon à corriger trois compositions françaises traitant d'un même sujet.

Max. 20	*Moyenne*	*Marge de variation*
Devoir I	$10\frac{1}{4}$	$4\frac{1}{2}$ - $13\frac{3}{4}$
Devoir II	6, 6	$2\frac{1}{2}$ - $12\frac{1}{2}$
Devoir III	11,6	$5\frac{1}{2}$ - $17\frac{1}{2}$

- L'intervention d'hommes cultivés, étrangers aux routines de l'enseignement, améliorerait-elle la situation ?

J. French, P. Diderich et S. Carlton[2] ont comparé la façon de corriger de dix professeurs de langue maternelle, de neuf professeurs de sciences naturelles, de dix écrivains ou rédacteurs de journaux, de neuf juristes et de sept directeurs d'entreprises commerciales.

Afin d'obtenir une très large distribution des résultats, plusieurs centaines d'étudiants de première année universitaire ont été invités à traiter un des deux thèmes suivants : « Qui doit fréquenter l'université ? » ou « A partir de quel moment l'adolescent doit-il être traité en adulte ? ».

1 C.R.D.P., *Docimologie et Examens*, Lyon, I.P.N., 1969.
2 J.W. FRENCH, *School of thought in judging excellence of English themes*, Princeton, Educational Testing Service, 1961.

On a extrait cent cinquante travaux pour chacun des deux thèmes. La loi des grands nombres jouant, chaque échantillon devait normalement inclure des copies très bonnes et de très médiocres, ce qui rendait la discrimination aussi facile que possible.

Les trois cents travaux ont été évalués par tous les correcteurs; ceux-ci travaillaient indépendamment à domicile et avaient simplement reçu les instructions suivantes:

1° Fiez-vous à votre propre jugement pour définir ce que vous pensez être la «faculté d'écrire».

2° Triez les travaux en neuf tas, par ordre de mérite.

3° Veillez à ce qu'au moins six des 150 rédactions, traitant de chaque sujet, figurent dans chaque tas.

4° Indiquez par un commentaire sur les travaux pourquoi vous les aimez ou ne les aimez pas.

On a calculé la corrélation entre toutes les paires possibles de correcteurs. La corrélation moyenne est de 0,31.

Sur 300 travaux:

- 101 se sont vu attribuer les neuf classements possibles;
- aucun travail n'a reçu moins de cinq classements différents.

b) *Mathématiques.*

Le C.R.D.P. de Lyon[1] a aussi demandé à 150 professeurs de mathématiques exerçant dans une classe de 3e de noter les copies de 3 élèves ayant à résoudre le même problème.

Max. 20	*Moyenne*	*Marge de variation*
Devoir I	5,70	0,5 - 11,5
Devoir II	16	11,5 - 20
Devoir III	8	3,5 - 11,5

Si, pour l'élève I par exemple, on élimine les deux notes extrêmes, il reste autant de correcteurs pour noter 2/20 que 8/20, alors qu'on pouvait attendre une notation objective.

1 *O.c.*

c) *Médecine.*

Au niveau universitaire, les résultats ne sont pas plus sûrs.

Un même candidat de 2e médecine a subi, en juin et en septembre 1965, une épreuve écrite comportant cinq questions simples et précises à traiter en 1 h 30[1].

Les réponses (anonymes) ont été corrigées indépendamment par cinq correcteurs possédant des titres et des qualités identiques.

Résultat total sur 100	*Marge de variation*	*Décision*	
		Admis	*Refusé*
Copie de juin	47 - 78	3	2
Copie de septembre	32,5 - 73	3	2

Si on considère les notes attribuées à chacune des 10 questions, on constate:

1) Ecart maximum:	12	(de 3 à 15 sur 20)
2) Ecart pour 4 questions:	11	
3) Plus petit écart maximum entre cinq correcteurs:	7,5	

d) *Divers.*

A. Agazzi[2] rappelle un exemple, plus frappant encore, concernant un ensemble de branches.

Six correcteurs ont, chaque fois, noté les examens du baccalauréat (note d'échec: moins de 10 sur 20):

	Refusés par les six correcteurs	Admis par les six correcteurs	Admis par les uns et refusés par les autres
Version latine	40%	10%	50%
Composition française	21%	9%	70%
Anglais	37%	16%	47%
Mathématique	44%	20%	36%
Philosophie	9%	10%	81%
Physique	37%	13%	50%

1 Voir *Le Monde* du 6 septembre 1966. Remarquons que les trois admis furent les mêmes dans les deux cas.

2 A. AGAZZI, *Les aspects pédagogiques des examens*, Strasbourg, Conseil de l'Europe, 1967, p. 119. La recherche est due à H. Laugier et D. Weinberg et portait sur le baccalauréat français.

e) *Aux interrogations orales, plus de discordances encore.*

Tous les docimologistes s'accordent pour reconnaître encore plus de discordances lors d'interrogations orales que pour les épreuves écrites. Rares sont cependant les expériences rigoureusement contrôlées en la matière. L'une des plus récentes a été réalisée par H. Piéron, M. Reuchlin et F. Bacher[1].

L'examen oral est, jusqu'à ces derniers temps, le moins systématiquement étudié, en raison de la difficulté d'enregistrer - sans perturber gravement la situation -, non seulement ce qui se dit, mais aussi les comportements non verbaux si importants dans la relation examinateur-examiné[2]. L'enregistrement télévisuel télécommandé offre aujourd'hui des possibilités qui ne manqueront pas d'être exploitées.

On doit à K. Ingenkamp[3] une bonne vue d'ensemble sur la question. Il rappelle d'abord une observation faite dans les années trente par Hartog et Rhodes[4]. Seize candidats à un poste dans l'administration anglaise ont été interrogés par deux jurys qui avaient reçu les mêmes directives. La corrélation entre les notes attribuées par les deux groupes n'est que de 0,41. Le candidat classé premier par un jury ne venait qu'en treizième place dans l'autre, et le candidat classé premier dans le second jury ne venait qu'en onzième place dans le premier...

La médecine et les disciplines connexes sont les seules branches pour lesquelles l'examen oral a été étudié de façon assez continue depuis les années trente jusqu'à ces derniers temps[5].

Nous ne retiendrons qu'une seule recherche à titre exemplatif: celle de Pokorny et Frazier[6] qui observent une corrélation de 0,57 entre deux équipes faisant passer des examens de psychiatrie aux mêmes candidats. «La corrélation entre la moyenne des notes obtenues à deux examens pratiques (interview de patients, etc.) et la note moyenne aux examens oraux et l'examen écrit s'élevait à 0,73, mais n'était que de 0,21 avec l'appréciation globale du maître de stage clinique.»

1 H. PIERON, M. REUCHLIN et F. BACHER, Une recherche expérimentale de docimologie sur les examens oraux de physique à une session du baccalauréat, in *Biotypologie*, 1962, 23, 48-73.

2 Voir à ce propos: G. DE LANDSHEERE et A. DELCHAMBRE, *Les comportements non verbaux de l'enseignant*, Paris, Nathan; Bruxelles, Labor, 1980.

3 K. INGENKAMP, *Pädagogische Diagnostik*, Weinheim, Beltz, 1975, pp. 36 sqq.

4 P. HARTOG et E.C. RHODES, *An Examination of Examinations*, Londres, McMillan, 1936, 2e éd.

5 Ingenkamp cite notamment T. COLTON et O. PETERSON, An essay on medical students abilities by oral examination, in *Journal of Medical Education*, 42, 1967, 1005-1014.
D. WAUGH et C. MOYSE, Medical education. Oral examination. A videotape study, In *Canadian Medical Assoc. Journal*, 100, 1969, 635-640.
P. HALLOWAY, The validity of essay and viva-voce examining, in *British Dental Journal*, 123, 1967, 227-232.

6 A. POKORNY et S. FRAZIER, An evaluation of oral examination, in *Journal of Medical Education*, 41, 1966, 28-40.

Reuchlin a enregistré sur bande vingt examens oraux en physique (2e partie du baccalauréat) et les a fait évaluer par seize professeurs de lycées expérimentés: la moyenne des notes varie de 8 à 13,4 sur 20[1].

De l'ensemble des données - assez limité, on l'a dit - dont on dispose actuellement, il se dégage que, le plus souvent, la corrélation entre les notes attribuées par différents juges à un examen oral est inférieure à la corrélation observée pour les épreuves écrites. Elle se situe le plus souvent en dessous de 0,50.

Toutefois, l'accord est nettement plus élevé quand les différents juges se réfèrent à un même syllabus ou manuel de cours et s'accordent sur des critères de contenus et de comportements définis de façon aussi opérationnelle que possible.

Ingenkamp conclut: «En Europe, les examens oraux jouent un rôle important dans les écoles et les universités. Pourtant, leur valeur est extrêmement douteuse. Et la recherche scientifique à ce propos est, en gros, très lacunaire. Nous ne savons que très peu de choses sur les interactions qui se produisent pendant les examens, sur la différence entre des jugements portés par des animateurs qui connaissent bien les candidats et des juges qui ne les connaissent pas, sur l'influence de la facilité verbale des candidats, sur les réactions de sympathie et d'antipathie, sur le degré de difficulté des questions posées à différents candidats, etc.»[2].

f) *Combien de correcteurs pour stabiliser la note?*

Calculer la moyenne entre deux correcteurs apporterait-il une amélioration considérable? Non, si aucune mesure stricte n'a été prise pour accorder les notateurs. Par combien de professeurs différents faudrait-il faire évaluer le même travail pour que la note se stabilise? Les nombres suivants, avancés par Piéron[1], n'ont certes pas de valeur absolue. Ils donnent néanmoins une idée de l'ordre de grandeur des nombres...

Composition française	78
Version latine	19
Anglais	28
Mathématiques	13
Dissertation philosophique	127
Physique	16

1 H. PIERON, M. REUCHLIN et F. BACHER, Une recherche expérimentale de docimologie sur les examens oraux de physique au niveau du baccalauréat de mathématique, in *Biotypologie*, mars-juin 1962, p. 5189.

2 *o.c.*, p. 41. A ces phénomènes, il faut aussi ajouter celui de la fluctuation que nous analysons p. 53.

L'histoire et la géographie ne figurent pas dans la liste; ces deux branches aussi donnent pourtant lieu à de grandes fluctuations.

Une remarque s'impose enfin: stabiliser la note, dans des circonstances d'examen particulières, n'implique en rien qu'on a ainsi obtenu *la vraie note*. Il va de soi que ni un individu accomplissant une tâche, ni la tâche elle-même ne portent en eux une note qui représenterait leur valeur essentielle, absolue. «Cette notion, ce mythe de la vraie note que mériterait un devoir, apparaît en effet, à l'analyse, fallacieuse sur le plan théorique, abusive sur le plan méthodologique et absurde sur le plan pratique; mais bien commode évidemment, et idéologiquement orientée pour perpétuer l'idée dénuée de tout fondement selon laquelle il existerait quelque part dans l'harmonie préétablie une note qui serait la juste mesure d'un devoir; ce qui impliquerait que cette note serait celle que mérite celui qui la reçoit»[1].

Noizet et Bonniol[2] doutent même qu'une stabilisation de note soit possible. Selon eux, il ne serait nullement établi que les causes de fluctuation sont aléatoirement distribuées parmi les correcteurs et donc se compensent. Laugier et Weinberg, suivis par H. Piéron, admettent cependant cette hypothèse forte; nous partageons leur avis.

On pourrait multiplier les exemples de désaccord entre notateurs. Mais à quoi bon se complaire dans l'accusation facile, comme trop de docimologistes l'ont fait? L'important est d'être conscient du danger et, surtout, de chercher les remèdes.

Quelle est l'explication des divergences observées?

Elle réside principalement dans la multiplicité des points de vue, des dimensions, selon lesquels une même épreuve peut être jugée, et à l'absence de directives précises (parfois, il n'y en a pas du tout) données pour la notation. F. Bacher[3] écrit:

«Dans une dissertation, on peut noter l'organisation des idées, leur originalité, la correction de l'expression, l'élégance du style, etc. Si l'on considère chaque aspect possible comme une dimension de la notation, le devoir d'un élève peut être situé dans un espace à *n* dimensions; la difficulté vient de ce que, dans cet espace, seul un ordre partiel s'établit entre les élèves: un élève occupant une position élevée sur toutes les dimensions peut être déclaré supérieur à un élève occupant une position moins élevée qui a une position élevée sur une première

1 J.-J. BONNIOL, *Les comportements d'estimation dans une tâche d'évaluation d'épreuves scolaires*, Aix-en-Provence, Université de Provence, 1972 (thèse de 3e cycle).
2 G. NOIZET et J.-J. BONNIOL, *o.c.*, p. 784.
3 F. BACHER, *La normalisation des notes*, o.c., pp. 53-54.

dimension et une position faible sur une deuxième dimension s'il est supérieur ou inférieur à un élève occupant des positions inverses.»

En outre, même s'ils considèrent le même aspect, les notateurs varient en sévérité, en stabilité de jugement, en résistance à l'effet de halo, etc.

Il ne suffit pas de constater des divergences entre correcteurs d'une même épreuve; il importe, en outre, de déterminer si on peut les considérer comme des «accidents» survenant au hasard, ou s'il s'agit, au contraire, de fluctuations systématiques, c'est-à-dire relevant de causes précises et donc identifiables.

Comme le dégagent clairement Noizet et Caverni[1], les divergences systématiques dépendent à la fois de variables de situation (selon les conditions dans lesquelles ils se trouvent, les évaluateurs ne traitent pas les mêmes informations) et de variables de personnalité (influence du style cognitif des évaluateurs)[2].

Fondamentalement, la démarche cognitive de l'évaluateur se ramène à extraire de l'objet à évaluer un certain nombre d'indices correspondant à des critères issus d'un «modèle de référence». Celui-ci est la résultante de ce que Noizet et Caverni appellent le «produit norme» (exemple: une «bonne rédaction» est un texte logique, original, écrit dans une langue correcte et sans fautes d'orthographe), le «produit attendu» (exemple: pour un même thème de rédaction, on n'attend pas le même produit à l'école élémentaire et au lycée) et l'utilisation faite d'une échelle de mesure.

On sait que, selon le degré de pratique de l'évaluateur et les événements qui interviennent en cours d'évaluation, l'usage fait du modèle de référence peut varier, parfois de façon considérable.

Nous étudierons, sous des rubriques particulières, des phénomènes déterminant le «produit attendu», notamment l'effet de halo, l'effet de stéréotypie, l'effet d'ordre. Nous nous arrêterons, plus spécialement ici, à la notion de «produit norme». Celui-ci est évidemment déterminé par les critères auxquels l'évaluateur accorde une importance particulière[3] ... ou croit la leur accorder.

1 G. NOIZET et J.-P. CAVERNI, *o.c.*, pp. 65 sqq.
2 Le style cognitif est la résultante entre le fonctionnement intellectuel et la personnalité. Exemples: impulsif-réfléchi; approche globale-approche analytique, rigide-flexible, etc. Sur le style cognitif, voir G. DE LANDSHEERE, *Dictionnaire de l'évaluation et de la recherche en éducation*, Paris, P.U.F., 1992, 2e éd.
3 Nous analysons ici la situation d'un évaluateur travaillant en toute indépendance. Sa liberté serait évidemment moins grande s'il devait respecter un barème de correction arrêté de commun accord avec des collègues.

L'étude expérimentale des critères (ou plus exactement, des indices de critères) utilisés par les évaluateurs revêt une importance considérable. Noizet et Caverni proposent à cet égard une méthodologie d'un grand intérêt, *la technique de construction de copies*[1].

Par exemple, pour la rédaction, ils construisent, à partir de phrases authentiques d'élèves, des copies susceptibles d'être décrites selon des modalités précises. Pour une modalité donnée, par exemple, la syntaxe, on soumet des phrases d'élèves à des juges expérimentés en leur demandant de les classer en correctes ou incorrectes, et l'on n'utilise par la suite que les phrases sur le classement desquelles les juges sont unanimes. Arrivé à l'expérience proprement dite, on pourra alors constituer des copies comprenant un taux d'erreurs variable.

Par ailleurs, *pour découvrir les critères utilisés* par les évaluateurs, trois méthodes sont proposées.

1° La correction par relais.

L'expérimentateur possède ou prétend posséder une copie, mais ne la montre pas. On demande à l'évaluateur de poser des questions qui lui permettent d'apprécier le travail sans le voir et l'expérimentateur répond: «Les questions posées par l'enseignant, l'ordre dans lequel il les pose, fournissent des indications sur les critères qu'il prétend utiliser»[1].

Les questions posées peuvent être catégorisées:

- selon l'ordre d'apparition (on peut penser que le premier critère cité est le plus important aux yeux de celui qui l'énonce);
- selon qu'elles permettent ou non une réponse objective, par comptage (nombre de mots, de fautes d'orthographe, etc.);
- selon qu'elles appellent ou non une réponse subjective (élégance du style).

2° Correction par relais à partir d'une liste de questions.

Même méthode que précédemment, mais cette fois l'évaluateur ne peut utiliser que les questions figurant dans une liste, d'ailleurs constituée en prolongement d'expériences du type a.

1 G. NOIZET et J.-P. CAVERNI, *o.c.*, pp. 120 sqq.

3° Recherches sur la notation de devoirs construits.

La méthode consiste à faire varier une même copie de base selon un certain nombre de modalités, par exemple:

- incorrection de style et fautes d'orthographe;
- incorrections de style sans fautes d'orthographe;
- sans incorrections de style et sans fautes d'orthographe;
- sans incorrections de style et avec fautes d'orthographe.

Si la moyenne des notes est significativement inférieure chaque fois qu'il y a incorrection de style, c'est que ce critère joue effectivement dans la notation[1].

Les critères étant identifiés, il reste à déterminer quelle est leur importance relative dans la notation et aussi dans quelle mesure ils sont indépendants ou non. On trouvera dans Noizet et Caverni (pp. 126 sqq.) des suggestions expérimentales à ce propos.

7. **Infidélité d'un même correcteur.**

Un même correcteur est-il au moins égal à lui-même? Non. A des phénomènes aussi évidents que les variations de santé physique et mentale et l'évolution du savoir s'ajoutent un grand nombre de facteurs plus ou moins bien définis: variation dans la qualité de la relation établie avec l'élève, dans le contexte de l'évaluation (si l'on vient de corriger un travail excellent, le suivant peut être sous-évalué, ...), dans l'échelle consciemment ou inconsciemment adoptée, etc.

Quatorze historiens ont été invités à noter une deuxième fois quinze compositions, douze à dix-neuf mois après les avoir notées une première fois. Toute trace de correction avait été effacée. Les professeurs accordaient non seulement des points, mais indiquaient la réussite globale ou l'échec.

Dans 92 cas sur 210, le verdict a été différent d'une fois à l'autre[2].

Il faut toutefois insister sur le fait que des résultats aussi pauvres sont sans doute en partie dus au manque de directives rigoureuses précisant les aspects à considérer par les notateurs.

1 G. NOIZET et J.-P. CAVERNI, *o.c.*, pp. 122-124.
2 HARTOG et RHODES, *An Examination of Examinations*, London, McMillan, 1935, p. 81 et p. 15.

Un schéma pour continuer la recherche.

Il est utile que, de temps en temps, des recherches soient faites sur la fidélité d'un notateur par rapport à lui-même ou sur l'accord entre notateurs, ne fût-ce que pour ranimer la conscience d'un danger menaçant. C'est pourquoi nous croyons opportun de signaler l'excellent plan de recherche dû à F. Yates et D. Pidgeon.

- Sept groupes de 50 enfants achevant l'école primaire ont été constitués. Ils ont été soumis à deux épreuves de langue maternelle:
  a) rédaction sur un sujet choisi parmi trois ou quatre;
  b) questions de compréhension de textes et questions de grammaire (usage de la langue).
- Sept examinateurs expérimentés - cinq hommes et deux femmes - ont noté les travaux.
- Chaque groupe d'enfants a d'abord subi trois examens différents, à une semaine d'intervalle; ensuite, la première des trois épreuves a été recommencée deux fois à une semaine d'intervalle.

SCHEMA[1]

GROUPE	Séances d'examen				
	I	II	III	IV	V
1	Aae	Bac	Dab	Aac	Aae
2	Bbf	Cbd	Ebc	Bbf	Bbf
3	Ccg	Dce	Fcd	Ccg	Ccg
4	Dda	Edf	Gde	Dda	Dda
5	Eeb	Feg	Aef	Eeb	Eeb
6	Ffc	Gfa	Bfg	Ffc	Ffc
7	Ggd	Agb	Cga	Ggd	Ggd

Sept séries de questions: A, B, C, D, E, F, G.

Sept examinateurs: a, b, c, d, e, f, g.

Nous ne reprenons pas le détail des résultats, mais avons simplement voulu montrer un exemple de contrôle bien mené, avec, notamment, une mesure de l'apprentissage.

1 Il est dû à D. FINNEY, cité par Yates et Pidgeon, *Admission to Grammar Schools*, Londres, N.F.E.R., 1957, p. 99.

8. **Stéréotypes et effets de halo.**

Dans la situation scolaire habituelle, le maître connaît chacun de ses élèves et peut donc doser, nuancer ses notes en fonction d'un effet souhaité: ici, on encourage en surestimant le travail; là, on fait preuve d'une sévérité exceptionnelle pour donner un choc que l'on espère salutaire. Dans ces cas, le maître agit délibérément, en toute conscience.

Il en va tout autrement dans les phénomènes de contagion des évaluations, de stéréotypie et de halo.

a) *Contagion des évaluations. Stéréotypie.*

La connaissance des résultats antérieurs d'un élève - même inconnu - tend à influencer l'évaluateur. On assiste à une sorte d'imitation par contagion.

Caverni, Fabre et Noizet[1] ont invité une population de professeurs de sciences de l'enseignement secondaire à noter chacun quatre copies, constituées par des réponses à une question. Avec chaque copie, on indiquait une série de cinq notes censées avoir été obtenues par l'auteur de la copie lors de devoirs effectués précédemment pendant l'année scolaire. C'est à ces séries de notes antérieures que des variations systématiques ont été apportées.

Les séries de notes se distinguaient selon deux descripteurs croisés: leur moyenne, forte (13/20) ou faible (7/20), et leur dispersion, forte (10 points d'écart entre les notes extrêmes) ou faible (2 points d'écart entre les notes extrêmes). La moyenne exprimait le niveau moyen de l'élève, tandis que la dispersion exprimait la régularité ou l'irrégularité de ses performances[2].

La variation systématique introduite dans cette expérience concerne *l'information donnée à l'évaluateur sur les notes* obtenues précédemment par l'élève. Les mêmes copies (cf. tableau), associées à une série de moyenne élevée et de dispersion faible, ont été notées en moyenne 9.69 sur 20, alors qu'elles n'ont été notées en moyenne que 7,69 sur 20 lorsqu'elles ont été associées à une série dont la moyenne était faible et la dispersion forte. Cet écart est statistiquement significatif. Il apparaît de plus que l'importance de l'effet dépend de la qualité reconnue à la copie: il est d'autant plus fort que la copie est jugée en moyenne de plus haut niveau.

1 J.-P. CAVERNI, J.-M. FABRE, G. NOIZET, Dépendance des évaluations scolaires par rapport à des évaluations antérieures, in *Le Travail humain*, 1975, 38, 213-222.

2 Un autre descripteur aurait pu être utilisé: la succession des notes peut marquer un progrès ou, au contraire, une régression. Dans l'expérience citée, cette source possible de variation a été maintenue constante: toutes les séries marquaient un progrès (la dernière note était toujours la plus élevée, mais la progression n'était pas monotone de la première à la dernière).

TABLEAU

Moyenne des notes (sur 20) obtenues
par les mêmes copies selon la série des notes antérieures
à laquelle elles étaient associées

	Série de notes antérieures	
	Moyenne forte et dispersion faible	Moyenne faible et dispersion forte
Copie a	12,00	9,75
Copie b	8,50	6.50
Copie c	15,25	11,75
Copie d	3,00	2,75

b) *Stéréotypie*

Par *stéréotypie*, on entend une immuabilité plus ou moins accusée qui s'installe dans le jugement porté sur l'élève.

La stéréotypie résulte d'une **contagion des résultats**. Un premier travail médiocre incline à penser que le second le sera aussi; si cela se vérifie, la tendance à accorder une note médiocre au troisième travail s'accroît encore, et ainsi de suite. Chez le professeur surchargé de corrections, la déformation se produit d'autant plus facilement.

On aurait tort de croire que la stéréotypie influence uniquement les évaluations à base subjective accusée (dissertation, composition d'histoire et, en général, réponses impliquant un jugement de valeur). Elle atteint des exercices aussi «objectifs» que la dictée orthographique. L'expérience suivante en témoigne.

Un professeur de langue maternelle fait régulièrement des dictées. Bientôt, il connaît les élèves qui réussissent habituellement le mieux et le moins bien cet exercice. Si l'on détermine la fréquence des fautes «oubliées», non perçues par le correcteur, on constate que les oublis en faveur des bons élèves sont significativement plus élevés que pour les élèves faibles. Dans le premier cas, le maître s'attend à ne pas rencontrer d'erreurs; dans le second, il les guette[1].

Si un manque de conscience professionnelle intervient, la situation peut devenir très grave. Un cas de stéréotypie accusée dont un élève était victime pour les versions latines nous avait été signalé

1 Voir à ce propos M. ZILLIG, Beliebte und unbeliebte Volksschülerinnen, in *Arch. f. d. ges. Psychologie*, 12, 1934, p. 32, cité par E. HOHN, *Der schlechte Schüler*, Munich, R. Piper, 1967, p. 105.

(enseignement secondaire). A titre de vérification, nous avons fait faire les devoirs ultérieurs successivement par un autre élève de la classe, par un élève de même niveau pédagogique, classé premier dans une autre école, puis par un licencié en philologie classique: la note n'a pas varié d'un demi-point sur vingt...

Nous avons répété ce genre d'expériences à d'autres occasions. Combien de parents, aussi, parfois très compétents dans le domaine où ils faisaient occasionnellement le travail de leur enfant, ont éprouvé quelque déception ou étonnement quand ils ont eu connaissance de «leur» note.

En déduire que tous les maîtres incriminés de telle façon manquent du sens des responsabilités serait pourtant injuste. C'est plus les méthodes d'évaluation que les hommes qu'il faut mettre en cause.

c) *Effet de halo.*

L'effet de halo présente un caractère affectif accusé. Souvent, on surestime les réponses d'un élève de belle allure, au regard franc, à la diction agréable. Il ne faut cependant pas généraliser. On connaît des professeurs qui, par anticonformisme, favorisent un certain débraillé ou des originalités qui ne sont pas toujours du meilleur goût.

Soit pour des raisons de lisibilité, soit pour des raisons nettement affectives, l'écriture peut aussi influencer le correcteur. Les spécialistes de la publicité savent depuis longtemps que la présentation du message exerce une influence considérable sur son rendement.

C. Chase[1] a étudié l'influence de la qualité de l'écriture sur les notes attribuées aux rédactions: elle est nettement significative.

On constate que, même dans des tests objectifs exigeant des indications manuscrites de la part de l'étudiant et pour la correction desquelles on utilise des grilles standardisées, la mauvaise qualité de l'écriture fait baisser le score.

Les expériences suivantes, dues à R. Weiss, montrent avec quelle facilité une combinaison de stéréotypie et d'effet de halo peut être artificiellement provoquée.

1 C. CHASE, The Impact of some obvious variables on essay test scores, in *Journal of Educational Mesurement, 1968, 5, 315-318.*

R. Weiss[1] a sélectionné deux rédactions faites par des élèves de 4e primaire. Les travaux ont été dactylographiés sans modification aucune, puis soumis pour correction à deux groupes de 46 instituteurs enseignant aussi en 4e primaire.

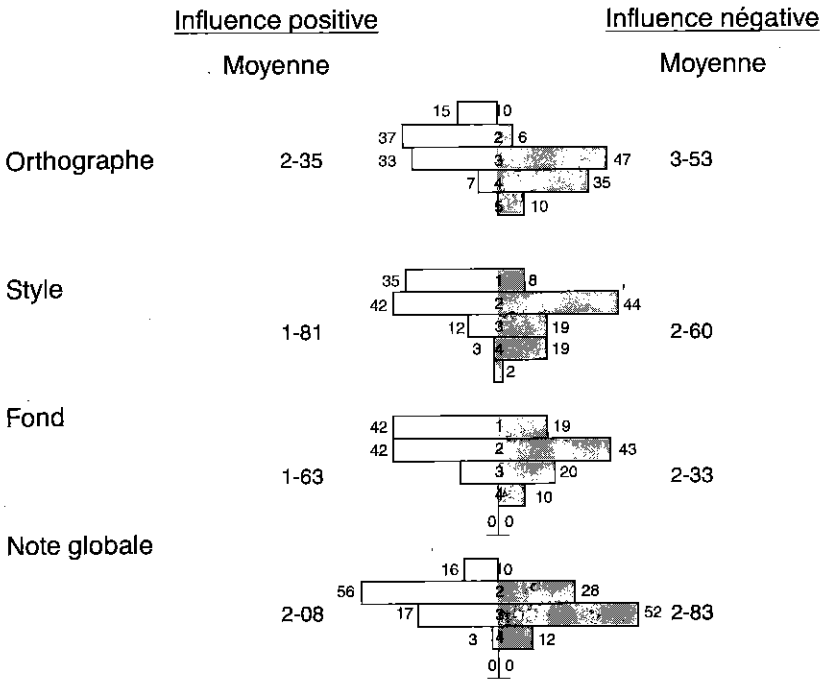
Le commentaire suivant accompagnait la distribution des travaux au premier groupe : « Voici deux compositions écrites par des élèves de 4e primaire. Le travail n° 1 est l'œuvre d'un élève moyen qui aime lire les bandes dessinées ; son père et sa mère sont employés. Le travail n° 2 a été fait par un enfant doué ; son père est rédacteur d'un quotidien connu. » Pour le second groupe de maîtres, les rôles ont été inversés.

La correction devait se faire selon une échelle à cinq degrés : très bien (1) ; bien (2) ; moyen (3) ; suffisant (4) et insuffisant (5). L'orthographe, le style, le fond, puis l'ensemble devaient être notés séparément.

Pour les quatre aspects considérés, les notes attribuées au travail pour lequel on a créé un préjugé favorable ont été significativement supérieures aux autres. Pour l'orthographe, qui semblait le plus devoir échapper à l'effet de halo, on observe qu'au travail de l'élève présenté comme doué, 16 % des correcteurs accordent la note *très bien* et aucun la note insuffisant ; si le même élève est présenté comme moyen, les correcteurs n'accordent aucun *très bien*, mais 11 % notent insuffisant.

1 R. WEISS, The Reliability of the Number Marking System : An Austrian Study. In J. LAUWERYS et D. SCALON, *Examinations*, Londres, Evans Brothers, 1969, pp. 101-107.

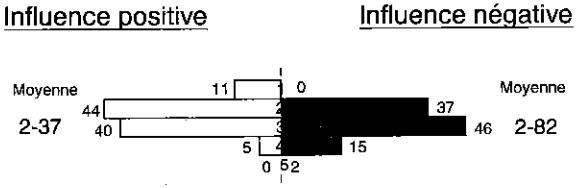
NOTATION DES COMPOSITIONS SOUS L'INFLUENCE D'UN PREJUGE FAVORABLE OU DEFAVORABLE[1]



Dans une expérience similaire sur des problèmes d'arithmétique, également en 4e primaire, R. Weiss observe :

Préjugé favorable	*Préjugé défavorable*
11 % de très bien	Aucun très bien
5 % de suffisant	15 % de suffisant
0 % d'insuffisant	2 % d'insuffisant

PROBLEME D'ARITHMETIQUE



1 Toutes les différences sont statistiquement significatives à P.01 sauf une (P.10).

On voit combien un artifice relativement grossier exerce déjà une influence. Comme le souligne Weiss, la déformation est très probablement bien plus grande encore en situation réelle où le maître connaît l'apparence de l'élève, sa conduite, sa façon de s'exprimer, la profession et le statut social des parents, etc.

9. Effets d'ordre de correction[1].

Les élèves rompus aux examens ont depuis longtemps découvert l'importance des contrastes : passer immédiatement après un candidat brillant se révèle défavorable ; succéder à plus faible que soi peut être avantageux, à condition que la médiocrité des réponses que l'interrogateur vient d'obtenir ne l'ait pas mis de trop méchante humeur. Bref, l'ordre d'évaluation ou de correction importe.

Ce phénomène a été spécialement étudié par J.-J. Bonniol[2] qui y trouve un des déterminants essentiels de l'évaluation.

Dans une première expérience, Bonniol constitue deux groupes de neuf correcteurs ; le premier reçoit une série de devoirs dans un ordre donné et le second dans l'ordre inverse. Bonniol observe que les divergences (importantes) entre les deux groupes « sont plutôt imputables aux deux ordres de correction qu'aux différences de critères dont les examinateurs font état ». Les examinateurs notent par contraste avec le devoir précédent ; ils sont plus sévères pour les derniers travaux notés que pour les premiers.

Pour les expliquer, J.-J. Bonniol traite les effets d'ordre de correction comme un problème de *perception* des stimuli d'une série : « Comme dans une tâche de perception, ce qui est demandé au sujet auquel on donne à noter une série de devoirs, c'est d'effectuer un codage, de mettre en correspondance des éléments de deux ensembles : celui des notes chiffrées qu'il lui est possible d'utiliser et celui des dimensions en fonction desquelles peuvent être estimés les stimuli. »

Pour vérifier expérimentalement si les examinateurs jugent bien par contraste, Bonniol introduit un ou plusieurs devoirs très bons ou

1 J.-M. FABRE utilise l'expression plus large *effets de contexte (Jugement et certitude*, Berne, Lang, 1980, p. 21).

2 J.-J. BONNIOL, *Les comportements d'estimation dans une tâche d'évaluation d'épreuves scolaires. Etude de quelques-uns de leurs déterminants.* Aix-en-Provence, Université de Provence, 1972 (note de présentation d'une thèse de 3e cycle).

très mauvais dans une série de devoirs moyens. Les effets d'ordre peuvent alors être considérés comme des effets d'ancrage[1].

On émet deux hypothèses principales : l'introduction des ancres exercera des effets de contraste, se traduisant par des *déplacements* dans l'échelle d'évaluation par surestimation ou sous-estimation des travaux succédant à l'ancre dans la série, et par des modifications de l'étendue de l'échelle utilisée.

Deux fois sept groupes équivalents de 17 notateurs interviennent dans l'expérience.

Bonniol constitue deux séries de devoirs moyens : 17 versions anglaises (notation en négatif) et 14 résolutions d'une opération logarithmique simple (notation en positif).

Après notation de ces séries moyennes, Bonniol introduit :

- une ancre haute (très bon devoir) à la fin du premier tiers de la série ;
- une ancre haute (très bon devoir) au début du troisième tiers de la série ;
- trois ancres hautes (très bons devoirs) à la fin du premier tiers de la série ;
- trois ancres hautes (très bons devoirs) au début du troisième tiers de la série ;
- une ancre basse (très mauvais devoir), etc.
- trois ancres basses, etc.

Pour les versions anglaises, l'ancre exerce un effet sur un ou plusieurs des devoirs qui suivent ; ces effets sont plus importants avec une ancre lourde (trois devoirs) ; l'ancrage bas exerce plus d'effet que l'ancrage haut (notation en négatif).

En mathématiques, l'ancre lourde est nécessaire pour exercer un effet, qui est plus important si les ancres sont hautes (notations en positif).

Un phénomène de fluctuation.

Mais l'ordre dans lequel un examen est passé peut aussi être influencé par un autre phénomène qui reste peu étudié : la fluctuation du niveau d'exigence d'un évaluateur.

1 Bonniol définit l'ancre comme « un stimulus privilégié qui joue comme un stimulus de référence, soit parce qu'il est présent plus fréquemment que les autres, soit parce qu'il est situé dans une position particulière, soit parce qu'il est signalé d'une manière ou d'une autre à l'attention du sujet ». L'introduction d'ancres vient modifier le modèle, le niveau de performance auquel l'examinateur se référait pour évaluer.

Betz écrit à ce propos[1] : « On peut démontrer l'existence de fluctuations affectant l'évaluateur au cours d'une série d'examens. En raison même de la régularité du phénomène, il est peu probable que la montée et la descente périodiques des notes reflètent la prestation réelle des candidats. » Il semble que l'importance des fluctuations varie selon les évaluateurs et aussi le nombre d'examens auxquels ils procèdent en une même séance ou un même jour. Betz note d'ailleurs que la fréquence des fluctuations augmente à mesure que la séance se prolonge.

Et à l'intérieur d'une même copie ?

On dispose encore de peu d'études sur l'évolution du comportement de l'évaluateur pendant un même examen. Or on retrouve ici aussi des phénomènes d'ancrage. Noizet et Caverni écrivent[2] : « Il est probable que les premiers indices recueillis, qu'ils soient positifs ou négatifs, vont provoquer des inférences et que ces inférences vont guider le recueil des indices. C'est ainsi qu'à une *recherche ouverte* en début de lecture peut succéder progressivement une recherche sélective, l'évaluateur cherchant davantage à recueillir des indices susceptibles de confirmer ses premières inférences que des indices susceptibles de les remettre en question. » Bref, il semble que s'il doit faire des fautes, l'élève a intérêt à les faire dans la seconde moitié de son examen. C'est en tout cas ce qu'une expérience rapportée par Noizet et Caverni (p. 142) confirme.

10. **Manque de validité.**

Les examens traditionnels ne sont pas seulement de pauvres moyens d'inventaire; leur valeur pronostique apparaît tout aussi contestable dans bien des cas.

Nous avons observé, en compagnie du préfet d'un des grands athénées de Belgique, que deux catégories d'élèves réussissaient brillamment dans l'enseignement universitaire : ceux qui, pendant toutes leurs études, semblent s'être joués de la difficulté et sont restés en tête de leur classe, et une partie, non négligeable, de ceux qui n'ont presque jamais enregistré d'échecs, mais se sont néanmoins maintenus juste au-dessus de la note fatidique. Dans les deux cas, les performances dans l'enseignement secondaire sont de bon augure, moins par leur nature apparente que par la facilité d'adaptation, par la plasticité intellectuelle qui y a conduit.

1 D. BETZ, Rhythmische Schwankungen als Fehler in der Notengebung bei mündlichen Prüfungen, in *Psychologie in Erziehung und Unterricht*, 21, 1974, pp. 1-4.

2 G. NOIZET et J.-P. CAVERNI, *Psychologie de l'évaluation scolaire*, Paris, P.U.F., 1978, p. 141.

Pronostiquer la réussite ou l'échec en faculté sur la base du pourcentage obtenu au terme de l'enseignement secondaire est donc hasardeux.

11. **Un instrument d'immobilisme social.**

L'examen traditionnel sert au moins autant de moyen de sélection que d'outil de promotion de l'éducation individuelle. De plus, cette sélection s'opère, dans beaucoup de cas, moins en fonction de la qualification technique que de la qualification sociale.

Ainsi, l'examen devient un instrument privilégié de l'immobilisme social, phénomène analysé de façon pénétrante par J.-C. Passeron à qui nous empruntons l'essentiel des considérations suivantes[1].

a) *Effets irréversibles de la certification scolaire.*

Une fois obtenu, un titre scolaire suit l'individu toute sa vie, lui assure *grosso modo* son grade dans la hiérarchie professionnelle, son niveau de rémunération et le pouvoir qu'il détiendra.

Or, au moment de l'obtention d'un diplôme, « la compétence se trouve mesurée non pas au travers de l'activité qui sera exigée du travailleur, mais au travers d'une sorte d'activité analogique, préparatoire, ludique et presque fictive. Ce que mesurent les examens par rapport à la demande professionnelle, ce n'est pas ce que les gens seront censés faire, mais plutôt le niveau auquel il sera socialement exigible de les payer » (p. 7).

Ces observations n'ont certes pas de valeur absolue; elles correspondent néanmoins à une réalité incontestable. Par exemple, jamais on n'attribue le grade d'ingénieur à un ouvrier ou au technicien qui en a acquis la compétence fonctionnelle.

Le dernier examen qu'il subit avec fruit fixe donc souvent un individu dans une zone socioprofessionnelle dont il est, encore aujourd'hui, malaisé de sortir. La précocité de cette fixation est en contradiction profonde avec toute la politique contemporaine d'éducation et donc de promotion permanente.

Que l'examen accorde plus une certification sociale qu'une garantie de compétence technique serait d'ailleurs confirmé par le fait que, dans beaucoup d'entreprises, deux personnes occupent le même poste :

1 J.-C. PASSERON, Sociologie des examens, in *Education et Gestion*, 1970, 2, pp. 6-16.

«L'une officiellement parce qu'elle a tous les titres scolaires requis pour l'occuper (avec la rémunération, le prestige et le pouvoir), l'autre effectivement, parce qu'elle assume la part technique de la tâche (...), c'est-à-dire qu'elle fait «tout le travail» (...). Parfois, le second conviendrait techniquement au poste, bien mieux que le titulaire. Mais c'est précisément ce qu'empêche l'effet de la certification produit par l'examen qui, une fois pour toutes, et pour toute la vie, tend à définir ce que vaut la prestation professionnelle d'un individu par la valeur de cet individu sur le marché des titres scolaires.»

De nouveau, cette constatation appelle des restrictions. Dans une entreprise, la seule compétence technique suffit rarement aux cadres: la capacité de communiquer, d'exercer des fonctions de *leader* et de relations publiques, d'harmoniser le travail d'administration et de production, etc., sont autant de qualités dont la manifestation échappe souvent à ceux qui ne voient que l'aspect technique des choses. Mais ces qualités ne sont pas toujours réelles, loin s'en faut. Souvent, on a affaire à un formalisme dont nous allons reparler.

b) *Les examens ne sont pas socialement neutres.*

Une enquête de l'INSEE[1], en France, révèle que si, au cours des dernières années, les chances d'accès à l'enseignement supérieur se sont accrues pour toutes les catégories sociales, «l'augmentation globale du taux de scolarisation s'est ventilée entre les diverses catégories socioprofessionnelles dans la même proportion que les inégalités antérieures. Bref, il s'agit d'une simple *translation vers le haut* de la structure des inégalités» (p. 8). Le tableau ci-dessous est révélateur.

Catégorie socioprofessionnelle du père	*Probabilités d'accès à l'enseignement supérieur*	
	1961-62	*1965-66*
Salariés agricoles	1,1	2,7
Agriculteurs	3,4	8
Ouvriers	1,3	3,4
Employés	9	16,2
Cadres moyens	24,9	35,4
Professions libérales et cadres supérieurs ...	38	58,7

1 Institut national de statistiques et d'études économiques.

Comment s'explique ce phénomène en apparence contradictoire avec l'obligation scolaire généralisée, de plus en plus longue, la gratuité des études et la répartition des aptitudes, selon la loi du hasard, dans toutes les couches de la société?

La différence d'éducation familiale, au cours des premières années, créerait rapidement des différences dans les mécanismes intellectuels, les moyens linguistiques et les attitudes devant le travail: «L'éducation donnée par les familles prédispose d'autant moins à la réussite aux examens que ces familles appartiennent à une classe plus éloignée de la culture scolaire, de la culture savante» (p. 11).

Selon la famille, le milieu rural ou urbain, le niveau de développement du pays où ils naissent, des individus théoriquement doués d'un même potentiel de départ ont donc des chances d'études très diverses.

L'ensemble impressionnant des dernières recherches de l'Association Internationale pour l'Evaluation du Rendement Scolaire (I.E.A.)[1] confirme que, dans les conditions actuelles, l'origine familiale prédit mieux la performance scolaire que toute autre variable.

Dans une communication récente, B. Bloom[2] a aussi indiqué que, pour une même durée d'études, par exemple douze ans, le rendement scolaire moyen peut varier du simple au double selon le degré d'industrialisation du pays.

Bref, pour en revenir à Passeron, les examens, théoriquement neutres, se borneraient à enregistrer les effets profonds de la première éducation. Force est de reconnaître, nous venons de le voir, que cette thèse se vérifie souvent. Même si on n'a pas affaire à un déterminisme inéluctable, il reste néanmoins incontestable que les examens actuels ne sont pas socialement neutres.

Passeron continue: les procédures de notation et les types d'épreuves utilisés prennent en compte, au moins autant que les aptitudes techniques, certains aspects gratuits de la performance, qui n'ont aucune importance technique, mais qui sont en revanche très fortement liés aux habitudes culturelles de telle classe sociale plutôt que de telle autre (p. 12). Ainsi les examens français les plus prestigieux sont moins des *épreuves de connaissances que des épreuves de manières ou des exercices d'usage lettré du langage.*

1 L.C. COMBER and J. KEEVES, *Science Education in Nineteen Countries*, I.E.A., Stockholm, Malmqvist, 1973.
A.C. PURVES, *Literature Education in Ten Countries*, I.E.A., Stockholm, Malmqvist, 1973.
R.L. THORNDIKE, *Reading Comprehension in Fifteen Countries*, I.E.A., Stockholm, Malmqvist, 1973.

2 B.S. BLOOM, *Time and Learning*, Communication au 81e Congrès de l'American Psychological Association, Montréal, 1973.

La dissertation reste, dans cette perspective, le moyen privilégié. Rejoignant la thèse sociologique de Durkheim, de Max Weber, d'A. Clausse, Passeron rappelle que le formalisme a toujours été un moyen de défense des classes privilégiées. Il est, en effet, frappant de constater que l'on accorde le plus d'importance à une épreuve, la dissertation, qui échappe le plus radicalement à toute notation objective, technique... Or, le formalisme consiste précisément à définir la culture, non par son contenu objectif, mais par ces impondérables que sont la manière, la nuance. Cette «indéfinition» permet d'exercer une *fonction de fermeture* au profit d'un groupe favorisé.

La neutralité des examens scientifiques n'est d'ailleurs pas tellement mieux garantie, car eux aussi font souvent intervenir le langage, et bien peu d'examinateurs restent indifférents à l'élégance de l'expression, pourtant étrangère au problème technique sur lequel porte fondamentalement l'examen.

Quant à l'examen oral, il ne fait que renforcer l'effet de la belle présentation, du langage châtié, de la diction élégante.

Notons, pour conclure, que la floraison des examens à caractère sélectif paraît caractéristique des conditions culturelles du XIX^{e} siècle, en particulier du développement d'une bureaucratie très hiérarchisée au service de l'économie capitaliste. Il est donc naturel que la nouvelle forme de civilisation vers laquelle nous évoluons s'accompagne d'une mise en question d'un type d'examen conçu pour d'autres conditions.

La conception très déterministe qu'exprime la *théorie de la reproduction* de P. Bourdieu et J.-C. Passeron[1] est fortement nuancée par P. Perrenoud[2]. Il souligne que l'école ne reproduit pas simplement les inégalités sociales observables dans le monde adulte, mais qu'elle en crée aussi. Sans contester les mécanismes de reproduction, Perrenoud montre que l'école joue également un rôle en créant elle-même des inégalités par ses choix curriculaires et ses pratiques d'évaluation. N'a-t-elle pas le pouvoir d'affirmer que tel élève est excellent, et de faire en sorte qu'il soit traité comme tel dans la société ? «... Les organisations ont le *pouvoir de construire une représentation de la réalité et de l'imposer à leurs membres et à leurs usagers comme la définition légitime de la réalité.* A aucun moment, le jugement de l'école ne se donne comme *un* point de vue sur l'élève parmi d'autres possibles. Dans le champ couvert par la norme d'excellence, l'école prétend attribuer à

1 P. BOURDIEU et J.-C. PASSERON, *La reproduction. Eléments pour une théorie du système d'enseignement*, Paris, Ed. de Minuit, 1970.

2 P. PERRENOUD, *La fabrication de l'excellence scolaire*, Paris, Droz, 1984.

chacun son *vrai* niveau d'excellences et fonder sur cette évaluation une décision sans appel. Le pouvoir de l'organisation scolaire, qu'elle tient évidemment du système politique, est de faire d'un enfant qui se trompe dans les soustractions, n'accorde pas le verbe avec le sujet ou ne maîtrise pas le passé simple, un «mauvais élève» (p. 17).

12. **Faiblesse de beaucoup d'expériences docimologiques.**

Depuis plus d'un demi-siècle, on accumule les expériences destinées à montrer les désaccords entre notateurs ou le manque de fidélité d'un même correcteur.

Le problème existe assurément. Mais on l'aggrave artificiellement, dans bien des cas, en omettant d'inviter les notateurs à s'entendre sur les aspects à considérer et l'importance relative à leur réserver. Autrement dit, on continue à enregistrer patiemment les résultats d'une politique anarchique des examens.

13. **Autres critiques.**

Bien d'autres critiques s'adressent encore aux examens. Le Rapport de la Commission Consultative sur les Examens dans l'Enseignement secondaire de Grande-Bretagne, *déposé dès 1911*, nous en fournit une liste. Nous l'allongeons un peu, tout en ayant l'impression de n'être guère exhaustif.

Pour l'élève:

- Il consacre trop d'énergie à reproduire les idées des autres au lieu de développer sa créativité.
- Il est récompensé pour des apprentissages souvent éphémères. (On a démontré que jusqu'à 80% des connaissances, surtout factuelles, apprises pour l'examen, ont disparu quinze mois après.)
- Il apprend l'obéissance passive aux consignes.
- Son aptitude à s'exprimer prend parfois le pas sur le contenu. (On montrera plus tard combien les handicaps socioculturels se marquent surtout dans le domaine verbal.)
- Il renonce souvent à exprimer un jugement personnel, pour se plier aux idées du professeur qui évaluera l'examen.
- Ses apprentissages peuvent être viciés par un esprit de compétition confinant parfois à l'esprit mercenaire.
- Il apprend à spéculer sur sa chance: dans l'examen traditionnel, le petit nombre de questions, reflet des idiosyncrasies du professeur, ouvre la voie au jeu des «tuyaux».

- Les examens traditionnels, leur longue préparation et la période d'essoufflement qui suit, raccourcissent considérablement l'année scolaire effective.
- Les examens traditionnels empêchent le travail en groupe et exaltent, au contraire, la valeur de la performance individuelle, source d'égoïsme.
- La menace d'échec qu'ils font peser - souvent doublée de la peur d'une sanction familiale - incite à la fraude.
- Le succès des fraudeurs constitue un mauvais exemple pour les élèves restés honnêtes.
- Les examens donnent une idée fausse du travail adulte où, plus la question est complexe et difficile, plus on s'entoure de conseils et d'ouvrages de référence.

Pour le professeur:

- Il enseigne en fonction de l'examen, voire des exigences particulières des membres d'un jury extérieur.

  Nous verrons qu'en bonnes conditions, l'esprit de l'enseignement et de l'examen doit être un. On fait allusion ici à des examens étrangers aux objectifs éducationnels.
- Il est jugulé dans sa méthodologie, si l'esprit de l'examen est étranger aux objectifs pédagogiques qui lui paraissent essentiels.

  Comment pratiquer une pédagogie de la découverte, de l'exploration personnelle - processus lents, mais générateurs d'apprentissages profonds - si l'examen imposé est de caractère encyclopédique?
- Esclave du programme, il ne laisse pas les élèves avancer à leur allure propre, et donc digérer la matière.
- Il tend à accorder trop d'importance aux aptitudes et aux connaissances utiles à l'examen.

Par après, on a baptisé *effet de reflux* le phénomène par lequel les professeurs modifient la méthode et le contenu de leur enseignement en fonction de l'évolution des examens imposés de l'extérieur.

Des aspects éducatifs importants peuvent ainsi être négligés et l'ont souvent été.

En résumé, les examens traditionnels présentent généralement de graves défauts de construction. Leur validité est contestable. L'évaluation des travaux est, de son côté, grevée de lourdes faiblesses. En outre, les examens peuvent nuire à la santé physique et mentale des élèves. On ajoutera enfin que, dans certains cas, les professeurs font leur cours en fonction d'un examen et non le contraire...

Mais il est temps de donner la parole à la défense.

CHAPITRE 2

DEFENSE DE LA NOTE SUBJECTIVE ET DE L'EXAMEN

Que l'on ait abusé des examens au point de vicier l'action éducative de l'école, personne ne le contestera. Que bien des concours furent surtout de sinistres loteries et duperies semble aussi hors de doute. Que des scores se virent attribuer une signification qu'ils n'avaient pas est également patent.

De là à conclure qu'examens et concours doivent être à jamais proscrits pour ne laisser subsister que des séries d'évaluations occasionnelles et les rapports qui les synthétisent paraît utopique.

Dans la première partie, nous avons vu que les différents modes d'évaluation correspondent à des fonctions particulières. Il suffit d'en reprendre la liste pour constater qu'elles ne peuvent être toutes remplies par une seule procédure.

Par ailleurs, tous les modes d'évaluation impliquent l'établissement d'un score ou d'une note que l'on souhaite parfaitement objectifs ou, mieux, parfaitement contrôlés: entendons par là que nous reconnaissons aux maîtres le droit d'adapter beaucoup de leurs évaluations, autant en fonction de l'élève que de la matière, à condition qu'ils sachent exactement ce qu'ils font.

Avant d'envisager certains avantages souvent reconnus aux examens, une discussion de caractère négatif s'impose.

1. La mesure rigoureuse est peut-être impossible.

Rechercher les voies de la parfaite validité des examens et de la fidélité des évaluations sur deux postulats rappelés par J. Guillaumin[1]:

1° «Que les productions de l'élève sont par nature mesurables, quantifiables.»

1 J. GUILLAUMIN, L'aspect interpersonnel de la notation scolaire: de la docimologie à la doxologie, in *Bulletin de la Société A. Binet et T. Simon*, IV, 1968, p. 270.

2° Que les différences qu'on peut trouver entre les mesures pratiquées par les notateurs sont susceptibles d'être réduites.

Si ces deux postulats sont faux, la docimologie classique s'écroule. Or, on constate que, bien que ses grands principes sont définis depuis plusieurs décennies, elle n'a guère pénétré à l'école. Cet échec serait dû, pour une bonne partie, à une incompatibilité entre la nature psychologique de la situation d'enseignement et le caractère mathématique dominant des procédés de mesure proposés.

Le débat n'est pas nouveau et l'issue est dans le compromis. Les progrès considérables de la psychologie et de la pédagogie contemporaines sont largement dus à l'objectivation de l'observation, donc à la mesure. Toutefois, les chercheurs en sciences humaines ont aujourd'hui une conscience assez claire de leurs limites et, en particulier, de l'impossibilité presque générale d'utiliser des échelles de mesure mathématiquement parfaites. Ils savent aussi que nous ne mesurons le plus souvent que des comportements isolés du tout humain. Comment évaluer rigoureusement des entités hypothétiques comme l'esprit critique, le sens de l'observation, etc., sinon à travers des performances particulières supposées représentatives de l'ensemble considéré ?

De même que l'on n'est jamais parvenu, jusqu'à présent, à mesurer avec précision le rendement de l'enseignement, tant les points de vue à considérer à des moments différents sont nombreux, de même on ne peut rendre analytiquement compte de la valeur réelle d'une performance scolaire complexe. Même si, en calcul, le résultat ne peut être que juste ou faux, la démarche de la pensée et l'effort produit peuvent varier considérablement d'un sujet à l'autre et donc être difficiles à chiffrer objectivement. Quant à la composition française, considérée comme œuvre d'art, elle échappe à l'évaluation parcellaire[1]. Aussi, le droit à la subjectivité, à la réaction globale, compte, on le comprend, d'ardents défenseurs.

Si la réponse à donner à des questions de géographie, d'histoire, de sciences naturelles contraint l'élève à s'exprimer oralement ou par écrit, faut-il toujours faire abstraction de l'élégance du langage, de l'ordonnance de la pensée, de la rigueur du raisonnement ?

1 Signalons néanmoins qu'Ellis Page a réussi à construire un programme d'ordinateur qui permet d'évaluer automatiquement les compositions de langue maternelle. La corrélation avec les notes traditionnelles est élevée. Toutefois, il faudrait savoir si, dans ce cas, le chercheur n'a pas simplement fait adopter par l'ordinateur les démarches imparfaites des notateurs ...

E. PAGE et D. PAULUS, *The Analysis of Essays by Computer*, Washington, U.S. Office of H.E.W., Project 6-1318, 1968.

Pour la première fois dans l'histoire, un échantillon national de rédactions a été corrigé par ordinateur aux Etats-Unis en 1971 à l'occasion du *National Assessment of Educational Progress*. Cf. NAEP *Newsletter*, vol. V, n° 2, 1972, p. 1.

Bref, dès que l'on ne se résout plus à ramener la réussite ou l'échec à des critères rigides et souvent schématiques, les productions des étudiants échappent à la quantification automatique, impersonnelle.

Cet argument ne suffit assurément pas pour renoncer à objectiver une partie de la notation des élèves, mais une partie seulement. L'appréciation globale du maître, tout en finesse, sa sensibilité tant à la performance matérielle qu'à l'effort de dépassement et à la faiblesse humaine, doivent garder leur place.

La richesse de l'enseignement réside avant tout dans la qualité de la relation humaine qu'il crée et l'évaluation est un des aspects de cette relation. Si elle s'appauvrit au point de ne laisser subsister qu'une communication impersonnelle, le maître peut être avantageusement remplacé par une machine à enseigner.

2. Les maîtres jugent bien leurs élèves.

Si les maîtres tendent à relativer leurs jugements par rapport au niveau moyen du groupe, leur classement à l'intérieur de celui-ci n'en possède pas moins une validité élevée.

Il suffirait donc d'appliquer des mesures permettant de rendre ces jugements comparables entre écoles (nous verrons comment au chapitre de la modération) pour disposer de points de repère importants.

La sûreté de jugement des maîtres s'explique par plusieurs facteurs :

1. ils fondent leur jugement sur une observation longue et continue;
2. ils considèrent un beaucoup plus grand nombre de facteurs (notamment de personnalité) que l'examen;
3. ils peuvent, en particulier, tenir compte de comportements exceptionnels (en telle occasion, tel élève a fait preuve d'une lucidité, d'une originalité peu communes) qui n'apparaîtront probablement pas à l'examen.

La valeur prédictive des jugements des instituteurs (ajustés pour les rendre comparables) a été bien mise en lumière par une recherche faite par la *National Foundation for Educational Research in England and Wales*[1].

1 Voir F. YATES and D. PIDGEON, *Admission to Grammar School, o.c.*, pp. 57 sqq.

En 1951 et en 1952, environ 1200 élèves de la région de Twickenham ont été examinés, lors de leur sélection, à l'entrée dans les lycées *(Grammar Schools)*. Leurs résultats ont été étudiés deux ans, puis trois ans après.

Epreuves subies au départ:

Tests:

1. test d'intelligence verbale (V);
2. test standardisé de connaissances en langue maternelle (E1) (questions à choix multiple);
3. test standardisé de connaissances en arithmétique (A) (calculs et problèmes);
4. test d'intelligence non verbale (N/V);
5. test d'aptitude spatiale (Sp. 1) (espace à deux dimensions);
6. test d'aptitude spatiale (Sp. 2) (espace à trois dimensions);
7. test de connaissances en langue maternelle (E2) (questions plus ouvertes que E1).

Jugement par un instituteur de 6e primaire:

1. prédit la réussite des études à la *Grammar School* sur une échelle à 15 degrés (M);
2. classe ses élèves selon l'ordre de leur réussite à la *Grammar School*. Ces jugements sont ajustés par les chercheurs en fonction d'un test d'intelligence verbale (voir description du système au chapitre de la modération) (F).

Epreuves utilisées pour mesurer le succès dans l'enseignement secondaire: ordre de mérite fourni par le préfet des études, ajusté en fonction des résultats à une batterie de tests objectifs d'aptitudes et de connaissances, administrés immédiatement après l'élaboration du classement (S.H.A.).

Le tableau ci-dessous montre, pour le groupe de 1951, la corrélation entre chaque épreuve de prédiction et les résultats en cours d'études:

Prédicteur	*Groupe de 1951: deux ans après*	*Groupe de 1951: trois ans après*
Jugement de l'instituteur (F)	.821	.748
(M)	.796	.722
Test Verbal (V)	.789	.704
E2	.749	.623
A	.734	.659
E1	.729	.622
N/V	.648	.535
Sp. 1	.565	.453
Sp. 2	.491	.361

Le jugement ajusté des instituteurs se révèle le meilleur prédicteur. La valeur pronostique élevée du simple test verbal est aussi confirmée.

Il faut y insister: ces observations ne sont valides que dans la situation considérée: la réussite dans les *Grammar Schools*. Nous ne savons pas ce qui se passerait dans d'autres formes d'enseignement. Nous disposons néanmoins ici d'une observation strictement contrôlée dont on n'a nulle raison de croire qu'elle ne puisse s'appliquer à d'autres cas.

3. **Validité limitée mais réelle des examens traditionnels.**

Quelles qu'en fussent les imperfections, les examens traditionnels n'ont pas empêché notre civilisation d'atteindre un niveau scientifique, jamais égalé dans l'histoire de l'humanité. Le filtrage qu'ils ont opéré s'est donc révélé, au moins partiellement, valide.

La chose est évidente. Avec la restriction faite au chapitre précédent, on peut affirmer que les examens traditionnels ont permis d'identifier les élèves les mieux et les moins bien doués, tri capital dans la forme de civilisation que nous avons connue dans la dernière centaine d'années.

Que la procédure ait été grossière (en ce sens qu'elle a ignoré ou gaspillé des talents, au détriment des couches sociales défavorisées surtout) et injuste envers la population des élèves moyens, mis brutalement dans le même sac, n'enlève rien au fait que, dans une démocratie peu évoluée, les examens ont bien joué leur rôle.

4. **S'endurcir pour la vie.**

Selon les behavioristes, une conduite ne s'apprend que si elle est effectivement produite. Il paraît donc souhaitable que, périodiquement, l'étudiant soit amené à faire un effort exceptionnel, à bander son énergie, à affronter l'ordalie des examens, avec toutes ses imperfections, voire ses injustices. Car la vie ne lui épargnera pas semblables épreuves, et il est bon d'y être préparé.

Par ailleurs, la concurrence, la compétition sont des traits fondamentaux de notre civilisation. L'examen et le concours existent et l'on peut même souhaiter qu'ils s'imposent de plus en plus là où il y a plus de candidats que de places ou d'emplois disponibles, sinon le favoritisme sous toutes ses formes et l'inefficacité qui l'accompagne ont libre jeu.

5. **Se situer par rapport aux autres.**

Si un classement défavorable à une épreuve peut donner un choc, il permet aussi de se situer par rapport aux autres (non dans l'absolu, mais dans les conditions de l'examen!). Celui qui a fait de son mieux peut ainsi mieux ajuster ses ambitions; celui qui ne l'a pas fait a l'occasion de découvrir, peut-être avant qu'il ne soit trop tard, les conséquences de ses faiblesses.

En règle générale, le succès profite mieux que l'échec, mais une certaine frustration constitue, pour certains, un utile aiguillon.

6. **Large synthèse et intégration des connaissances.**

Il est indéniable que les examens portant sur de vastes ensembles de connaissances obligent l'étudiant à construire des synthèses à l'occasion desquelles il perçoit, parfois pour la première fois, l'économie de tout l'édifice, les relations entre parties et, éventuellement, les points communs entre différentes disciplines.

7. **L'examen externe contrôle le professeur.**

Un examen régional ou national constitue, dans une mesure qui reste à définir, un moyen de contrôle du travail des professeurs.

S'il ne peut plus être question d'asservir les maîtres à des programmes surabondants, fixés en détail, il n'en reste pas moins nécessaire que l'enseignement se déroule selon un plan d'étude précis dans ses objectifs et défini dans ses matières principales.

Même si l'on parvient à proposer des programmes individualisés ou semi-individualisés, les buts à atteindre devront être définis, au moins provisoirement, avant que ne s'engage l'action pédagogique.

Toujours, la communauté éducative, en particulier les parents et les pouvoirs organisateurs auront le droit de vérifier si la mission enseignante a été accomplie.

L'existence d'un contrôle portant sur tous les éléments importants du plan d'étude semble aussi une saine sauvegarde contre la laxité, le relâchement. Consacrer le temps nécessaire à l'apprentissage est louable; musarder est condamnable. Il me souvient d'un professeur de langues qui, en deux ans, n'avait pas dépassé l'introduction du cours... Apparemment, aucun inspecteur ne s'aperçut de la chose.

Enfin, l'examen permet d'établir, au moins dans une certaine mesure, si le professeur ne concentre pas trop ses efforts sur certains élèves, au détriment des autres.

8. **L'examen externe, feed-back pour le professeur.**

Tant pour notre santé mentale que pour la meilleure efficacité de notre action, nous avons besoin d'être informés de la validité de nos comportements.

L'examen bien conçu permet au professeur de juger de la valeur de certains aspects de son enseignement à travers les apprentissages réalisés par des élèves.

Il serait erroné d'évaluer les maîtres uniquement sur les résultats obtenus à court terme. Toutefois, seul ce genre d'évaluation semble actuellement pouvoir être effectué avec quelque rigueur scientifique. Or, tous ceux qui exercent une profession se rattachant aux sciences humaines souffrent, à vrai dire à des degrés fort divers, du manque d'information sûre sur la pertinence de leur action. C'est le souci du magistrat comme de l'assistant social, du professeur comme du prêtre.

Quand un maître peut-il dire avec certitude qu'il a bien rempli sa mission?

★
★ ★

En résumé, nous ne sommes partisans ni de supprimer complètement les examens, ni de renoncer entièrement à la notation subjective. Il importe d'adopter une façon de faire qui emprunte à chaque procédure ce qu'elle a de meilleur et de plus sûr. Nous allons essayer de voir comment pareil but pourrait être atteint.

TROISIEME PARTIE

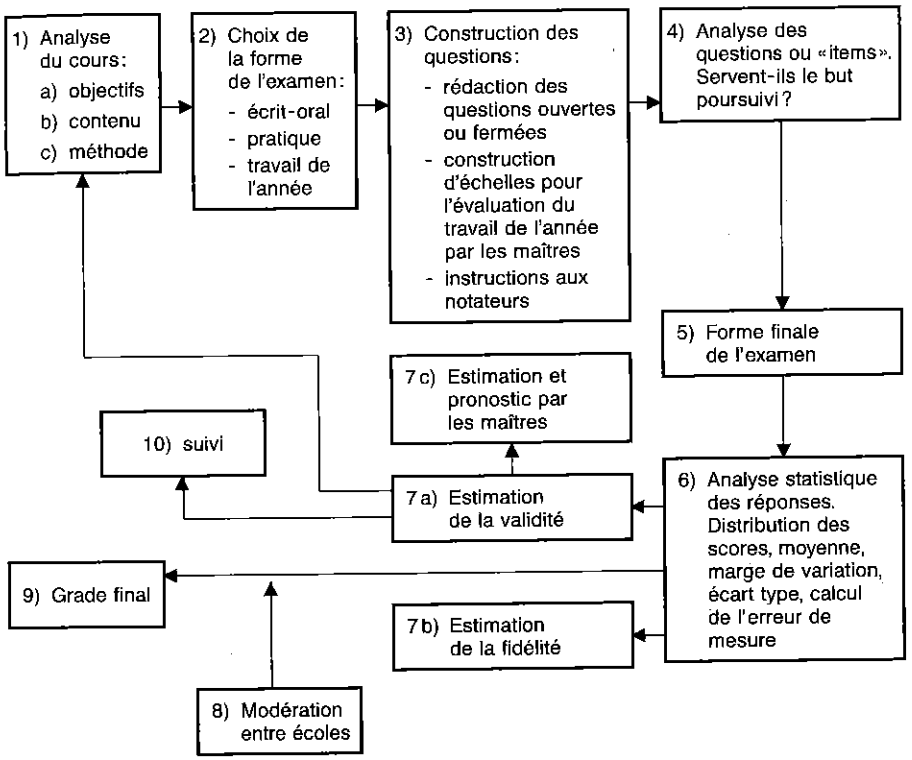
CONSTRUCTION DE L'EXAMEN

Comme les tests standardisés, les examens préparés par les maîtres devraient se construire par étapes bien définies. Il n'est, certes, pas possible de déployer, pour une simple interrogation de routine, les mêmes efforts que pour une épreuve destinée à des centaines d'élèves. Pourtant, les exigences de base restent les mêmes.

Les grandes phases de la construction d'un examen sont:

1. définir l'objet et les objectifs;
2. rédiger les questions;
3. standardiser la présentation, l'administration et la correction;
4. étalonner;
5. éprouver la fidélité de l'examen;
6. établir la validité.

Le schéma suivant montre le chemin que nous allons parcourir[1].



1 Adapté d'après *Examinations Bulletin* n° 3, 1964, p. 33.

CHAPITRE 1

L'OBJET ET LES OBJECTIFS

Voir clairement la raison d'être de l'examen que l'on prépare et définir, avec le plus de précision possible, les types d'apprentissages à évaluer est probablement l'étape la plus décisive dans la construction d'une épreuve. Non seulement sa validité en dépend, mais aussi le mode d'évaluation à adopter.

I. L'objet

Trois objets sont possibles : le *pronostic*, l'*inventaire* et le *diagnostic*. Il est rare qu'une épreuve scolaire puisse se ranger exclusivement dans l'une de ces trois catégories. Le plus souvent, les examens visent à faire le bilan, l'inventaire des acquisitions, mais ils remplissent en même temps une fonction pronostique à court ou à moyen terme (L'élève est-il prêt à aborder un nouveau chapitre du cours ? L'élève pourra-t-il suivre avec succès les cours de la classe supérieure ?), et une fonction diagnostique (Où l'élève achoppe-t-il ? Pourquoi ?).

Par ailleurs, il importe de savoir si l'examen s'insère dans l'action éducative quotidienne - cas où l'élève sera comparé à lui-même et au groupe qui l'entoure (le plus souvent, sa classe) - ou s'il sert à informer l'individu de sa valeur relative dans sa communauté ou dans son pays ou à le classer dans un concours - cas où des normes adéquates devront être utilisées.

A. Le pronostic.

A long terme (plus d'un an, au moins), le pronostic formulé à partir des résultats d'un examen scolaire est souvent décevant. Les changements de maître, de méthode, d'esprit d'un cours peuvent provoquer une transformation profonde de l'attitude chez l'élève. De plus, surtout chez les sujets jeunes, les intérêts manifestés connaissent d'importantes fluctuations. Il faut aussi tenir compte de l'évolution psychologique des élèves : par exemple, les professeurs d'enseignement secondaire connaissent bien l'incidence de la crise d'adolescence sur le rendement scolaire.

On a maintes fois démontré que, pour le pronostic à long terme, les tests d'intelligence sont au moins d'aussi bons prédicteurs que les tests de connaissances et sont beaucoup plus sûrs que les résultats scolaires[1]. Ces tests explorent des aptitudes qui, surtout après l'âge de 8 ou 9 ans, connaissent une grande stabilité[2] ; elles touchent le raisonnement, les facteurs verbal, spatial, numérique, ...

En outre, si les dominantes profondes de la personnalité, y compris les zones d'intérêt (et non des intérêts particuliers, manifestés occasionnellement) sont identifiées, la prédiction du succès scolaire et académique peut atteindre un niveau élevé de sûreté.

A court terme, le pronostic pédagogique s'établit de trois façons :

1) *Tests de maturité spécifique ou tests de préparation (readiness).*

Ils portent sur les formes de raisonnement, sur les aptitudes qui interviennent dans un apprentissage déterminé, par exemple : apprentissage de la lecture. La construction et souvent aussi l'utilisation de ces instruments complexes sont du ressort du spécialiste. Quelques tests permettant un premier dépistage sont toutefois conçus pour les maîtres, par exemple : le *Test d'Inizan* pour l'apprentissage de la lecture.

Dans les pays de langue française, il serait souhaitable que les centres de recherche augmentent considérablement leurs efforts dans le domaine de la maturité spécifique. Nous avons montré ailleurs combien nous sommes encore mal équipés pour la plupart des branches et à presque tous les niveaux[3].

2) *Vérification des connaissances-clés ou notions critiques* nécessaires aux acquisitions nouvelles (tests de prérequis).

Des épreuves de ce genre peuvent être assez aisément construites par les maîtres, pour autant qu'ils aient dressé systématiquement la liste des erreurs commises lors d'un enseignement antérieur.

Ici aussi, des fiches de recherche précisant le prérequis pour aborder les points importants de la matière rendraient aux maîtres d'inestimables services.

---

1 Voir F. HOTYAT (1962), W. McCLELLAND (1945), E. PEEL et D. RUTTER (1951), P. VERNON (1967), F. BACHER (1965), etc.

2 Certaines modifications peuvent toutefois encore se produire. On sait aujourd'hui que le *Quotient Intellectuel* n'est pas fixé une fois pour toutes.

3 Voir G. DE LANDSHEERE, *Les tests de connaissances*, Bruxelles, Editest, 1965.

3) *Essai.*

Pour déterminer si un élève est prêt à aborder une matière nouvelle, on lui présente les premiers éléments de cette matière et l'on observe systématiquement les réactions. Une leçon programmée de type mixte (Skinner-Crowder) constitue souvent un heureux alliage d'enseignement et de contrôle analytique permettant un pronostic à court terme.

B. **L'inventaire** (épreuves de rendement). **L'évaluation sommative.**

Il a pour but premier de mesurer les apprentissages réalisés au cours d'une période plus ou moins longue. Sauf dans les cas de concours ou d'enquêtes normatives *(surveys)* préparant des réformes, les inventaires remplissent aussi, à l'école, une fonction pronostique et diagnostique.

Il semble, en effet, contraire à toute attitude éducative saine qu'un maître, constatant des faiblesses graves, n'essaie pas d'en localiser la cause pour y remédier et ne s'inquiète pas des difficultés probables que les faiblesses annoncent pour les apprentissages ultérieurs.

Pour cette raison, notamment, il importe de distinguer soigneusement dans tout inventaire les notions critiques, indispensables pour certains apprentissages ultérieurs, des notions marginales.

Depuis quelques années, les expressions *évaluation sommative, évaluation formative* sont entrées dans l'usage. L'évaluation sommative revêt le caractère d'un bilan. Elle intervient donc après un ensemble de tâches d'apprentissage constituant un tout, correspondant, par exemple, à un chapitre de cours, à l'ensemble du cours d'un trimestre, etc. Les examens périodiques, les interrogations d'ensemble sont donc des évaluations sommatives.

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, ...

C. **Le diagnostic. L'évaluation formative.**

L'évaluation formative, à caractère diagnostique, intervient, en principe, au terme de chaque tâche d'apprentissage. Elle a pour objet d'informer élève et maître du degré de maîtrise atteint et, éventuellement, de découvrir où et en quoi l'élève éprouve des difficultés d'apprentissage, en vue de lui proposer ou de lui faire découvrir des stratégies qui

lui permettent de progresser. L'expression «évaluation formative», due à Cronbach et à Scriven, marque bien que l'évaluation fait, avant tout, partie intégrante du processus éducatif normal, les «erreurs» étant à considérer comme des moments dans la résolution d'un problème (plus généralement comme des moments dans l'apprentissage), et non comme des faiblesses répréhensibles ou des manifestations «pathologiques». L'évaluation formative permet aussi de déterminer si un élève possède les prérequis nécessaires pour aborder la tâche suivante, dans un ensemble séquentiel.

D. Durrell[1] estime que les causes les plus communes des difficultés scolaires sont les suivantes:

1. le manque d'aptitudes pour effectuer une tâche;
2. la connaissance imparfaite des éléments de base;
3. un enseignement mal ajusté au niveau d'aptitudes de l'enfant et à sa vitesse d'apprentissage;
4. l'acquisition de mauvaises habitudes qui freinent le progrès;
5. l'inaptitude à transférer, à découvrir le «système», à généraliser les techniques de solution;
6. le manque de vigueur dans l'attaque des problèmes, résultant d'échecs répétés et du manque d'intérêt.

Si des difficultés scolaires surgissent, l'état physique et la personnalité requièrent aussi la plus grande attention. On n'oubliera pas toutefois que beaucoup de problèmes émotionnels de l'élève proviennent de l'échec au lieu d'en être la cause.

Dans la pratique quotidienne de l'école, le diagnostic fin devrait être une des préoccupations dominantes des professeurs et donc occuper un temps important. Contrairement à ce qu'une sorte de mythe de la courbe de Gauss (voir p. 235) a parfois fait croire, l'enseignement idéal serait celui qui, pour les notions clés au moins, amènerait tous les élèves d'un groupe à une performance parfaite.

Quelques bons tests diagnostiques ont été publiés et les maîtres ne peuvent les ignorer. Mais, le plus souvent, ils pourraient construire eux-mêmes des instruments analytiques simples et efficaces, spécialement adaptés à leurs besoins.

On imagine aisément l'utilité d'un tableau d'ensemble où, pour chacun des élèves d'une classe, on voit apparaître par + ou par - si la réponse a été ou non correcte pour trois exercices portant sur un des

1 D. DURRELL, *Analysis of Reading Difficulties*, New York, Harcourt, 1955.

points de la matière[1]. Il suffit de colorier en rouge les cases qui contiennent deux ou trois signes moins pour obtenir une première impression de la situation[2]. Il y a quelques années, nous avons dressé et tenu, avec la collaboration des élèves, un tableau de ce genre pour la prononciation d'une langue étrangère. Les résultats furent positifs.

Pour que l'épreuve puisse jouer son rôle diagnostique, il faut évidemment que les exercices portent sur un aspect très précis de la matière et qu'ils soient conçus, si possible, de façon à ne mettre en cause qu'une difficulté à la fois.

Comme l'élaboration de pareils contrôles devient facilement lourde, la collaboration de plusieurs professeurs est souvent souhaitable.

D. **Psychologie cognitive et perspectives nouvelles.**

Tant pour le pronostic que pour les épreuves de rendement ou de diagnostic, les progrès récents de la psychologie cognitive ouvrent des perspectives nouvelles. Même si les notions théoriques qui vont être évoquées - de façon très simplifiée - ne peuvent pas encore déboucher aisément sur des pratiques quotidiennes, elles aident au moins à prendre conscience de problématiques fondamentales.

Au lieu de se contenter de constater qu'un élève n'a pas trouvé la solution d'un problème, n'a pas été capable d'accomplir telle tâche, la psychologie cognitive veut savoir quelle démarche mentale explique le phénomène. Grâce surtout aux travaux de Piaget, tout éducateur sait que les possibilités d'un élève sont liées au stade de développement cognitif, affectif, psychomoteur qu'il a atteint.

Par une comparaison avec l'ordinateur, le traitement de l'information par le cerveau peut être décrit de la façon suivante:

- Les structures du sytème cognitif peuvent être activées par des stimuli reçus par les organes des sens ou par d'autres structures de connaissances (pensée); des informations sont ainsi *entrées.*
- La *mémoire à court terme* ou *mémoire active* retient un nombre limité d'informations pendant une courte période de temps. Les faits isolés sont oubliés les premiers.

1 Dans une épreuve diagnostique, on vérifie au moins trois fois la connaissance de la même notion, à des endroits différents du test, et en faisant varier la forme des questions, afin de réduire le rôle éventuellement joué par une distraction momentanée, par une réponse correcte trouvée par hasard, etc.

2 Pour un exemple d'application de ce système, voir: BONGRAIN et al., *Epreuves analytiques d'arithmétique* (fin du primaire), Morlanwelz, Institut Supérieur de Pédagogie, 1961.

- La *mémoire à long terme* ou *mémoire active* stocke de façon codée et permanente les concepts et les constructions mentales dérivées de l'expérience passée de la personne (événements, relations, processus, affects, ...) et dirige les opérations de tout le système de traitement de l'information. Cette *base de données* est le cœur du système cognitif. La capacité de cette mémoire est illimitée.
- Le *système musculaire* exécute les actes moteurs: lire, parler, courir, ...

L'extraordinaire performance du cerveau ne peut s'expliquer que par l'existence:

- de procédures tellement rapides qu'elles ne demandent plus beaucoup d'attention (automaticité);
- de stratégies permettant de sélectionner, planifier, piloter et, au besoin, modifier les actions (habiletés métacognitives);
- de structures mentales (schémas) qui mettent les faits et les habiletés en relation. Les schémas sont des structures abstraites qui représentent des ensembles significatifs d'informations stockées en mémoire (par exemple, tout ce qui constitue le déroulement normal d'un match de football).

L'apprentissage consiste en automatisation de procédures, en acquisition et en développement d'habiletés métacognitives, et en construction, révision et remplacement des schémas.

Par rapport aux «novices», les «experts» se caractérisent par le nombre croissant d'interconnexions qu'ils établissent entre un nombre de faits de plus en plus grand, par l'ampleur et la qualité des schémas qu'ils organisent et par une automatisation poussée des procédures, ce qui leur permet de s'attaquer à de nouveaux aspects des problèmes, de se placer dans des perspectives différentes.

Plus techniquement, on parle:

- de *connaissances déclaratives:* connaissances relatives aux propriétés des objets et des événements du monde: faits, concepts, formules; ...
- de *connaissances procédurales:* elles articulent les connaissances déclaratives en procédures et représentent les savoir-faire pratiques. Par le travail sur les connaissances déclaratives (processus métacognitifs) et aux feed-back qui l'accompagnent, des règles se dégagent et des procédures se combinent de mieux en mieux en séquences. Dans les cas favorables, les procédures se combinent de mieux en mieux en séquences et deviennent automatiques; la place est ainsi laissée libre pour l'acquisition de comportements nouveaux, plus élaborés. Par exemple, celui qui, à la simple lecture d'une partition musicale, est capable de l'exécuter automatiquement peut alors consacrer l'essentiel de ses efforts et de son attention à la qualité de l'interprétation;
- de *schémas:* ce sont des structures cognitives qui spécifient les propriétés générales d'un objet ou d'un événement et abandonnent tout aspect contingent. Ils permettent à des objets d'être rattachés à des catégories générales (fonction d'abstraction) et

exercent aussi des fonctions d'anticipation : ils guident la pensée dans des tâches de résolution de problèmes (exemple : prévoir le contenu d'un texte que l'on commence à lire). Les schémas s'organisent et se réorganisent par l'expérience et l'acquisition d'informations nouvelles;

- de *représentations :* ce sont des « théories personnelles », des systèmes de croyances relatifs à des phénomènes. C'est la façon dont un individu ou un groupe les comprend, se les explique. Par exemple, croire que l'électricité coule dans les fils comme l'eau dans les tuyaux, puisque, dans les deux cas, on parle de *courant.* Souvent tenaces, les représentations ne sont pas faciles à débusquer, ni à évaluer. Plus elles sont incorrectes et inadéquatement articulées, plus elles rendent difficiles de nouveaux apprentissages.

Implications pour l'amélioration des tests de performances scolaires.

A la lumière des apports de la psychologie cognitive, H. Walberg[1] estime que l'évaluation scolaire devrait remplir les fonctions suivantes :

- Mesurer la quantité de connaissances déclaratives que possède l'élève et identifier le type d'organisation qu'il utilise pour les stocker.
- Mesurer la vitesse d'exécution des tâches cognitives. Le taux d'attention requis peut être manipulé par l'introduction de tâches concurrentes ou par augmentation de la difficulté des tâches.
- Apporter rapidement un feed-back aux élèves et attirer leur attention sur les erreurs de processus ou de contenu.
- Diagnostiquer la source et le type des erreurs.
- Déterminer, par l'analyse des erreurs, le degré d'adéquation, d'étendue et de flexibilité des schémas.
- Identifier les stratégies de résolution de problèmes; elles sont, en général, spécifiques à un domaine.
- Observer dans quelle mesure l'élève est capable de varier ses stratégies de résolution de problèmes lorsqu'il ne trouve pas directement les solutions.
- Identifier les stratégies métacognitives utilisées par l'élève; repérer celles qui l'aident et celles qui ne l'aident pas. Il importe de distinguer les habiletés métacognitives générales et les habiletés métacognitives spécifiques à un contenu.
- Pour évaluer la qualité des stratégies de résolution de problèmes, prendre pour référence la façon dont les experts perçoivent le problème, élaborent et choisissent un plan de résolution et mettent la stratégie arrêtée en œuvre.

1 H. WALBERG, *The implications of cognitive psychology for measuring school achievement,* Chicago, University of Illinois, 1991, ronéotypé, pp. 19-22.

- Evaluer la qualité des réseaux sémantiques.
- Evaluer le degré d'automaticité des habiletés intellectuelles.

En conclusion :

- Des techniques d'évaluation permettent de connaître les processus cognitifs de l'élève.
- L'utilisation de tests de connaissances à facettes s'impose[1]. En particulier, la façon de questionner doit être variée, de même que le niveau de difficulté. On peut ainsi estimer dans quelle mesure la façon de poser les questions influence les performances observées.
- L'évaluation doit être dynamique. On ne doit pas se contenter d'observer ce que les élèves sont capables de faire par eux-mêmes, à un moment donné. Il faut aussi déterminer jusqu'où ils peuvent aller lorsqu'on les met sur le chemin de la solution (étude du *développement proximal,* selon l'expression de Vygotsky[2]).

Il s'impose, toutefois, de rester prudent à propos de toutes ces propositions. On ne dispose que de peu d'études statistiques à leur sujet. La validité de contenu, la généralisabilité des épreuves doivent être établies. De surcroît, les démarches proposées sont coûteuses.

1 Voir la présentation de la théorie de la généralisabilité.

2 L.S. VYGOTSKY, *Mind in society : the development of higher psychological processes,* Cambridge, Mass., Harvard University Press, 1978.

II. Les objectifs

> «Il est curieux de constater que lorsqu'on demande aux personnes ayant charge d'éduquer les enfants de préciser les fins qu'elles poursuivent, on les plonge souvent dans la perplexité.»
>
> P. Osterrieth, *Faire des adultes*, p. 9.

Affirmer que les maîtres doivent non seulement *instruire*, mais aussi *éduquer* est devenu un truisme. Encore importe-t-il de ne pas se contenter de vœux pieux, de préoccupations vagues. Certes, on enseigne ce que l'on est: notre personnalité, notre façon de penser et d'agir influencent directement, et presque malgré nous, nos élèves.

Mais, si nous voulons systématiser notre action, une définition précise des buts poursuivis devient nécessaire.

Eduquer, c'est mettre en œuvre les moyens propres à développer l'intelligence et la personnalité dans le sens voulu par le milieu culturel; sans oublier que le bonheur, la santé physique de l'élève, exigent aussi l'attention de l'éducateur.

Les tests d'intelligence s'affinent de plus en plus. En particulier, le *Modèle tridimensionnel de l'intellect* proposé par Guilford (voir infra) a beaucoup contribué à la prise de conscience de la multiplicité des aspects de l'intelligence. En outre, les progrès de la psychologie cognitive incitent à une évaluation de plus en plus fine des processus intellectuels qui «expliquent» les résultats scolaires observés.

Instruire signifie mettre quelqu'un en possession de connaissances nouvelles.

Les *objectifs généraux* relèvent de l'éducation, les *objectifs spéciaux*, de l'instruction. Mais les deux sont inséparables; du moins, ils devraient l'être[1].

Evidemment, c'est *avant le début de l'année* scolaire que le professeur doit s'interroger sur les objectifs à atteindre (sinon il agit en aveugle) et *donc aussi définir la matière des examens.*

Certes, les plans d'études et les programmes scolaires officiels apportent une partie des réponses. Mais il est clair qu'elles n'apparaîtront qu'après une étude minutieuse et une longue méditation de ces documents de base. Il n'est d'ailleurs pas rare, surtout dans l'enseignement secondaire, que chaque branche fasse l'objet d'un programme séparé et que seuls les objectifs privilégiés soient explicitement traités.

1 Pour une étude d'ensemble sur ce problème, voir: V. et G. DE LANDSHEERE, *Définir les objectifs de l'éducation*, Paris, P.U.F.; Liège, Dessain, 1992, 7e éd.

Il appartient alors au professeur d'ajouter les objectifs plus généraux qu'il poursuit en commun avec ses collègues des autres disciplines. Si le maître n'a pas une conscience vive des objectifs et ne les a pas faits siens, il a bien peu de chances de les atteindre.

Dresser un tableau des objectifs à poursuivre aide beaucoup le professeur au moment où il prépare sa matière de l'année. Ce tableau (voir exemple p. 66) contiendra autant de colonnes que d'objectifs généraux. Pour chaque chapitre ou chaque étape du programme, on indique les objectifs que l'on souhaite atteindre et l'importance relative à leur réserver (échelle à 3 ou 5 degrés).

L'appréciation de l'importance relative des différents points du programme échappe à la quantification rigoureuse: s'agit-il d'une matière de base qu'il faut nécessairement connaître pour en aborder d'autres? S'agit-il de connaissances ou d'habiletés indispensables pour l'exercice d'une profession ou pour l'insertion dans la vie?

Si subjective l'évaluation de l'importance puisse-t-elle parfois être, l'effort de réflexion qu'elle exige entraîne néanmoins, presque toujours, une clarification utile.

A chaque occasion, les maîtres se poseront une question plus grave encore: «Les objectifs qu'ils s'apprêtent à poursuivre ne sont-ils pas en contradiction avec le projet éducatif, ne trahissent-ils pas les buts fondamentaux, les fins de l'éducation?» Toute contradiction entre la philosophie de l'éducation, spécialement les valeurs épousées, et l'action pédagogique doit être évitée.

Enfin, les considérations qui précèdent pourraient faire croire que les objectifs sont arrêtés une fois pour toutes, avant d'engager l'enseignement. Il ne devrait pas en être ainsi. D'une part, parce que, chemin faisant, des objectifs d'une importance au moins égale à ceux que l'on voulait privilégier peuvent apparaître. D'autre part, les élèves doivent, le plus tôt possible, être associés au choix des objectifs, lors d'une négociation, plus ou moins poussée selon le niveau de développement et, l'âge venant, selon le projet éducatif de chacun.

A. LES OBJECTIFS GENERAUX.

Au fond, l'objectif de l'enseignement est unique: aider l'élève à se développer harmonieusement et à accéder à l'état adulte dans les meilleures conditions.

Plus spécialement, les maîtres doivent, d'une part, au moins ne pas nuire à la santé physique et mentale de leurs élèves et, autant que possible, aider à la développer, et, d'autre part, poursuivre systématiquement des objectifs cognitifs et affectifs que nous allons essayer de préciser.

Séparer les domaines cognitif (penser), affectif (être satisfait ou irrité ; aimer ou rejeter) et conatif (vouloir, désirer) est théorique. Peut-on penser sans éprouver de sentiment, agir de façon responsable sans penser ? On le sait bien, c'est toujours l'organisme total qui répond à une stimulation.

Mais ne perdons pas notre propos de vue : nous sommes ici à la recherche de points de repère qui serviront à jalonner enseignement et examens. C'est pourquoi nous allons couper arbitrairement le domaine cognitif du domaine affectif et donner à ce dernier mot une acception très large.

1. Les objectifs cognitifs.

On doit à deux chercheurs américains, B. Bloom[1] et J.-P. Guilford[2], des classifications hiérarchisées des objectifs cognitifs qui, malgré certaines faiblesses[3], se révèlent d'utiles outils.

a) La taxonomie de Bloom.

En voici d'abord les grandes articulations ; un exemple d'application pratique est proposé pages 84-86.

1. Connaître de mémoire.
2. Comprendre.
3. Appliquer.
4. Analyser.
5. Synthétiser.
6. Evaluer.

1 B. BLOOM et al., *Taxonomie des objectifs pédagogiques. I. Domaine cognitif*, traduit par M. Lavallée, Montréal, Education Nouvelle, 1969.

2 J.-P. GUILFORD, *Modèle tridimensionnel de l'intellect.* Voir *The Nature of Human Intelligence*, New York, McGraw-Hill, 1967.

3 Voir spécialement la critique de J. CARROLL.

Appliquer, par exemple, appelle une analyse plus ou moins fine. Ne faudrait-il donc pas placer l'analyse avant l'application ? Bloom et ses collaborateurs ont rétréci le sens d'*application* pour échapper à cette difficulté. Ils indiquent d'ailleurs à plusieurs endroits de leur ouvrage les recouvrements entre certains échelons et analysent, avec beaucoup de finesse, les difficultés d'utilisation de la *Taxonomie*. Notre but n'est pas d'en restituer toutes les nuances, mais bien d'attirer l'attention sur l'instrument.

Chacun des échelons doit être accepté dans le sens retenu par les auteurs de la taxonomie. Par ailleurs, même si connaître de mémoire et évaluer apparaissent comme les comportements cognitifs, le moins et le plus nobles, cela n'implique en rien que la mémorisation doive disparaître de notre enseignement. Théoriquement au moins, il faut avoir franchi chaque échelon pour accéder au niveau supérieur. La taxonomie a été créée pour *aider l'éducateur à ne pas oublier* certaines étapes et pour l'inciter à élever graduellement le niveau de son enseignement. La voici d'abord telle que Bloom la présente.

1. CONNAITRE DE MEMOIRE.

   1.1. Des données particulières : appellations, faits, dates, symboles.

   1.2. Des façons de traiter des données particulières (sans les appliquer) : conventions, classifications, critères, méthodes.

   1.3. Des données universelles : principes, lois, théories, ...

La différence entre les trois sous-catégories est plus quantitative que qualitative : savoir par cœur, sans plus, les dates de naissance des rois de France ou tel exposé de la philosophie de Kant nécessite surtout des efforts de mémorisation différents. Le niveau de pensée reste, dans les deux cas, très bas. Or, combien de questions d'examen, du primaire à l'enseignement supérieur, ne se situent-elles pas à ce niveau ? (Que savez-vous de ... ? Quelles sont les clauses de ... ? Comment prépare-t-on le ... ?).

2. COMPRENDRE.

Il s'agit ici de la compréhension au niveau le plus bas. Par une formulation nouvelle du donné, l'individu montre qu'il a dépassé le psittacisme, que le message a, pour lui, une signification. Toutefois, il n'est pas encore question d'appliquer, donc de percevoir les rapports du donné avec d'autres matériaux, d'autres situations.

Bloom distingue deux échelons :

2.1. *Traduire, transposer.*

Le contenu de la communication est conservé sans que son ordre soit modifié, mais la forme est changée.

Exemples : Paraphraser un récit, une proposition :
« Une *taxonomie* est une *classification* ».
Exprimer verbalement des symboles mathématiques :
A > B signifie que A est plus grand que B.

2.2. *Interpréter.*

Expliquer ou résumer une communication.
L'interprétation implique un nouvel arrangement, une nouvelle vue du matériel. Elle suppose donc la capacité de reconnaître et de saisir les idées maîtresses d'une communication et de comprendre les rapports existant entre elles. « A cet égard, l'interprétation devient synonyme de l'analyse et possède certaines caractéristiques de l'évaluation[1]. »

1 B. BLOOM et al., *o.c.*, p. 104.

Exemple :
Interpréter des données représentées sous forme de tableaux ou de graphiques en tirant des déductions tenant compte des relations entre données, ou de leur signification d'ensemble.
Aller au-delà des données et des renseignements fournis : extension des tendances, généralisation.

2.3. *Extrapoler.*

3. APPLIQUER.

L'application suppose que le sujet distingue les traits communs à deux situations, à deux problèmes; une abstraction s'est donc produite.

Des principes ou des généralisations sont appliqués à des problèmes nouveaux.

Exemple : emploi de procédés expérimentaux pour résoudre des problèmes de travaux ménagers.

4. ANALYSER.

4.1. Rechercher des éléments.

Exemple : distinguer les faits des hypothèses dans une communication.

4.2. Rechercher des relations.

Exemple : les hypothèses sont-elles logiques par rapport aux renseignements dont on dispose ?

4.3. Rechercher des principes d'organisation.

Exemple : identifier les techniques de propagande utilisées dans des tracts.

5. SYNTHETISER.

5.1. Produire une œuvre personnelle.

Exemple : narration captivante d'une expérience vécue.

5.2. Elaborer un plan d'action répondant aux exigences fixées.

5.3. Dériver un ensemble de relations abstraites. Induire une règle.

6. EVALUER.

Des jugements qualitatifs ou quantitatifs établissent dans quelle mesure le matériel et les méthodes répondent aux critères (internes ou externes).

Exemples : Déceler les sophismes dans une discussion. Apprécier un travail par comparaison à un modèle.

L'expérience montre que la taxonomie telle qu'elle est ainsi présentée soulève de nombreux problèmes d'interprétation quand on tente de l'utiliser dans la pratique.

Tenant compte des clarifications apportées par différentes recherches[1], nous l'avons reformulée de la façon suivante :

1 Voir V. et G. DE LANDSHEERE, *Définir les objectifs..., o.c.*

1. *Connaissance*

Simple restitution de mémoire.
Exemple : comment s'appelait le troisième président des Etats-Unis ?

2. *Compréhension*

Montrer par la réponse fournie que l'on sait accomplir une tâche pour laquelle *toutes les données nécessaires figurent dans l'énoncé du problème.*

Exemple : dette publique des Etats-Unis.

Année	Dette nationale totale (en dollars)	Dette par tête d'habitant (en dollars)
1915	1 101 264 068	11,85
1920	24 299 321 467	228,23
1925	20 516 193 888	167,12
1935	28 700 892 625	225,55
1940	42 967 531 038	325,59

La colonne «dette par tête d'habitant» indique l'argent que chaque personne vivant aux Etats-Unis aurait dû si la dette nationale avait été divisée également entre tous.

Voici deux propositions relatives au tableau ci-dessus. Indiquez, pour chaque proposition, si vous pensez qu'elle est :
1. juste,
2. probablement juste,
3. si les données ne sont pas suffisantes pour que vous puissiez vous prononcer,
4. probablement fausse,
5. fausse.

a) En 1940, la dette par tête d'habitant aux Etats-Unis était approximativement deux fois plus grande qu'en 1925.

b) La dette nationale totale était plus grande en 1916 qu'en 1911.

3. Application

L'élève doit utiliser un modèle général de solution, appris antérieurement, pour résoudre un problème concret, particulier. Toutes les données nécessaires à la résolution ne se trouvent donc pas dans l'énoncé du problème; l'élève doit apporter les informations supplémentaires nécessaires.

Exemple : calculez la surface d'un triangle dont la base mesure 20 cm et la hauteur 15 cm.

4. Analyse

L'élève doit découvrir les composantes d'une situation ou d'un document, les moyens employés par un auteur pour arriver au résultat (texte, objet, ...) que l'on observe. En particulier, connaissant des conditions ou des critères, l'élève doit découvrir s'ils sont ou non réunis dans l'objet de l'observation. Il n'existe qu'une réponse possible au problème ainsi posé.

Exemples :

1. Distinguez, dans le texte suivant, les propositions factuelles et les propositions normatives.
2. Un bac à fleurs est placé devant une fenêtre exposée au sud. Toutes les plantes du bac se penchent vers la fenêtre.
   Indiquez, pour chacune des propositions suivantes, si
   A. elle aide à expliquer la cause du phénomène;
   B. elle décrit seulement le phénomène;
   C. elle décrit une conséquence du phénomène;
   D. elle ne concerne pas directement le phénomène.

   1) La division cellulaire se fait plus vite à l'ombre.

   2) La vitesse de photosynthèse est plus grande du côté exposé au sud.

   3) Les plantes présentent un phototropisme positif.

Evaluation

Il s'agit d'une analyse, mais il existe plusieurs réponses au problème parce que les critères ne sont pas des faits ou des règles univoques, mais des croyances, des valeurs personnelles.

Exemples :

1. Lequel des trois dessins suivants trouvez-vous le plus beau ?
2. Voici une courte biographie d'un personnage célèbre. Dressez la liste de ses comportements que vous trouvez immoraux.

5. *Synthèse - Créativité*

La synthèse consiste à disposer et à combiner des éléments afin de former un plan ou une structure que l'on ne distinguait pas clairement auparavant. La synthèse implique nécessairement la production de comportements personnels originaux. Plusieurs solutions sont toujours possibles.

Exemples :

1. Trouvez un titre qui convient à l'histoire suivante.
2. Quelles qualités devrait posséder un bon journal ?

b) Le modèle de Guilford.

Moins utilisé, jusqu'à présent, que la taxonomie de Bloom pour la construction des examens, le modèle de Guilford offre peut-être de plus grandes possibilités encore, en raison de sa rigueur.

Voici d'abord comment J.-P. Guilford et R. Marrifiels définissent les trois dimensions de l'intellect et leurs composantes[1] :

LES OPERATIONS.

Ce sont les activités ou les processus intellectuels principaux; c'est ce que fait l'organisme à partir de la matière première informationnelle, à partir de ce qu'il discrimine.

1 J.-P. GUILFORD, *The Nature of Human Intelligence*, New York, McGraw-Hill, 1967.

1. Cognition.
   Conscience, appréhension, découverte ou redécouverte, reconnaissance, compréhension d'informations sous diverses formes.
2. Mémoire.
   Rétention d'informations.
3. Production convergente.
   Génération d'informations uniques, conventionnellement acceptées, à partir d'un donné. L'usage, la coutume, la règle sont respectés.
4. Production divergente.
   Génération d'informations variées à partir d'un même donné. Originalité, créativité.
5. Evaluation.
   Prise de décisions ou formulation de jugements concernant l'exactitude, l'adéquation, la désirabilité, ... conformément à des critères, à des idéaux, à des objectifs adoptés.

LES CONTENUS.

1. Figuratifs.
   Information dans sa forme concrète, perçue ou rappelée en images.
   Un minimum d'organisation, de structuration est nécessaire. Intelligence pratique.
2. Symboliques.
   Informations sous forme de signes dépourvus de signification par et en eux-mêmes : lettres, nombres, notes de musique.
   Intelligence théorique.
3. Sémantiques.
   Informations sous forme de significations attachées à des mots.
   Intelligence verbale.
4. Comportementaux.
   Informations, essentiellement non verbales, intervenant dans les interactions humaines, où la perception d'attitudes, de besoins, de désirs, d'intentions, de pensées d'autrui et de soi-même jouent un rôle.
   Intelligence sociale.

LES PRODUITS.

Ce sont les résultats du traitement des informations par l'organisme.

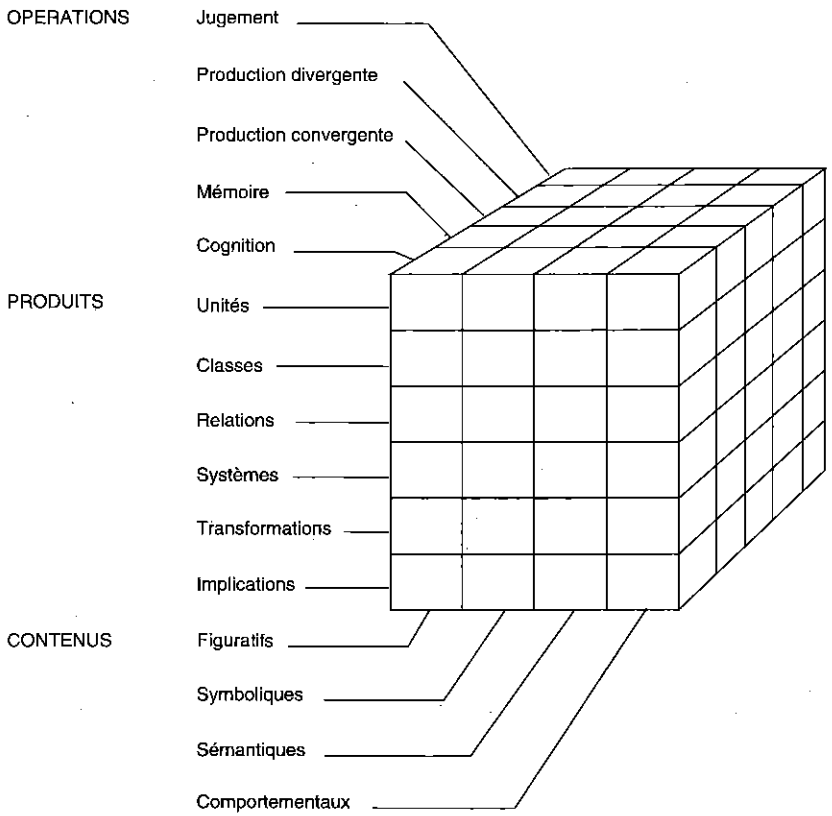
1. Unités.
   Portions d'information relativement isolées ou circonscrites.
2. Classes.
   Unités groupées en raison de leurs propriétés communes.
3. Relations.
   Connexions reconnues entre des unités.
4. Systèmes.
   Groupements d'unités organisées ou structurées; complexes de parties se trouvant en interrelation ou en interaction.
5. Transformations.
   Changements apportés dans des informations ou dans leur utilisation.
6. Implications.
   Extrapolation d'informations : prédiction, conséquences, antécédents.

Chacune des composantes des trois dimensions se combine avec toutes les autres. Ainsi:
Mémoire des

→ { unités, classes, relations, systèmes, transformations, implications } → { figuratifs, symboliques, sémantiques, comportementaux }

soit 24 combinaisons.

Comme on distingue cinq types d'opérations, il existe donc, en tout, cent vingt combinaisons dans le modèle.



L'expérience montre qu'il est difficile de traduire toutes les combinaisons en termes utilisables pour l'enseignement.

Dans un premier temps, les quatre types d'opérations retiendront surtout l'attention. Idéalement, et dès le début de la scolarité, le maître devrait s'efforcer de les introduire dans toutes les activités.

1. La restitution de mémoire témoigne de la fixation de la notion.
2. La production convergente témoigne de la capacité d'appliquer les acquisitions en conformité avec les règles, les usages.
3. La production divergente témoigne de la capacité de découvrir des solutions ou des usages nouveaux.
4. Enfin, l'évaluation témoigne de la capacité de critiquer, de porter un jugement en fonction de critères bien définis et d'agir en conséquence.

Trop souvent, l'activité scolaire et les examens qui la sanctionnent se limitent aux deux premières opérations.

S'efforcer de garder un équilibre entre les quatre catégories des contenus importe autant dans l'enseignement général que dans l'enseignement technique ou artistique.

2. **Les objectifs affectifs.**

En schématisant l'analyse fine de P. Osterrieth[1], nous dirons qu'un individu accède pleinement à la condition d'adulte:

1° a) si son comportement a trouvé sa logique, sa cohérence et échappe à la versatilité;
   b) s'il a, par ailleurs, acquis une saine tolérance au changement, à la contradiction, à l'échec;

2° a) s'il a conquis son indépendance, son autonomie intellectuelle et affective;
   b) s'il est cependant capable de faire le don de soi, de rester fidèle à ses engagements et à ses sentiments.

L'éducation est un long acheminement vers cet ultime équilibre.

Nous proposons ci-dessous une adaptation et une interprétation de la taxonomie des objectifs de D. Kratewohl et B. Bloom[2]. On obtient de nouveau cinq échelons, cinq étapes qui acheminent du comportement le plus passif au plus actif.

1 P. OSTERRIETH, *Faire des adultes*, Bruxelles, Dessart, 1964.
2 D.R. KRATEWOHL, B.S. BLOOM, B. MASIA, *Taxonomie des objectifs pédagogiques. Domaine affectif*, Montréal, Education nouvelle, 1970.

a) **L'individu répond à une stimulation extérieure.**

1. *Est simplement réceptif.*

C'est une sorte d'état affectif amorphe où le sujet perçoit la beauté ou la laideur, les sentiments divers, sans réagir, un peu comme un miroir qui ne renverrait pas l'image.

Ce comportement est d'ailleurs difficile à distinguer de la simple cognition qui précède la mise en mémoire. Seul un certain éveil de l'attention est observable. Exemple : écoute la musique, écoute parler les autres.

2. *Reçoit et réagit.*

L'individu réagit nettement soit en obéissant, soit en manifestant du plaisir, par la parole, par le geste ou l'attitude. A ce stade, on n'observe pas encore de rejet explicite qui témoignerait d'un choix délibéré.

Pour le professeur de littérature, c'est le moment où les élèves n'ont pas encore le goût assez formé pour faire un choix personnel, où leur sensibilité n'est pas encore assez raffinée pour leur permettre de partir seuls à la découverte, mais où, mis en présence de belles œuvres, ils commencent à en sentir la grandeur.

3. *Reçoit et réagit en acceptant ou en refusant.*

Maintenant, l'individu sait ce qu'il veut ou ce qu'il aime, à condition d'être mis en présence des personnes ou des choses; il s'engage.

b) **L'individu prend l'initiative.**

4. *Essaie spontanément de comprendre, de juger, de ressentir.*

L'individu éprouve assez d'intérêt, de curiosité, pour s'instruire sans y être invité, assez de sensibilité pour prendre une initiative sentimentale ou, encore, il a suffisamment découvert le sens des valeurs pour se choisir une philosophie ou une religion.

5. *Agit selon ses options.*

C'est le stade psychologiquement adulte, comme l'a défini P. Osterrieth.

Par exemple, l'individu vit en fonction de ses options morales, sentimentales, esthétiques, mais il est aussi capable de changer de conduite à la lumière de preuves, d'arguments convaincants.

Cette ultime étape de l'ascension affective correspond à l'évaluation dans le domaine cognitif.

B. LES OBJECTIFS SPECIAUX.

Théoriquement, toutes les matières, tous les points des programmes scolaires offrent l'occasion de se rapprocher des objectifs généraux et de vérifier s'ils ont été atteints.

Trois remarques s'imposent toutefois :

1° la pratique montre qu'une même épreuve ne peut porter que sur un nombre limité d'objectifs;

2° certaines matières se prêtent mieux que d'autres à la poursuite de certains objectifs;

3° le grave problème du transfert des apprentissages continue à se poser avec acuité. Par exemple, cultiver la divergence à l'occasion des activités artistiques ne garantit en rien que ce trait se manifestera dans les domaines scientifiques ou dans la vie pratique.

Quoi qu'il en soit, la première démarche visant à l'identification des objectifs spéciaux reste fondamentalement toujours la même et consiste en l'élaboration d'un tableau à double entrée. En haut, on porte les objectifs généraux; sur le côté, on inscrit les matières du cours. Chaque intersection de colonne et de rangée indique, en principe, un objectif spécial.

Exemple : CHIMIE[1].

	Connaissance des faits, méthodes et techniques	Application	Evaluation
Equilibres ioniques[1]			
1. Généralités			
1. Degré d'ionisation ou fraction ionisée			
2. Mesures du degré d'ionisation			
3. Application de la loi d'action des masses aux équilibres ioniques			
4. Loi de la dilution d'Ostwald			
2. Produit de solubilité			
1. Définition			
2. Calcul du produit de solubilité en fonction de la solubilité			
3. Diminution de la solubilité d'un électrolyte peu soluble			
4. Précipitation d'un électrolyte par adjonction, à sa solution saturée, d'un électrolyte apportant un ion commun			

1 Extrait de : BERGER et DIGHAYE, *Chimie IV,* Liège, Sciences et Lettres, 1967.

	Connaissance des faits, méthodes et techniques	Application	Evaluation
5. Dissolution des précipités			
Appendice: Application du produit de solubilité à la précipitation des sulfures			
3. Produit ionique de l'eau et pH			
1. Produit ionique de l'eau			
2. Le pH			
a La notation pH			
b pH de l'eau pure			
c L'échelle des pH			
d Détermination du pH			
e Les indicateurs colorés			
f Méthodes expérimentales de détermination du pH			
g Distinction entre électrolytes (acides et bases) forts et faibles, par détermination du pH de solutions diluées de concentrations connues			
h Calcul du pH des solutions d'acides et de bases à partir de leurs concentrations molaires et - éventuellement - de leurs constantes d'ionisation			
4. Solutions tamponnées			
1. Introduction expérimentale			
2. Cas des sels dérivant d'un acide fort et d'une base forte			
3. Cas des sels dérivant d'un acide fort et d'une base faible			
4. Cas des sels dérivant d'un acide faible et d'une base forte			
5. Cas des sels dérivant d'un acide faible et d'une base faible			
6. Facteurs influençant l'hydrolyse			
6. Neutralisation mutuelle des acides et des bases			
1. Chaleur de neutralisation			
2. Courbes de neutralisation			
a Acide fort - base forte			
b Acide faible - base forte			
c Base faible - acide fort			
d Base faible - acide faible			
7. Les méthodes d'analyse quantitative par voie chimique			
1. Méthodes gravimétriques			
2. Méthodes volumétriques			
Acides - bases			
Choix des indicateurs			
Titrages rédox			

La matière de chimie figurant dans le tableau ci-dessus constitue un des quatre chapitres d'un cours destiné à l'année supérieure de l'enseignement secondaire. Parmi bien d'autres objectifs généraux possibles, nous n'en avons retenu que trois.

Un simple coup d'œil sur l'ensemble révèle que, même dans ces conditions très simplifiées, le professeur se trouve devant un choix difficile. Sur quels points va-t-il faire porter l'examen pour obtenir un échantillonnage suffisant de la matière? Quelle importance relative va-t-il réserver aux matières retenues et aux divers objectifs spéciaux qui y correspondent? En quoi consiste exactement l'évaluation dans un cours comme celui-ci?

On le voit, le problème est loin d'être simple. Il dépasse presque toujours la compétence d'un seul homme et débouche sur des questions restées sans réponse. Il est souhaitable de créer des commissions de définition d'objectifs où enseignants, psychologues et spécialistes de la recherche en éducation unissent leurs efforts.

Voici enfin quelques questions se rapportant à trois des *productions* telles qu'elles sont définies dans le modèle de Guilford.

Branche	*Production convergente*	*Production divergente*	*Evaluation*
Sciences	Expliquez pourquoi il ne peut pas y avoir de vie sur Mercure?	En quoi la vie sur Mars pourrait-elle différer de la nôtre?	Pensez-vous qu'il existe une vie sur Mars?
Géographie	En quoi le détroit de Bering a-t-il influencé le peuplement de l'Amérique du Nord?	Que serait-il arrivé si le détroit de Bering n'avait pas existé?	Actuellement, quel est, à votre avis, l'importance stratégique et commerciale du détroit de Bering?
Histoire	Expliquez l'importance de la découverte de l'Amérique par C. Colomb pour la vie économique de l'Europe.	Que serait-il arrivé si C. Colomb avait découvert la route des Indes au lieu de l'Amérique?	Quelles sont, à votre avis, les deux conséquences les plus importantes du voyage de C. Colomb?
Langue maternelle	Expliquez pourquoi la nouvelle s'est beaucoup plus développée aux Etats-Unis qu'en Europe?	Voici le début d'une nouvelle. Imaginez autant de dénouements que vous le pouvez.	Qu'est-ce qui importe le plus dans une nouvelle: les caractères ou l'intrigue?

1 D'après J.-R. VERDUIN Jr., *Conceptual Models in Teacher Education*, Washington, A.A.C.T.C., 1967, p. 93.

C. LES OBJECTIFS OPERATIONNELS[1]

Comme nous le verrons par la suite, le recours aux taxonomies n'exclut pas la subjectivité. La preuve en est que, pour chaque intersection du tableau matières-objectifs (p. 91), divers examinateurs travaillant isolément formuleront des questions parfois fort différentes. Il reste d'ailleurs à prouver qu'une question déterminée met bien en œuvre des comportements commandés par l'objectif.

Le problème est déjà considérablement clarifié si les objectifs sont directement exprimés en termes de comportements observables.

Comparez:

- L'élève connaîtra le nom des rois des Belges qui se sont succédé depuis 1830 (objectif non comportemental).

  L'élève nommera correctement les rois des Belges... (objectif comportemental).

- L'élève acquerra une connaissance active du magnétophone (objectif non comportemental). Disposant d'un magnétophone et d'une bobine d'enregistrement vierge, l'élève enregistrera sa propre voix (objectif comportemental).

Dans les deux cas ci-dessus, le verbe «connaître» peut être diversement interprété. Par contre, «nommer» ou «enregistrer» sont dépourvus d'ambiguïté.

On remarquera en outre qu'ainsi formulés, les objectifs portent en eux-mêmes la question qui permettra de vérifier s'ils sont atteints.

Sous l'impulsion particulière de W. Popham[2] et, plus généralement, des spécialistes de la technologie de l'enseignement, les efforts de clarification au niveau des micro-objectifs ont été considérables ces derniers temps.

Les micro-objectifs ne suffisent toutefois pas à constituer une unité d'enseignement; si étroitement limitée, elle deviendrait artificielle. Par contre, vérifier si une batterie de micro-objectifs relatifs à un même objet sont atteints offre une démarche diagnostique de premier choix.

1 Selon H. PIERON, la définition opérationnelle d'un comportement, d'un facteur, etc., est «l'énoncé des procédures qui permettent de le mesurer, de le produire ou simplement de le reconnaître parmi d'autres». (*Vocabulaire de la psychologie*, Paris, P.U.F., 1968.)
Nous ne nous attachons ici qu'au caractère *observable* des comportements sans traiter de leur mesure.

2 W. POPHAM, *Objectives and Instruction*, Chicago, Rand Mc Nally, 1967.

Voici deux exemples.

1. Biologie - enseignement primaire.

Liste publiée par la «Bourse d'objectifs» animée par Popham (extrait)[1].

Catégories principales.

I. L'organisme individuel - 7 principes - 46 objectifs spécifiques.
II. La population - 5 principes - 25 objectifs spécifiques.
III. L'environnement - 2 principes - 10 objectifs spécifiques.
IV. La communauté - 3 principes - 11 objectifs spécifiques.
V. L'écosystème - 4 principes - 16 objectifs spécifiques.

Division de la catégorie I: L'organisme individuel.

Principe A: Les objets sont classables en vivants et non vivants.

Objectif spécifique A1:
Dans une série de dessins, l'élève saura distinguer les objets vivants.

Objectif spécifique A2:
On présente le dessin d'une semence non germée. L'élève saura expliquer par écrit ou oralement comment on peut savoir si cette semence est morte ou vivante.

Objectif spécifique A3:
L'élève saura distinguer les plantes et les animaux dans une série de dessins.

Objectif spécifique A4:
Dans une série de silhouettes d'animaux, l'élève saura indiquer les animaux qui possèdent des vertèbres; il saura dessiner la colonne vertébrale par une ligne sur la silhouette.

Etc.

Principe B:

Pour se procurer de l'énergie et remplir des fonctions nécessaires à la vie, les organismes recherchent et utilisent une variété de substances nourrissantes. L'énergie extraite de ces substances est stockée dans l'organisme puis est utilisée pour la croissance, le mouvement et la reproduction.

Objectif spécifique B1:
Savoir reconnaître, sur un dessin, les sources d'énergie qui conviennent aux animaux qui s'y trouvent.

Etc.

1 La «Bourse d'objectifs» (*Instructional Objectives Exchange*) a été créée en 1969 par l'Université de Californie à Los Angeles (U.C.L.A.). Ses buts sont:
- servir de clearinghouse permettant à toutes les écoles du pays d'échanger leurs objectifs d'enseignement; les efforts des éducateurs sont ainsi conjugués et non dispersés;
- réunir et construire des techniques de mesure permettant de déterminer si les objectifs sont atteints;
- formuler de façon appropriée des objectifs d'enseignement dans des secteurs importants pour lesquels on n'en dispose pas encore; il s'agit donc de combler des lacunes existantes.

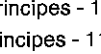
2. Enseignement de la lecture et de l'écriture[1].

Voici une des vingt-quatre batteries d'objectifs relatifs à l'enseignement élémentaire de la lecture.

OBJECTIF D.

L'ELEVE DOIT POUVOIR DISCRIMINER LES LETTRES ENTRE ELLES.

D_1 Etant donné un ensemble de lettres différant par la taille et le dessin, l'élève doit pouvoir identifier celles qui sont les mêmes.

Ex.: A  B

D_2 Etant donné différentes lettres qui ont un élément commun, l'élève doit pouvoir identifier ce qui les différencie.
Ex.: d/b; g/j; m/n ...

D_3 Etant donné un chiffre ou une lettre modèles, l'élève doit pouvoir repérer dans une série de chiffres ou de lettres l'élément identique au modèle donné.
Ex.: b: d b p q.

D_4 Etant donné une lettre ou un groupe de lettres écrites, l'élève doit pouvoir souligner dans une phrase donnée toutes les lettres identiques à la lettre ou aux lettres proposées.
Ex.: ch: Charles cherche le chat.

D_5 Etant donné une liste de mots contenant diverses lettres initiales ou finales, l'étudiant devra classer les mots avec initiale ou finale identiques.

D_6 Etant donné une séquence de lettres suivie par une série de séquences de lettres, l'élève doit pouvoir repérer la séquence qui est identique à la séquence donnée comme modèle.
Ex.: par: por pra rap par.

La formulation complète d'un objectif opérationnel comprend cinq indications précises:

1. Qui produira le comportement souhaité.
2. Quel comportement observable démontrera que l'objectif est atteint.
3. Quel sera le produit de ce comportement (performance).
4. Dans quelles conditions le comportement doit avoir lieu.
5. Quels critères serviront à déterminer si le produit est satisfaisant.

1 W. POPHAM, *Language Arts: Decoding Skills*, Los Angeles, Instructional Objectives Exchange, 1972.

Exemple :

1. l'élève
2. saura construire
3. un poste de radio à transistors
4. en choisissant lui-même les pièces au magasin, en se référant au schéma adopté.
5. L'appareil devra capter correctement des émissions d'au moins cinq émetteurs différents sur ondes moyennes et de cinq émetteurs sur ondes longues.

R.F. Mager concentre ces exigences en trois points : « Pour décrire le comportement final (ce que l'élève fera) :

1. Identifiez et nommez le comportement.
2. Définissez les conditions dans lesquelles le comportement doit se produire (ce qui est donné ; quelles sont les restrictions ou, à la fois, le donné et les restrictions).
3. Définissez les critères de la performance acceptable[1]. »

D. L'ENSEIGNEMENT PAR OBJECTIFS : UNE MISE EN GARDE

Ce qui vient d'être dit à propos des objectifs opérationnels concerne l'évaluation, et plus spécifiquement l'évaluation diagnostique, et non l'enseignement.

La distinction de différents domaines et niveaux taxonomiques pourrait susciter la résurgence de la gymnastique intellectuelle de l'époque où l'on ambitionnait d'entraîner séparément la mémoire, l'attention, ... Or beaucoup de nos actes, en apparence les plus simples, sont d'une complexité considérable. Par exemple, au moment où il s'apprête à lancer la boule, le joueur de pétanque doit *se souvenir* des règles du jeu, *analyser* la situation, *synthétiser* ses observations pour prendre une décision, *évaluer* ses chances de réussite. Des *comportements psychomoteurs* de divers niveaux vont aussi intervenir, et *l'affectivité* n'est pas absente de la situation. Tout cela se passe généralement en moins d'une minute !

1 R.F. MAGER, *Preparing Instructional Objectives*, Palo Alto, Fearon, 1962, p. 53.

De façon plus générale, un grave malentendu s'est produit à propos de l'enseignement par objectifs. D'aucuns ont cru voir dans leur définition rigoureuse le point de départ et d'arrivée d'une méthodologie rationnelle de l'enseignement. On a vu paraître, peu après la publication des taxonomies de B.S. Bloom et de ses associés, et du petit ouvrage que R.F. Mager a consacré à la définition opérationnelle des objectifs[1], des listes d'objectifs « terminaux » ou « comportementaux » couvrant tout un programme d'enseignement. De là à imaginer qu'il suffirait d'enseigner dans l'ordre de ces listes bien ordonnées pour être assuré du succès, il n'y avait qu'un pas que certains ont semblé prêts à franchir...

On a vu ainsi la vieille pédagogie impositive, fondée sur la structuration logique des savoirs achevés et sur le principe du pas à pas militairement ordonné, trouver un souffle nouveau avec, en plus, l'étiquette de la modernité.

D'où la réaction aussi aveugle que l'action. La définition des objectifs de l'éducation, identifiée à une technicité desséchée, fut déclarée incompatible avec la pédagogie fonctionnelle. Il n'en est évidemment rien. En réalité, les objectifs diffèrent profondément selon que l'on veut transmettre un savoir préfabriqué ou susciter la découverte, la construction active des savoirs, des savoir-faire et des savoir-être.

1 R.F. MAGER, *Comment définir les objectifs pédagogiques ?*, Paris, Gauthier-Villars, 1972.

CHAPITRE 2

LA REDACTION DES QUESTIONS

I. Observations générales

Pour être équitable et valide, un examen doit presque toujours comporter un grand nombre de questions. Si le domaine à couvrir est étendu, la chose semble évidente. N'interroger que sur une fraction, parfois minime, de la matière repose sur une exigence difficilement justifiable, surtout dans notre système d'enseignement actuel, non individualisé[1] : tous les élèves sont censés avoir tout appris et la qualité de tous leurs apprentissages est supposée homogène. Des questions portant arbitrairement sur une partie du tout permettraient donc de porter un jugement valable.

En réalité, si, par malchance, le professeur n'interroge que sur une des rares parties qu'un étudiant n'a pas étudiée, c'est la catastrophe. Fréquemment aussi, une simple distraction, une fatigue passagère expliquent une mauvaise réponse dans un domaine pourtant bien connu.

Nous avons déjà vu que, pour prévenir ce danger, les testeurs essaient, en général, de poser au moins trois fois la même question, sous des formes différentes, ce qui n'est certes guère l'usage dans les examens traditionnels.

Par ailleurs, la multiplicité des objectifs à poursuivre, ainsi que les taxonomies nous l'ont révélé, rend difficilement concevable un petit nombre de questions.

Pourtant, il serait irréaliste de préconiser une construction savante pour le moindre exercice de contrôle ou de diagnostic.

Savoir poser des questions est probablement la capacité la plus nécessaire au professeur. Mais c'est un art difficile, les erreurs dans le choix du niveau de langage et les obscurités fréquemment rencontrées dans le libellé des exercices d'application et les questions d'examen en témoignent.

1 Voir 5e partie : Une pédagogie de la maîtrise.

Il n'existe évidemment pas de recette universelle pour la rédaction des questions. Quelques règles, proposées par R. Thorndike[1], peuvent néanmoins servir d'introduction :

1. Avant de commencer à rédiger une question, ayez clairement à l'esprit quel processus mental vous souhaitez que l'élève utilise pour répondre.
2. Utilisez des matières nouvelles ou une présentation nouvelle des matières dans les questions.
3. Commencez les questions par les mots ou les expressions suivants : « Comparez - Opposez - Donnez les raisons de... - Expliquez comment... - Critiquez - Dites ce qui arriverait si... ». Ne commencez pas vos questions par des mots comme : « Quoi ? Qui ? Quand ? Citez... ».
4. Ecrivez les questions de façon qu'elles soient claires et précises pour chaque élève.
5. Une question portant sur une matière controversée doit demander des arguments en faveur d'une position plutôt qu'une simple prise de position.
6. Assurez-vous que la question appelle un comportement que vous souhaitez réellement voir produit par l'élève.
7. Adaptez la longueur et la complexité des questions au niveau de maturité des élèves.

A. Des questions compréhensibles.

Bien des réponses erronées ne sont pas dues à l'ignorance de la matière, mais bien à des malentendus, à une mauvaise compréhension des questions. Le danger apparaît spécialement en un domaine comme les mathématiques parce que deux difficultés s'y combinent aisément : la difficulté inhérente au problème et la difficulté du langage abstrait des mathématiques.

Mais, on le sait, l'abstrus n'épargne pas non plus les historiens, les géographes, les professeurs de sciences... ou de langues.

En ne s'assurant pas de la clarté des questions, les maîtres risquent, en particulier, de commettre une injustice sociale. La recherche contemporaine confirme que beaucoup d'enfants issus de milieux socialement

1 D'après R. THORNDIKE et E. HAGEN, *Measurement and Evaluation in Psychology and Education*, Londres, Wiley, 1969, 3e éd.

défavorisés souffrent de handicaps graves dans le domaine du langage. Davis et Haggard ont, par exemple, montré qu'il suffit de modifier la forme d'un problème, sans en changer le sens, pour que la différence de réussite entre enfants provenant de milieux socio-économiques favorisés ou non passe de 12 à 32 %[1].

B. **Tenir compte du niveau d'information.**

Pour vérifier la capacité à transférer les apprentissages, il est souvent nécessaire de situer les problèmes dans un contexte non encore évoqué en classe. Encore faut-il que ce contexte ait un sens pour l'élève.

Tels petits problèmes, récemment rencontrés dans un manuel, supposaient la connaissance des règles du jeu de tennis ou de la manière de remplir des bulletins de pronostic de football. Que signifieront ici les échecs si les élèves n'ont pas la possibilité de s'informer du sens réel des questions avant d'essayer d'y répondre ?

C. **Essayer ou prétester les questions.**

Autant que possible, il faut essayer les questions avant de les utiliser pour un examen. On découvre ainsi les imprécisions, les défauts de rédaction, les erreurs matérielles et, aussi, le niveau de difficulté.

Pareil essai est difficilement réalisable par un enseignant isolé ; de nouveau, le travail en groupe offre bien des possibilités en ce domaine, l'examen étant préparé longtemps à l'avance.

D. **Calcul de la facilité des questions.**

S'il est possible de prétester sur un grand nombre de sujets, ce calcul se fera avant l'examen. Sinon, il est néanmoins utile de procéder à l'opération, soit pour mieux percevoir la physionomie des réponses, soit pour un usage ultérieur.

Le *pourcentage de réussite* est l'indice le plus simple. Rappelons, toutefois, qu'il n'est pas correct de calculer des pourcentages à partir de petits nombres. Normalement, le nombre de sujets devrait excéder cent.

Il faut y insister : une légère différence entre des pourcentages est loin d'être toujours significative. Des procédés statistiques simples permettent de vérifier cet aspect.

1 A. DAVIS, Education for the Conservation of Human Ressources, In *Progressive Education*, 1950, 27, 221-224.

Une fois l'indice de facilité connu, le professeur sait mieux comment doser l'examen.

E. **Calcul de l'efficacité - Pouvoir discriminatif.**

L'indice de facilité seul peut induire en erreur, car il résulte parfois de facteurs accidentels. Il est donc prudent de l'accompagner de l'*indice d'efficacité*, qui révèle dans quelle mesure une question déterminée *discrimine* les élèves forts des élèves faibles.

1. *Méthode simple.*

Pidgeon et Yates proposent une méthode de calcul simple, fort utile pour les enseignants :

- Diviser la classe en trois groupes : supérieur, moyen et inférieur, sur la base des scores totaux à l'examen considéré.
- Pour chaque *item*, voir quel pourcentage de chacun des trois groupes a réussi.
- Pour chaque *item*, la différence entre le pourcentage de réussite du groupe supérieur et celui du groupe inférieur donne une bonne estimation de l'efficacité.
  Si la différence est faible ou nulle, c'est que la question ne distingue pas bien les élèves forts des élèves faibles.

Plus tard, nous verrons que, plus la discrimination est fine, plus la distribution des résultats est large (la courbe de Gauss s'aplatit).

2. *Méthode plus fine.*

Voici une autre méthode, utilisée pour l'examen de fin du secondaire en Angleterre[1].

Quatre groupes A, B, C, D sont constitués.

On donne les consignes suivantes :

1) Déterminer, pour chacun des *items* et dans chacun des quatre groupes, le nombre d'élèves qui répondent correctement.

2) Calculer la moyenne obtenue par chacun des quatre groupes pour chaque *item* ou question.

1 D. MATHER et al., *The C.S.E., A Handbook or Moderators*, Londres, Collins, 1965, p. 108.
2 A partir d'un ensemble de 80 copies, D. Mather constitue quatre groupes de 20 : A = les 20 premiers ; B = les 20 suivants ; etc.

3) Réunir toutes ces données en un tableau d'ensemble et comparer les moyennes pour chaque *item* dans les quatre groupes.
Si les quatre moyennes s'ordonnent de la même façon que les scores moyens pour le test entier, dans les quatre groupes, on peut affirmer que l'*item* considéré contribue à la discrimination totale du test. Sinon, l'*item* est suspect.

Exemple :

Question n°	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*Total*
Maximum	6	6	6	6	10	8	8	50
Groupe A	4,9	5,1	5,3	5,2	4,3	5,8	4,3	34,9
Groupe B	3,2	4,9	5,8	4,1	4,7	5,9	3,8	32,3
Groupe C	3,4	4,6	5,4	3	3,4	3,6	3,9	27,3
Groupe D	2	3,4	5	2,8	2,3	4,1	2,3	21,8
Ordre des scores moyens	A C B D	A B C D	B C A D	A B C D	B A C D	B A D D	A C B D	A B C D
Discrimine	Non	Oui	Non	Oui	Non	Non	Non	Oui

La réponse que l'on obtient ainsi manque de finesse. Pour les *items* considérés comme acceptables, nous ne pouvons pas dire s'ils discriminent finement ou si, au contraire, ils sont tout juste acceptables. De plus, nous ne savons trop comment améliorer ou la rédaction de la question, ou la façon de l'évaluer.

Pour obtenir des informations plus fines, les statisticiens utilisent des techniques compliquées[1]. La solution graphique suivante est simple et apporte les nuances souhaitées.

Considérons les résultats individuels de chacun des 20 élèves de chaque groupe. Nous prendrons pour exemple les questions 2, 3 et 4. Au lieu d'écrire les résultats en chiffres, nous les pointons dans un tableau.

1 Encore que l'utilisation de l'ordinateur les mette aujourd'hui à la portée de chacun.

Question 2 - Maximum 6

Score obt.	*Groupes* A	B	C	D
6	𝍸 I	𝍸 I	II	
5	𝍸 𝍸	𝍸 I	𝍸 I	𝍸 III
4	IIII	IIII	𝍸 III	II
3		II		𝍸 I
2				IIII
1				
0				II

Question 3 - Maximum 6

Score obt.	*Groupes* A	B	C	D
6	𝍸 𝍸	𝍸 III	𝍸 𝍸	𝍸 III
5	𝍸 I	IIII	II	𝍸 III
4	IIII		IIII	IIII
3				II
2				
1				
0				

Question 4 - Maximum 6

Score obt.	*Groupes* A	B	C	D
6	𝍸 𝍸	𝍸 I	II	𝍸 III
5	𝍸 IIII	IIII	II	
4	II	II	IIII	II
3	II	II	II	
2		II	II	II
1				II
0		II	IIII	𝍸 III

Si, sur des tableaux semblables, nous indiquons le quartile supérieur, le médian et le quartile inférieur[1], le comportement des élèves apparaît clairement.

1 Le MEDIAN : note du milieu dans une série de notes ordonnées : si le nombre de notes est pair, on calcule la moyenne arithmétique entre les deux notes médianes

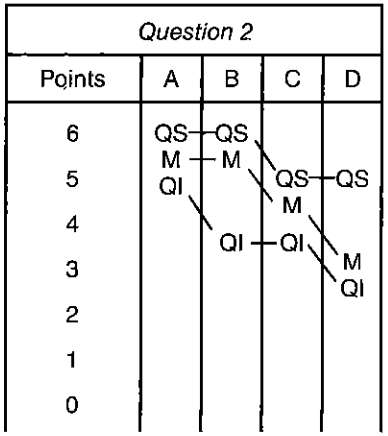
2 4 8 10 11 13

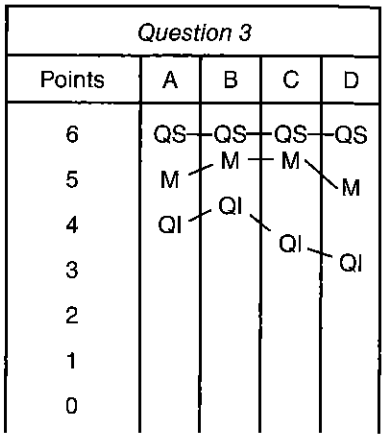
- 9 -

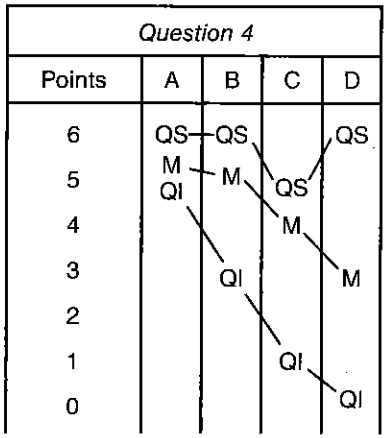
Dans une échelle de mesure, le médian est le point au-dessus duquel se trouve exactement la moitié des cas.

Le QUARTILE SUPERIEUR ou Q_3 est la note du milieu de la moitié supérieure de la série.

Le QUARTILE INFERIEUR ou Q_1 est la note du milieu de la moitié inférieure.







Idéalement, pour chacune des mesures, nous devrions observer une descente continue de A en D. La descente est plus ou moins accusée, selon le pouvoir discriminatif. Une montée indique qu'un groupe inférieur (d'après le résultat à l'ensemble de l'examen) obtient de meilleurs points que le groupe supérieur.

On observe:

Question 2:

- Il n'y a aucune remontée et une tendance à la descente, mais pas très nette.

Item faiblement discriminatif.

Question 3:

- Aucune descente pour le quartile supérieur.
- Remontée pour le médian.
- Remontée pour le quartile inférieur.

Item non discriminatif. A rejeter.

Question 4:

- Aucune remontée.
- Descente bien marquée pour le médian.
- Descente très bien marquée pour le quartile inférieur.

Bonne discrimination. Une réserve, cependant, pour le QS du groupe D: les correcteurs ont été trop généreux pour les meilleurs élèves du groupe faible (D).

II. Epreuves de performance ou épreuves de récitation? Un débat fondamental.

La plupart des examens et des tests continuent à demander, sous forme plus ou moins déguisée, une simple récitation écrite (épreuves dites «papier-crayon[1]») ou orale.

Dans une première acception, on appelle *épreuves de performance* celles qui appellent des réponses motrices, manuelles ou non: manipulation d'objets, construction d'après un modèle, assemblage, ... Cette limitation au domaine moteur est cependant erronée, du moins dans la perspective actuelle. En effet, on considère plus généralement comme test de performance toute épreuve qui oblige l'élève à construire effectivement sa réponse, au lieu de n'avoir qu'à la choisir toute faite dans sa mémoire ou en appliquant une démarche de résolution de problème routinière et stéréotypée.

Ne relèvent pas des épreuves de performance des questions telles que:

- Quelle est la formule de l'acide sulfurique? (Pure mémoire).
- Sachant que l'aire d'un triangle se calcule par la formule *base x hauteur/2*, quelle est l'aire d'un triangle de 25 cm de base et de 30 cm de hauteur? (Compréhension).
- Calculez l'aire d'un triangle de 25 cm de base et de 30 cm de hauteur. (Application).

1 On classe aussi dans la catégorie *papier-crayon* les examens ou tests non verbaux dont les consignes sont données oralement et où les questions sont posées sous forme de dessins, de symboles, de photos parmi lesquels il faut faire un choix ou sur lesquels il faut indiquer une caractéristique.

Dans la taxonomie de B.S. Bloom (domaine cognitif) et selon le sens conventionnel qui y est attribué aux termes *mémoire, compréhension, application,* ces trois questions font appel à des processus cognitifs inférieurs.

En revanche, bien que papier-crayon, la question à choix multiple suivante exige la mise en œuvre de processus cognitifs supérieurs et relève donc bien des épreuves de performance:

Le parquet d'une chambre doit être fait de lattes entières. A cet effet, M lattes de a cm sur b cm sont nécessaires. Combien faudrait-il de lattes mesurant x sur y cm ?

A. $\frac{Mab}{xy}$ B. $\frac{ab}{M\,x\,y}$ C. $\frac{(a+b)\,M}{x+y}$ D. $\frac{ab\,.\,xy}{M}$ E. $\frac{M\,x\,y}{ab}$

Ne relèvent pas non plus des épreuves de performance des examens où l'on peut répondre, pour l'avoir appris de mémoire, ce qu'il faudrait faire dans un cas donné, mais où l'on ne le fait pas effectivement. Exposer oralement ou par écrit la pédagogie progressiste que l'on pratiquerait si l'on était professeur ne prouve en rien qu'on est capable de la pratiquer effectivement, voire qu'il est possible de le faire étant donné les limites circonstancielles qui existent dans la plupart des situations scolaires.

Depuis peu s'est développé un courant très fort en faveur d'examens qui portent sur la performance réelle, non seulement dans des matières relativement simples comme la dactylographie ou la capacité de régler un moteur, mais à propos d'apprentissages plus abstraits et plus complexes. La difficulté réside dans la conception d'épreuves qui permettent d'évaluer de façon valide ce que le sujet est effectivement capable de faire[1].

Trois grandes voies s'offrent en matière de testing de performance:

1. Pour les disciplines théoriques (philosophie, logique, algèbre, physique, ...), poser des questions qui exigent la mise en œuvre de processus cognitifs supérieurs.
2. Faire effectuer la performance cible (par exemple, faire une leçon de langue étrangère) devant les examinateurs.
3. Examiner des productions réelles à travers lesquelles la qualité de la performance peut être estimée. Il s'agit alors d'un *examen sur dossier* (le *portfolio* des Anglo-Saxons). Par exemple, pour faire valoir ses capacités pédagogiques, un professeur peut constituer un dossier qui réunit des préparations de leçons, du matériel didactique qu'il a confectionné, des enregistrements vidéo de ses leçons, des travaux d'élèves, des enregistrements de conversations avec des parents d'élèves, des articles publiés, etc.

1 R.L. LINN, E.L. BAKER et S.B. DUBAR, Complex, performance-based assessment: explications and validation criteria, *Evaluation Comment,* Hiver 1991-92, 3-9.

Parmi les aspects qui menacent la validité des épreuves de performance, relevons[1]:

- Si le type de performance demandé est connu, le bachotage peut réapparaître.
- Un dossier ne se constitue pas naïvement. Si l'intéressé le prépare seul, il y placera sélectivement des témoignages à son avantage. Et si le dossier est constitué de façon négociée entre évalué et évaluateur, ce sont les meilleures pièces qui seront choisies de préférence. La solution théorique consiste à choisir les témoignages au hasard en s'efforçant de constituer un échantillon représentatif des performances. C'est plus facile à dire qu'à faire!
- L'évaluation des performances peut, elle aussi, être biaisée par la perception qu'en ont les évaluateurs.
- Il est important d'établir si les performances observées sont réellement représentatives de la capacité ou de l'habileté.
- Les épreuves de performance sont censées mieux se prêter que d'autres à l'évaluation des processus cognitifs supérieurs: résolution de problèmes, esprit critique, raisonnement, métacognition, ... Si la tâche proposée est très familière au sujet, il peut l'exécuter de façon routinière, sans mettre en œuvre des processus supérieurs.
- Les tâches doivent être significatives pour les sujets.
- Enfin, les épreuves de performance prennent beaucoup de temps, ce qui empêche souvent de proposer une quantité et une diversité de tâches suffisantes.

III. Réponses ouvertes ou fermées?

A une question à *réponse ouverte,* l'élève répond spontanément, en utilisant son propre vocabulaire. La *réponse* est dite *fermée* si le sujet est tenu d'opérer un choix parmi plusieurs réponses proposées.

Traditionnellement, l'école s'en est exclusivement tenue au premier type de questions; elle négligeait ainsi un outil de grande utilité.

1 R.J. STIGGINS, Design and development of performance assessment, *Educational measurement: Issues and practice,* 1987, 6, 3, 33-42.

A. **Réponses ouvertes** (orales ou écrites).

Ce sont les questions les plus naturelles, celles que nous posons à tout instant dans la vie.

Elles conviennent spécialement, soit pour des épreuves de contrôle faites rapidement, en toute spontanéité par les maîtres, en cours d'enseignement, soit pour la vérification d'apprentissages tellement complexes qu'ils échappent à l'analyse rigoureuse.

L'évaluation des capacités supérieures (créativité, jugement, esprit critique,...) semble de leur ressort. Pas exclusivement, toutefois. D'abord parce que, comme nous allons le voir, les questions à choix multiple bien construites permettent des explorations beaucoup plus subtiles qu'il n'y paraît. Ensuite, parce que, par souci d'objectivité, de rigueur, les notateurs tendent peut-être, de façon inconsciente, à ne retenir des réponses ouvertes que les éléments les plus concrets, les plus factuels qu'elles contiennent. La divergence des évaluations n'est-elle pas fonction de la subtilité de l'objet sur lequel elles portent?

«Nous devons admettre, écrit Vernon[1], que les élèves excellents ont l'occasion de montrer certaines qualités exceptionnelles dans l'examen traditionnel, et qu'un notateur perspicace peut, parfois, s'en apercevoir, alors que les autres peuvent très bien pénaliser la réponse à cause de son anticonformisme.» Le débat est loin d'être clos.

En construisant entièrement la réponse, l'élève tente de prouver deux choses: sa connaissance de la matière et sa capacité à l'exprimer verbalement.

Longtemps, on a cru que les deux allaient de pair. Le vieil adage: «Ce que l'on conçoit bien s'énonce clairement, Et les mots pour le dire arrivent aisément» fut vérité reçue par bien des maîtres jusqu'à aujourd'hui. Pourtant, les choses sont loin d'être aussi simples. La traduction verbale de la pensée n'est qu'une forme d'expression parmi d'autres, dédaignées pendant des siècles par les classes sociales privilégiées auxquelles tout travail manuel, tout traitement du réel répugnaient.

Renoncer entièrement aux questions à réponses ouvertes serait une erreur dans une civilisation où la communication verbale reste dominante. Mais il est parfois bon d'isoler, au moins partiellement, les connaissances et les capacités d'expression verbale, ce que permettent les réponses fermées. Les professeurs doivent d'autant moins

1 *Bull*, 4, p. 7.

hésiter à y recourir que, comme le remarquent Pidgeon et Yates, on n'a jamais démontré que leur utilisation, même fréquente, nuisait au développement verbal.

Une réponse formulée en toute liberté présente un autre inconvénient grave: son caractère unique se prête mal à l'évaluation par comparaison avec les réponses d'autres individus. C'est assurément une des principales sources des désaccords entre correcteurs, si souvent dénoncés par la docimologie.

Remarquons enfin que, même si les questions ouvertes se prêtent spécialement bien à la mise en œuvre des processus mentaux supérieurs, les examinateurs sont loin de toujours faire usage de cette possibilité.

Krumm et Seidel[1] ont analysé 2.825 exercices faits en classe et 3.827 questions ouvertes d'examens. Tous se rapportaient à l'enseignement commercial dans tous les Länder d'Allemagne fédérale. L'appel à la mémoire («connaissances» selon Bloom) était respectivement de 96,6 % et de 93 %...

B. **Réponses fermées. Questions** *(items)* **à choix multiple.**

1. *Utilité.*

L'examen intensif vise à vérifier en détail la qualité des acquisitions; il a souvent un but diagnostique. L'examen extensif porte sur une matière vaste.

Dans les deux cas, de nombreuses vérifications sont nécessaires. D'où le recours aux questions à «réponses fermées», surtout du type «à choix multiple».

Exemple: Le premier test d'intelligence utilisable dans la pratique courante a été construit par:

a) Binet

b) Galton

c) Goddard

d) Spearman

e) Terman.

1 V. KRUMM et G. SEIDEL, *Wirtschaftslehrtest*, Weinheim, Beltz, 1970, cité par INGENKAMP Ed., *Tests in der Schulpraxis*, Weinheim, Beltz, 1972, p. 112.
Pour une étude approfondie, voir R. WOOD, *Multiple choice*, et D. LECLERCQ, *La conception, les questions à choix multiple*, Bruxelles, Labor; Nathan, Paris, 1986.

2. *Constituer une provision de questions.*

Trouver plusieurs dizaines de questions précises, pour un seul examen, met l'imagination à rude épreuve. Constituer un fichier d'*items* que l'on enrichit à chaque occasion facilite considérablement la tâche.

On n'inscrit qu'une question par fiche afin de pouvoir porter aussi les «distracteurs», c'est-à-dire les réponses fausses, mais vraisemblables, à mesure de leur découverte. Ces distracteurs sont notamment fournis par les erreurs commises fréquemment par les élèves; on est ainsi assuré de respecter leur «logique» et la vraisemblance est garantie.

Les professeurs d'une même branche peuvent aisément unir leurs efforts en cette matière. On semble aussi s'acheminer vers la création de *banques d'items*, offices centraux pouvant mettre à la disposition des maîtres plusieurs centaines de questions à choix multiple, bien mises au point pour des populations déterminées.

3. Exploiter la gamme des possibilités logiques.

Répondre à un *item* à choix multiple n'est pas nécessairement un simple exercice de mémoire. Loin s'en faut[1].

Noizet et Caverni[2] écrivent justement:

> «L'avantage présenté par le QCM comme instrument d'évaluation, c'est qu'il ne permet pas l'esquive, à la différence des épreuves dites de production qui permettent toujours au candidat de masquer une absence de savoir. Il exige non seulement des connaissances précises, mais des connaissances organisées. Tel candidat capable de donner une définition de X se trouvera embarrassé s'il se trouve affronté à la question «Pour qu'un X soit Y, **il suffit que**». Il est alors contraint, pour répondre, de délimiter précisément la part du nécessaire et la part du suffisant dans la définition de X. Le fait de mettre le candidat en difficulté en lui faisant prendre pour une condition suffisante ce qui est une condition simplement nécessaire montre que le QCM oblige à un degré de plus dans l'élaboration des connaissances. Encore faudra-t-il, pour que les QCM soient utilisés dans toute la plénitude de leurs possibilités, que soient diminués sinon supprimés les biais qui interviennent dans les stratégies de réponse. Pour les supprimer éventuellement, il faut d'abord les connaître.» C'est ce à quoi doivent tendre les recherches docimologiques.

Les recherches entreprises par les services de l'enseignement supérieur en vue de la réforme des examens de médecine, en France, ont bien mis en lumière la richesse du système[3]. Nous choisissons à

1 La technique a même été appliquée avec succès à l'analyse littéraire - voir B. CHOPPIN et A. PURVES, A comparison of open-ended and multiple choice items dealing with literary understanding, *Research in the Teaching of English*, 3, 1, 1969, 15-24.

2 G. NOIZET et J.-P. CAVERNI, *o.c.*, pp. 186-187.

3 Paris, Ministère de l'Education nationale. Enseignement supérieur, Examens et concours, section médicale 1961. Tous les exemples médicaux suivants sont empruntés à cette publication.

dessein des exemples dans le domaine des examens universitaires, dans l'espoir de convaincre les maîtres que, quelle que soit la matière qu'ils enseignent, le recours à l'examen objectif n'est pas exclu *a priori*.

Dans le présent examen, on distingue neuf types de questions:

a) **Question à complément simple.**

L'*item* se présente, dans son esprit sinon toujours dans sa forme, comme une phrase à compléter:

Ex.: Parmi les caractères suivants, celui qui s'applique à toutes les enzymes est:

A. Elles contiennent toujours une coenzyme dissociable.
B. Elles sont thermostables.
C. Elles contiennent toujours de l'azote dans leur molécule.
D. Elles contiennent toujours du phosphore dans leur molécule.
E. Elles sont dialysables.

Remarquons qu'il n'est pas toujours nécessaire que la réponse correcte figure parmi les choix proposés. L'un d'eux peut être: «Aucune des réponses précédentes.»

L'*item* peut prendre la forme négative:

Ex.: Un hydrosol métallique a tous les caractères suivants, sauf un. Indiquez lequel.

A — B — C — D — E

Pour répondre, le candidat doit connaître tous les caractères de l'hydrosol. Toutefois, la forme négative contraint à une gymnastique intellectuelle qui se superpose à la difficulté inhérente à la matière.

b) **Association simple.**

Elle sert à vérifier la connaissance d'un certain nombre d'entités qui peuvent être ou ne pas être en relation.

Il faut ici faire correspondre un élément précédé d'une lettre à un élément précédé d'un chiffre (association ou appariement).

Ex.: *Branche postérieure du nerf radial*

A. Segment d'origine
B. Groupe des rameaux postérieurs
C. Groupe des rameaux antérieurs
D. Nerf interosseux postérieur

1. est appliqué sur la face postérieure du ligament interosseux
2. contourne le col du radius
3. innerve les muscles de la couche superficielle de la région antibrachiale postérieure
4. passe entre les deux chefs du court supinateur
5. innerve les muscles de la couche profonde de la région antibrachiale postérieure

c) **Association composée.**

C'est une simple variante de la précédente.

Exemple:

A. Paludisme à plasmodium vivax
B. Paludisme à plasmodium falciparum
C. Les deux à la fois (A et B)
D. Aucun des deux

Question 1:

1. Le développement clinique a toutes chances d'être moins grave chez un homme de race noire que chez un homme de race blanche (réponse B)
2. Une association de primaquine et de chloroquine est le traitement de choix pour une attaque aiguë (réponse A)
3. Les épisodes cliniques sont supprimés par l'ingestion de chloroquine une fois par semaine en zone endémique (réponse C)
4. Guérit d'une manière définitive par le traitement avec la chloroquine (réponse B)
5. La contamination est évitée par l'ingestion de chloroquine une fois par semaine (réponse D)

Remarquons que le nombre de questions peut être ici augmenté ou diminué selon l'importance accordée à la matière.

d) **Association à terme exclu.**

Dans l'*item* suivant, 4 des 5 phénomènes numérotés sont communs à un des troubles A.B.C. Il faut indiquer le trouble (A) et le phénomène qui n'y correspond pas (2).

Exemple:

A. éosinophilie d'importance diagnostique
B. plasmocytose d'importance diagnostique
C. lymphocytose d'importance diagnostique

1° trichynose
2° myélome multiple
3° syndrome de Lœffler
4° maladie de Hodgkin
5° schistosomiase

e) **Analyse de relations de cause à effet.**

Ex.: L'articulation radio-cubitale supérieure permet des mouvements de rotation limités PARCE QUE la tête du radius est entourée par le ligament annulaire.

A. La constatation et la raison proposée sont toutes les deux vraies et il existe une relation de cause à effet entre elles.
B. La constatation et la raison proposée sont toutes les deux vraies et il n'y a pas de relation de cause à effet entre elles.
C. La constatation est vraie, mais la raison proposée est fausse.
D. La constatation est fausse, mais la raison proposée est un fait ou un principe accepté.
E. La constatation et la raison proposée sont toutes les deux fausses.

f) **Analyse d'observations.**

L'ensemble complexe suivant met le candidat dans une situation comparable à l'expérience réelle.

Exposé de la maladie : Le malade est un homme de 21 ans qui se plaint de malaises, d'une toux et de fièvre.

La maladie a débuté dix jours avant l'admission par un malaise et une toux sans expectoration, suivis dans les 24 heures d'une température variant de 37.8 à 38.3 qui a persisté jusqu'au moment de l'admission.

Le quatrième jour de la maladie, la toux s'accentua, produisant de petites quantités d'expectoration blanche et visqueuse.

Trois jours avant l'admission, des accès paroxystiques de toux commencèrent, parfois suivis de vomissements. Des sensations de frissons furent notées, mais non pas de véritables frissons avec tremblements. Une douleur parasternale antérieure à la toux existe depuis le cinquième jour de maladie.

A l'examen physique, la température est à 38.3, le pouls à 110, le rythme respiratoire 32, la tension maxima 10 ½, minima 8.

Le malade est bien développé, sans maigreur, sa maladie semble aiguë, il est dyspnéique, mais non cyanosé.

L'examen physique de la cage thoracique montre des vibrations vocales à la palpation et à l'auscultation. Le murmure vésiculaire est normal. Dans l'aisselle gauche, on entend quelques râles fins et la qualité bronchique du son est augmentée, bien que d'intensité normale.

La formule sanguine est la suivante: globules blancs 3.400 (polynucléaires 30 %, lymphocytes 62 %, monocytes 5 %, éosinophiles 3 %).

La radio du thorax révèle une augmentation de la densité de la région périhilaire avec des aires mal définies de densité inégale, nuageuses aux deux bases et dans un champ pulmonaire supérieur gauche.

Questions:

1. *Quel est le diagnostic le plus probable ?*
   a) tuberculose
   b) pneumonie à pneumocoques
   c) pneumonie (primaire atypique) à virus
   d) coccidiomycose
   e) bronchopneumonie

2. *Quel est le signe physique qui s'y ajoute probablement ?*
   a) spénomégalie
   b) signe de souffrance méningée
   c) bruit de frottements pleuraux
   d) changements fréquents dans la distribution des symptômes thoraciques
   e) signe de condensation lobaire gauche

3. *Lequel des examens de laboratoire suivants va de pair avec le diagnostic ?*
   a) l'élévation et l'augmentation des agglutinines froides
   b) hémoculture positive
   c) leucocytose marquée au début de la convalescence
   d) examen des expectorations
   e) cuti-réaction positive

4. *Quelle thérapeutique devra être employée ?*
   a) repos au lit et streptomycine
   b) repos au lit et pénicilline
   c) streptomycine et acide para-amino-salicylique
   d) repos au lit et auréomycine
   e) psychothérapie et rééducation physique

5. *Quelle est l'issue probable de cette maladie sans traitement ?*
   a) la fièvre va disparaître spontanément par une crise terminale
   b) la convalescence va être progressive avec une rechute prévisible
   c) un empyème résiduel va se développer
   d) une fibrose résiduelle va apparaître après guérison
   e) une caverne pulmonaire peut apparaître

g) **Comparaisons quantitatives.**

Comparez X à Y:

X — pression mécanique dans le capillaire veineux
Y — pression oncotique dans le capillaire veineux

Et dites si:

A. X est plus grand que Y
B. Y est plus grand que X
C. X est égal à Y

h) **Relations.**

Soit:

1. Le débit circulaire cutané — ET
2. La quantité de chaleur perdue par unité de temps

A. L'augmentation du premier est accompagnée d'une augmentation du second, ou la diminution du premier est accompagnée d'une diminution du second.
B. L'augmentation du premier est accompagnée d'une diminution du second, ou la diminution du premier est accompagnée d'une augmentation du second.
C. Les variations du second sont indépendantes des variations du premier.

i) **Compléments groupés.**

Procédé utilisé quand une question peut avoir plus d'une réponse correcte.

Ex.: Cinq conscrits mesurent: 1,65 m - 1,67 m - 1,69 m - 1,63 m et 1,61 m.

1. La moyenne des tailles de l'échantillon est 1,65 m
2. L'écart type est proche de 8
3. L'écart type est proche de 2,8
4. L'échantillon a de bonnes chances de renseigner sur la population des tailles des individus en général

A. 1, 2 et 3 sont corrects
B. 1 et 3 sont corrects
C. 2 et 4 sont corrects
D. 4 est correct
E. Une seule des propositions 1, 2 ou 3 est correcte

On trouvera, en annexe, une comparaison entre un examen conduit selon la méthode traditionnelle et selon la méthode à choix multiple.

4. *Calcul de l'efficacité des distracteurs.*

Dans une question à choix multiple, il importe de vérifier si les distracteurs jouent bien leur rôle. A cet effet, on calcule, pour chaque question, quel pourcentage des choix chaque possibilité de réponse a recueilli.

a) **Situation idéale.**

A	B	C	D	E
10%	10%	10%	60%	10%

D est la bonne réponse.

Les autres choix ont été également attractifs.

b) **Situation à corriger.**

A	B	C	D	E
3%	0%	35%	60%	2%

Seul le distracteur C a réellement joué. Les autres se révèlent sans pouvoir.

5. *Critiques et réfutation partielle.*

Bien que sa valeur soit prouvée depuis longtemps, l'examen par questions à choix multiple suscite de vives critiques. Certaines peuvent être aisément réfutées; d'autres paraissent justifiées; d'autres encore ne peuvent être ni infirmées, ni confirmées, faute de critères scientifiques.

a) **Une objectivité trompeuse.**

La notation des questions à choix multiple est indiscutablement objective: les réponses correctes sont définies d'avance; l'élève les a trouvées ou non. Les avantages de cette méthode sont évidents.

Pourtant, la subjectivité est loin d'avoir été complètement éliminée: elle subsiste, au moins en partie, dans la rédaction des questions et dans la décision concernant la réponse à considérer comme correcte.

Dans la rédaction des questions d'abord. Elles sont le fruit de la réflexion, de l'invention des examinateurs qui, en dernière analyse, agissent subjectivement. A côté des questions auxquelles ils ont pensé, combien d'autres, peut-être plus valides, n'aurait-on pas pu imaginer?

Dans la rédaction des réponses considérées comme idéales, ensuite. On a rarement affaire à des réponses aussi évidemment correctes que 2 x 2 = 4. Au cours de la construction de tests avec nos étudiants, nous avons plus d'une fois rencontré des réponses proposées comme correctes, passées victorieusement au crible de l'analyse mathématique, et qui, pourtant, n'étaient, au mieux, que grossières approximations exprimées dans un langage douteux.

Comme Vernon le note, non sans malice, la supériorité indiscutable des examens «objectifs» sur les épreuves traditionnelles est sans doute plus souvent due à la préparation très soignée et à l'union des efforts et de la compétence de nombreux enseignants et psychopédagogues, qu'à la nature même de l'examen[1].

On oublie aussi qu'entre l'utopie de l'objectivité et la subjectivité totale, bien des stades intermédiaires existent.

b) **Choix «corrects» contestables.**

Ce point aussi a été excellemment discuté par Vernon[2] qui écrit:

«Il est fréquent que des personnes hautement éduquées accueillent les questions à choix multiple par des critiques telles que: «Cet *item* est stupide» ou prétendent que les choix, considérés comme «faux» par le testeur, sont tout aussi admissibles, sinon plus, que la réponse dite «bonne». Cette critique vient principalement du fait que l'*item* objectif ne s'appuie pas sur les mêmes capacités que les questions traditionnelles. Evidemment, des examens construits par des amateurs peuvent contenir pas mal d'*items* défectueux. On en trouve aussi dans les épreuves construites par des professionnels, mais plus rarement, car les *items* insatisfaisants sont ou bien rejetés, lors du contrôle préliminaire, par des spécialistes de la discipline concernée, ou bien éliminés lors de l'analyse statistique des réponses (...). Les critiques risquent aussi d'oublier que leurs réactions sophistiquées peuvent différer très fort de celles des élèves sur lesquels les *items* ont été essayés. Enfin, les critiques peuvent lire dans les questions des choses qui ne viennent même pas à l'esprit d'élèves intelligents.»

1 VERNON, The C.S.E.: *An Introduction to Objective-Type Examinations*, Londres, H.M.S.O., 1964, pp. 4-5.
2 VERNON, o.c., p. 6.

c) **Un jeu de hasard.**

Dans une question fermée à deux choix de réponses, l'une correcte et l'autre fausse, la probabilité de réussite de l'élève qui travaille au pur hasard est, théoriquement, de 50%. Les constructeurs de tests le savent depuis longtemps.

Aujourd'hui, on a le plus souvent recours aux questions à cinq choix, dont l'un est correct et les autres seulement vraisemblables (distracteurs)[1]. Dans ce cas, le jeu du hasard est considérablement réduit. La démarche des élèves variera d'ailleurs selon leur niveau de connaissance de la matière. En vertu de la loi de la probabilité, l'étudiant qui ignore tout n'a qu'une chance minime de tomber aveuglément sur la solution correcte. Par contre, une connaissance partielle permet d'éliminer sciemment un certain nombre de distracteurs; la chance de réussite dans le choix aveugle parmi les possibilités restantes est alors beaucoup plus grande que dans le cas précédent, et c'est justice.

On peut d'ailleurs réduire largement le rôle de la chance en corrigeant les résultats obtenus à l'aide d'une formule simple[2], laquelle, il est vrai, pénalise généralement de façon exagérée, puisque la part du choix aveugle varie selon les élèves.

Pour pallier cet inconvénient, voire cette injustice, il semble préférable d'inviter le sujet à indiquer dans quelle mesure il est certain de ses réponses.

Cette évaluation de la *certitude* se fait, par exemple, selon l'échelle suivante:

absolument certain	2
doute	1
ignore (pas de réponse)	0

Un système complexe de notation récompense le sujet qui évalue correctement la sûreté de ses connaissances et pénalise celui qui se trompe à cet égard[3].

1 Parfois aussi, plusieurs choix conviennent, mais à des degrés divers. Il faut alors sélectionner le plus adéquat.

2 Nombre de réponses correctes - $\frac{\text{Nombre de réponses incorrectes}}{\text{Nombre de choix - 1}}$

3 Pour un traitement approfondi de l'évaluation de la certitude, voir D. LECLERCQ, Confidence marking: Its use in testing, *Evaluation in Education: An International Review Series*, vol. 5, 2, 1982 (Oxford, Pergamon).

Par exemple, le système suivant repose sur la théorie des décisions. Il est dû à D. Leclercq et a été utilisé à l'Université de Liège.

Si vous considérez que votre réponse a une probabilité d'être correcte comprise entre	Ecrivez	Vous obtiendrez les points suivants en cas de réponse	
		correcte	incorrecte
0% et 25%	0	+ 13	+ 4
25% et 50%	1	+ 16	+ 3
50% et 70%	2	+ 17	+ 2
70% et 85%	3	+ 18	+ 0
85% et 95%	4	+ 19	- 6
95% et 100%	5	+ 20	- 20

d) **Acrobatie mentale.**

On reproche aux examens traditionnels de faire la part trop belle à la facilité d'expression. On peut toutefois se demander si les *items* à choix multiple un peu compliqués (nous venons d'en rencontrer des exemples) ne présentent pas un inconvénient au moins aussi grave en ajoutant, à la difficulté inhérente à la matière de la question, l'obligation de démêler des doubles négations[1], de saisir des subtilités logiques ou, plus généralement, en accordant une prime à une certaine aptitude à l'abstraction à partir de données verbales. P. Vernon constate, en tout cas, que la compréhension de la lecture joue un rôle important dans la réussite de ces épreuves[2].

Il importe, non seulement de connaître le niveau de développement mental des élèves, mais aussi de savoir clairement ce que l'on veut: vérifier la connaissance de la matière, la capacité en compréhension de la lecture et en raisonnement, ou la combinaison des deux.

On ne voit pas comment le sondage des qualités intellectuelles supérieures irait sans un accroissement de la complexité des *items*. C'est pourquoi plusieurs auteurs estiment que les questions ouvertes continueront à jouer un rôle important dans les examens de niveau avancé.

1 On constate que les sujets trouvent plus facilement les énoncés vrais que les énoncés faux et que, plus généralement, une proposition négative est plus difficilement comprise qu'une affirmation.

2 P. VERNON, The Determinants of Reading Comprehension, in *Educational Psychological Measurement*, 1962, 22, 269-286.

En cas de recours à la correction automatique, une autre difficulté vient encore s'ajouter : l'utilisation de cartes de réponses où ne figurent généralement que les numéros des questions et les lettres A, B, C, D, E, représentant les cinq choix. Ce système exige un supplément d'attention de la part de l'élève, en particulier quand il ne peut répondre à certaines questions et décide de les sauter. S'il oublie de sauter aussi la ligne correspondante sur la carte de réponses, les conséquences peuvent être désastreuses. La difficulté ne doit toutefois pas être exagérée : Remmers, Gage et Rummel ont montré expérimentalement que le système est utilisable avec des enfants, à partir de 9 ou 10 ans[1].

De toute façon, les élèves doivent être soigneusement entraînés à la technique des examens à choix multiple avant de subir une épreuve décisive. Moyennant cette précaution, et une construction rigoureuse, les questions à choix multiple donnent de bons résultats, leur utilisation intensive dans les pays anglo-saxons, depuis de nombreuses années, et de plus en plus répandue ailleurs, le prouve à suffisance.

e) **Inconvénients incertains.**

Nous qualifions les reproches suivants d'incertains parce que, à notre connaissance, aucune recherche scientifique rigoureuse n'en a établi le bien-fondé. Il semble que certaines critiques ne se justifient que dans la mesure où l'examen est mal construit.

- Choisir la bonne réponse parmi d'autres est plus facile que la construire. La mémoire intervient trop. Une certaine paresse intellectuelle, une répugnance à l'effort nécessaire, à la formulation claire de la pensée risquent de s'installer.
- Sachant que, pour subir l'examen en langue maternelle, il suffirait de souligner ou d'indiquer la réponse choisie par une croix, les maîtres négligeraient les exercices d'expression.
- La préférence irait à de nombreux petits exercices à faire en un temps très court ; on négligerait ainsi les problèmes qui exigent une très longue réflexion. Or, les études supérieures et la vie réelle ne nous épargnent pas ce genre d'épreuves.

C. En guise de conclusion : un compromis.

Un compromis nous paraît s'imposer à deux niveaux au moins.

1 H. REMMERS, N. GAGE et RUMMEL, *Educational Measurement and Evaluation*, New York, Harper, 1955, p. 246.

Puisque questions ouvertes et questions fermées semblent posséder chacune des avantages particuliers et mettre en cause certains apprentissages différents, on ne voit aucune raison de revendiquer un monopole pour l'une des deux. Elles peuvent parfaitement coexister. Les meilleurs docimologistes estiment cependant qu'il n'est pas souhaitable de mélanger les deux types de questions dans une même épreuve.

Par ailleurs, une formule intermédiaire entre questions ouvertes et fermées existe et s'imposera peut-être de plus en plus, selon les progrès de la recherche pédagogique.

L'examen consisterait en un nombre assez élevé de questions *ouvertes*, d'une portée relativement limitée, ayant été prétestées. L'éventail des réponses probables serait donc connu d'avance, ce qui permettrait de proposer un schéma de notation s'approchant très fort de la rigueur de la question à réponse fermée.

Cette possibilité est déjà confirmée par plusieurs travaux expérimentaux. Ainsi, des corrélations presque parfaites (.98 et .99) ont pu être obtenues entre sept groupes de notateurs représentant chacun une commission d'examens de Grande-Bretagne. L'expérience a été faite sur des épreuves d'histoire de niveau supérieur du G.G.E. (fin du secondaire)[1].

Sans aller aussi loin, des professeurs peuvent parfaitement s'entendre sur des points à exiger dans les réponses à des questions ouvertes.

Par ailleurs, la façon de rédiger les questions peut fermer partiellement des réponses et donc diminuer le jeu de la subjectivité. Une question comme : écrivez un début approprié à la phrase : « ... quand il commença à pleuvoir » est, en quelque sorte, à mi-chemin entre l'épreuve subjective et l'épreuve objective. Car ici, les cas de désaccord entre notateurs invités à se prononcer sur la correction de la réponse seront rares.

Nous faisons nôtre la conclusion de Wood[2] :

« Les questions à choix multiple remplissent une fonction évaluative particulière. Elles obligent le candidat à se concentrer sur certains problèmes, sans devoir écrire longuement (...). Mais les critiques veulent que les QCM soient ce qu'ils ne sont pas, lorsqu'elles regrettent qu'ils ne permettent pas de mesurer « la tolérance à l'égard de l'avis

1 JOINT MATRICULATION BOARD, *The Marking of Scripts in Advanced Level History*. Universities of Manchester, Liverpool, Leeds, Sheffield and Birmingham, 1964.

2 R. WOOD, *Multiple choice : A state of the art*, Oxford, Pergamon Press, 1977 (coll. Evaluation in Education : International Progress), p. 202.

d'autrui». Or, personne n'a jamais prétendu que les QCM jouaient ce rôle. Les QCM ont leurs faiblesses comme toutes les autres techniques d'examen. Mais, ici au moins, le candidat sait exactement ce qu'il doit faire, alors que, dans l'examen traditionnel, il doit souvent deviner ce que l'examinateur attend de lui.»

IV. Subjectivité - Objectivité

A. Théorie.

S'appuyer sur une taxonomie au lieu de s'abandonner à la simple inspiration pour rédiger des questions réduit le jeu de la subjectivité. Elle est toutefois loin d'être totalement éliminée.

Nous avons vu que, pour savoir quelles questions il doit poser, le constructeur d'un examen ou d'un test de connaissances se base généralement sur un tableau à double entrée: objectifs-matières, chaque intersection suggérant un type de questions ou d'*items*.

La subjectivité va d'abord jouer dans l'interprétation des catégories taxonomiques, rarement définies en termes de comportements observables. Le rédacteur des questions, par exemple, estime que tel *item* relève de la mémoire pure et simple, alors que tel autre appelle l'application, l'analyse, la synthèse, etc. Mais comment savoir sûrement s'il en est bien ainsi?

La question est d'autant plus redoutable que, nous l'avons vu, une tâche peut exiger la mise en œuvre de divers processus n'appartenant pas au même niveau taxonomique. En outre, selon les apprentissages antérieurs réalisés par différents individus, le niveau taxonomique change. Une même question exige de l'un une synthèse, c'est-à-dire au sens bloomien une production originale, divergente, alors qu'elle ne demande d'un autre individu que le rappel d'une solution trouvée antérieurement et déjà utilisée plusieurs fois.

De plus, la question jaillit ici aussi de l'inspiration de son rédacteur. Pour la même matière et le même niveau taxonomique, on pourrait évidemment formuler bien d'autres questions, de même difficulté, plus faciles, plus difficiles. Sauf dans des cas qui restent exceptionnels, le rédacteur ne réunit pas toutes les questions possibles avant de faire un choix (qui pourrait alors être un échantillon représentatif de l'ensemble).

Bref, l'introspection, l'inspiration continuent à jouer un grand rôle, et l'appel à des experts, dont la démarche ne diffère pas fondamentalement, ne résout pas tout le problème, loin s'en faut.

Pour surmonter cette difficulté, J.R. Bormuth[1] propose l'opérationalisation intégrale de la méthode de construction des questions. Sa théorie illustre bien les efforts d'objectivation entrepris actuellement et, même si elle n'est pas intégralement applicable dans la pratique, elle indique néanmoins une direction de recherche, voire un idéal vers lequel il faut tendre.

Pour Bormuth, opérationaliser la rédaction des questions, c'est proposer un ensemble de manipulations observables par tous, permettant de dériver de l'enseignement un *item* de test: «Une définition d'*une classe d'items de test* de connaissances consiste en une série de directives indiquant au rédacteur comment il doit réarranger des segments d'enseignement pour obtenir des *items* du type désiré. A aucun moment, ces directives ne doivent faire appel à l'introspection du rédacteur[2] ...» Et Bormuth continue: «Si l'expérimentateur ne peut pas vérifier si ses *items* sont bien du type qu'il prétend, et si d'autres expérimentateurs ne peuvent pas construire des *items* qu'ils puissent certifier d'un même type, ces autres expérimentateurs ne peuvent en rien prétendre réfuter ou confirmer les résultats originaux; en pareil cas, l'étude originale est aussi sans valeur[3].»

Bormuth propose un modèle de définition opérationnelle de questions:

1. Dans un premier temps, une structure syntaxique est donnée à la matière à explorer.
2. Ensuite, des opérations définies opèrent sur cette syntaxe et transforment les segments d'enseignement considérés en *items* de tests.

W. Hively, H. Patterson et S. Page prennent une position[4] proche de celle de J. Bormuth. S'appuyant sur la théorie de la «généralisabilité» de L. Cronbach (voir p. 196), ils définissent un test de connaissances comme un échantillon tiré d'un ensemble d'*items* relatifs à un ou plusieurs domaines soigneusement définis[5]. Pour obtenir des formes parallèles au test initial, il suffit, dans ce cas, de tirer au hasard un nouvel échantillon d'*items*.

Le problème est évidemment de définir l'ensemble, le domaine; une analyse logique ou psychologique (classes comportementales) devrait le permettre[6].

---

1 J.R. BORMUTH, *On the Theory of achievement test items*, Chicago, University Press, 1970.
2 J.R. BORMUTH, *Ibid.*, p. 5.
3 Idem.
4 W. HIVELY, H. PATTERSON and S. PAGE, *A Universe defined system of arithmetic achievement tests*, in *Journal of Educational Measurement*, 5, 1968, 275-289.
5 Ceci suppose toutefois que l'ensemble des *items* possibles est fini, ce qui est loin d'être vrai pour toutes les branches, du moins au niveau des *items* spécifiques.
6 Sur la définition du domaine, voir aussi p. 186.

W. Hively et al. distinguent:

- *La forme d'items*, c'est-à-dire des règles permettant de générer un ensemble d'*items* (= classe d'*items* chez Bormuth).
- *L'univers d'items*, c'est-à-dire une collection de formes d'*items*. Ex.: l'univers de la soustraction des nombres entiers est l'ensemble des formes d'*items* permettant d'en explorer tous les aspects.
- *La famille de tests parallèles aléatoires.* Ensemble de tests constitués en fonction d'un plan d'échantillonnage aléatoire portant sur un univers d'*items*.

  Ex.: 1. Générez un *item* pour chaque forme.
  2. Ordonnez les *items* au hasard.
  3. Recommencez les deux opérations précédentes autant de fois que vous souhaitez de tests parallèles.

La limite du système de construction opérationnelle des *items* proposé par Bormuth est clairement reconnue par Diedrich[1]. Il a en effet calculé que, pour un manuel de physique comptant environ 16 000 phrases, l'ensemble des transformations fournirait 960 000 *items*. Jamais un constructeur de test ne fera pareil détour. C'est donc vers une position intermédiaire qu'il faut s'orienter et comme l'écrit H. Rupprecht: «La formulation d'*items* opérationnellement définis ne dispense pas le constructeur de test d'analyser les textes d'enseignement et de choisir les unités à tester[1].» Une fois encore, l'intervention réfléchie, avec ce qu'elle réintroduit inévitablement de subjectivité, tente de compenser les limites pratiques d'une théorie.

Pour éviter ce retour à la subjectivité, on peut toutefois tirer au hasard les unités à tester, puis appliquer alors les principes de Bormuth. Une position en retrait, beaucoup plus proche de la pratique scolaire actuelle, est adoptée par Popham et son équipe. Aidé, dans chaque cas, d'au moins un spécialiste de la matière concernée, de plusieurs docimologues et de plusieurs maîtres enseignant au niveau auquel on destine les questions, le groupe de Popham définit d'abord des objectifs de contenu ou d'attitudes relativement généraux, puis les éclate en micro-objectifs. Suit alors la rédaction de questions dont la congruence avec les objectifs est chaque fois discutée par un groupe d'éducateurs, d'enseignants et de chercheurs.

Cette façon est naturellement moins rigoureuse que celle de Bormuth ou de Hively, mais elle offre néanmoins plus de garantie que la

1 H. RUPPRECHT, Konstruktion von Testaufgaben nach einem Verfahren von Bormuth, in K. Klauer et al., *Lernzielorientierte Tests*, Düsseldorf, Verlag Schwann, 1972.

démarche subjective classique et présente l'avantage d'être presque immédiatement réalisable dans tous les cercles scolaires.

Petit à petit s'élaborent des banques d'objectifs susceptibles de rendre d'inestimables services à la pratique scolaire, au moins sur le plan diagnostique.

B. **Quelques exemples.**

1. *Le test de closure*[1]

Traditionnellement, pour tester la compréhension d'un texte, l'évaluateur fait porter un certain nombre de questions sur ce qu'il croit nécessaire de vérifier (choix subjectif de la matière). Il détermine, subjectivement aussi, le degré de difficulté du contenu du problème posé et de la forme utilisée pour l'exprimer. En conséquence, une question logiquement ou formellement facile peut porter sur un texte difficile, tandis qu'une question difficile peut concerner un texte facile.

Le test de closure semble éviter ces écueils. Il consiste à supprimer un mot sur cinq et à inviter l'élève à combler les lacunes. Dans une première forme du test, on supprime, par exemple, le 1er, le 6^{e}, le 11^{e} mot, etc.; dans une deuxième forme, on supprime le 2^{e}, le 7^{e}, le 12^{e} mot, etc.

Avec cinq formes différentes, tous les mots du texte sont couverts. On a donc obtenu toutes les questions possibles dans cette *forme d'items*.

Cette épreuve, d'une simplicité de construction parfaite, permet une bonne mesure de la compréhension et de la lisibilité.

2. *Test de compréhension de la lecture.*

Ne fût-ce que pour compléter les indications fournies pour le test de closure, on peut souhaiter construire un test de compréhension, de facture plus classique.

Une opposition entre la classification abstraite de Davis et quelques règles opérationnelles inspirées de Bormuth montre bien, croyons-nous, la différence fondamentale entre les deux approches.

Exemple (Davis): Matière = compréhension de la lecture.

Objectifs.

Niveau I: Questions portant sur le sens d'un mot.
Niveau II: Questions portant sur l'organisation d'un passage.
Niveau III: Questions portant sur des informations contenues dans le texte.
Niveau IV: Questions portant sur des informations devant être inférées à partir du texte.
Niveau V: Questions portant sur le point de vue de l'auteur.

1 Voir G. DE LANDSHEERE, *Le test de closure, mesure de la lisibilité et de la compréhension*, Paris, Nathan; Bruxelles, Labor, 1973.

Dans cet exemple, on voit que Davis propose cinq niveaux auxquels le rédacteur du test va situer subjectivement ses questions.

Bormuth inverse partiellement la façon de procéder dans le domaine des objectifs. Ayant les objectifs généraux à l'esprit, on recherche des règles de génération d'*items* puis, celles-ci étant trouvées, on s'interroge sur les objectifs intermédiaires comme, par exemple, ceux de la taxonomie de Bloom, qu'elles peuvent desservir.

Voici quelques exemples directement inspirés de P. Menzel[1].

Supposons que l'on veuille tester la compréhension de La cigale et la fourmi. L'échelle de niveaux de Davis est remplacée par une structuration à base linguistique qui portera notamment sur:

a) **La compréhension des mots.**

Elle pourrait être testée en relevant un échantillon des mots du texte qui n'appartiennent pas à un vocabulaire de base donné et en construisant pour chacun un même type de question (fondée, par exemple, sur un ensemble de mots appartenant au même champ sémantique).

- La bise, c'est: - la pluie?
  - la neige?
  - le vent?
  - le brouillard?

b) **La compréhension de la structure de la phrase.**

On tire au sort un certain nombre de phrases et une même règle est appliquée à leurs différents constituants (par exemple, utilisation de tous les interrogatifs).

- Qui se trouve fort dépourvue?
- Quand la cigale se trouva-t-elle fort dépourvue?

Etc.

c) **La compréhension des relations anaphoriques.**

Qui dit: «Je vous paierai»?
Qui est «cette emprunteuse»?

d) **La compréhension des relations entre les phrases et entre les paragraphes.**

La succession temporelle ou les relations causales pourraient, par exemple, donner lieu à des questions ou des phrases ou des paragraphes présentés dans le désordre doivent être réarrangés.

Nous ne présentons nullement ces exemples comme les meilleurs possibles, mais tâchons simplement d'illustrer une direction de recherche.

1 P. MENZEL, *The linguistic bases of the theory of writing items for instruction stated in naturel language,* Annexe à J. BORMUTH, *op. cit.*

3. *Formes d'items pour la soustraction.*

Dans leur article déjà cité, W. Hively et al. proposent un programme complet d'arithmétique élémentaire, réparti en neuf «univers» avec la liste des formes d'*items*. Voici, pour l'univers «Soustraction de nombres entiers», quelques exemples de formes accompagnées des règles de génération.

Exemples de formes d'items pour l'univers de la soustraction[1]

Titres descriptif	*Exemple d'item*	*Forme générale*	*Règle de génération*
Grand nombre $>$ 10	13 - 6	A - B	1. A = 1a; B = b. 2. $(a < b) \in U$. 3. { H, V }.
Un emprunt; le petit nombre n'a qu'un chiffre	13 - 7	A - B	1. A = a_1a_2; B = b. 2. $a_1 \in U - \{ 1 \}$. 3. $(b > a_2) \in U_0$.
Emprunt sur 0.	403 - 138	A - B	1. $N \in \{ 3, 4 \}$. 2. A = a_1a_2...; B = b_1b_2. 3. $(a_1 > b_1)$, $(a_3 < b_3)$, $(a_4 \geq b_4) \in U_0$. 4. $b_2 \in U_0$. 5. $a_2 = \emptyset$. 6. P { { 1, 2, 3 }, { 4 } }.
Equation: le terme soustractif manque.	42 - = 25	A - = B	1. A = a_1a_2; B = b_1b_2. 2. $a_1 \in U$. 3. $a_2, b_1, b_2 \in U_0$. 4. Vérifier: $0 < B < A$.

Explication de la notation.

Les lettres majuscules A, B, ... représentent des nombres.
Les lettres minuscules a, b, ... représentent des chiffres.
$(a < b) \in \{ \ldots\ldots\ldots\ldots \}$: choisir 2 nombres au hasard sans remplacement: a doit être plus petit que b.
{ H, V }: choisir une disposition horizontale ou verticale.
$U = \{ 1, 2, \ldots\ldots 9 \}$.
$U_0 = \{ \emptyset, 1, \ldots\ldots 9 \}$.
N_A = nombre de chiffres dans A.
N = nombre de chiffres dans chaque nombre du problème.
P { A, B, ... }: choisir une permutation des éléments dans l'ensemble (si l'ensemble consiste en indices, permuter ces éléments indexés).
$x \in \{ \ldots\ldots \}$: choisir au hasard une valeur pour x dans l'ensemble donné.
a, b, $c \in \{ \ldots\ldots \}$ = a, b, c sont choisis dans l'ensemble donné avec remplacement (tirage aléatoire simple).

1 W. HIVELY et al, *o.c.*, p. 281.

4. *Exemple de système de génération d'items pour la mathématique nouvelle au début de l'école primaire*[1]

En une série de sept tableaux de départ, les dimensions et les valeurs possibles des principales unités de matières sont synthétisées. Voici une de ces unités:

III. **A. Etude des qualités d'objets et d'ensembles d'objets.**

Point d'application	*A. Objets isolés*			*B. Ensembles d'objets*		
Nombre de propriétés envisagées simultanément / Type d'opération	1	2	plus de 2	1	2	plus de 2
Découverte des propriétés *statiques* (S) définies						
positivement						
négativement						
Découverte des propriétés *dynamiques* (D)						

Le tableau comporte trois dimensions:

- le point d'application de l'opération;
- le nombre de propriétés envisagées simultanément;
- le type d'opération.

Parallèlement au tableau axé sur la matière, on dresse aussi le tableau des comportements attendus des élèves. Les comportements sont déterminés par les caractéristiques de la situation stimulus proposée. Trois dimensions ont été retenues:

- le canal par lequel est fournie l'information de base sur laquelle l'élève devra travailler (M);
- le canal par lequel l'élève est invité à fournir sa réponse (E);
- le type de réponse attendu (0).

2 J. PAQUAY-BECKERS, *D'une philosophie de la compensation à une pédagogie de la maîtrise*, Laboratoire de Pédagogie expérimentale de l'Université de Liège, 1973 (ronéotypé).

TABLEAU IV: Dimensions comportementales.

Mode de réponse: E / O		Mode de présentation: M — Objet concret M_1	Dessin M_{2a}	Symbole M_{2b}	Signe M_{2c}	Enoncé verbal M_3	«Invente»
Objet concret E_1	Choix: 0_1						
	Production: 0_2						
Dessin E_{2a}	Choix: 0_1						
	Production: 0_2						
Symbole E_{2b}	Choix: 0_1						
	Production: 0_2						
Signe E_{2c}	Choix: 0_1						
	Production: 0_2						
Enoncé verbal: E_3	Choix: 0_1						
	Production: 0_2						

Enfin pour les différentes cases du tableau III A, on développe la matrice de tous les exercices possibles en combinant des dimensions comportementales du tableau IV.

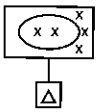
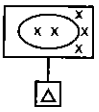
Considérons par exemple les cases B - S du tableau III A (cases hachurées): étude des propriétés statiques d'ensembles d'objets, définies positivement ou négativement.

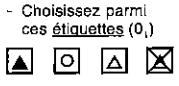
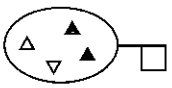
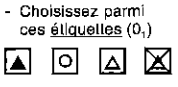
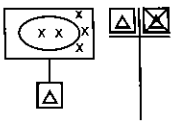
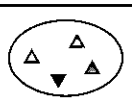
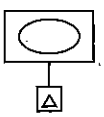
Chaque cellule de la matrice ainsi obtenue définit une classe de comportements équivalents. Certaines cellules de cette matrice sont supprimées soit parce qu'elles décrivent un comportement de simple copie (ex. M_1E_1) soit parce qu'elles se réfèrent à un comportement non introduit pour ce contenu (ex. utilisation de signes conventionnels pour l'étude des qualités d'ensembles d'objets isolés).

Toutes les cases sont exploitées par les maîtres pendant leur enseignement. Nous avons marqué d'un C les cases décrivant des comportements susceptibles d'être contrôlés par un test collectif papier-crayon. Les autres comportements doivent être contrôlés oralement pour chaque enfant individuellement. Le maître construit à cet effet des tableaux de dépouillement qu'il remplit au fur et à mesure du déroulement des activités de la classe.

OBJECTIF IIa :

- Etude des qualités statiques d'ensembles d'objets.
- Une seule qualité définie positivement ou négativement.

Mode de présentation / Mode de réponses E et O	Objets concrets : M_2	Dessins : M_{2a}	Symboles : M_{2b}	Signes conventionnels : M_{2c}	Enoncé verbal : M_3	Invente :
Objets concrets : E_1 Sélection : O_1 et production : O_2	Copie	Copie	 - Choisissez - parmi ces quelques blocs logiques (0_1) ou - dans tout votre matériel (0_2) les blocs (concrets) qui peuvent être rangés à la place des croix.	Non introduits à ce niveau	- Choisissez parmi ces ensembles de blocs (concrets) *l'ensemble des triangles* (0_1) - Construisez *un ensemble de triangles* (0_2) - Classez ces blocs (dessins) suivant *la couleur* (0_1)	*Invente un ensemble* que tu pourrais construire avec les blocs de notre boîte (0_2)
Dessins : E_{2a} Sélection : O_1 et production : O_2	Copie	Copie	 - Choisissez - parmi ces dessins de blocs (0_1) ou - dessinez (0_2) les blocs qui peuvent être rangés à la place des croix C	Non introduits à ce niveau	- Choisissez parmi ces ensembles de blocs (concrets) *l'ensemble des triangles* (0_1) - Construisez *un ensemble de triangles* (0_2) - Classez ces blocs (dessins) suivant *la couleur* (0_1) C	*Invente un ensemble* et dessine-le (0_2) C

Symboles : E_{2b} Sélection : 0_1 et production : 0_2	un ensemble de triangles : objets concrets (Dienes) (entourés d'une corde ou sous forme de tableaux) - Choisissez parmi ces étiquettes (0_1)  ou trouvez (0_2) une étiquette représentant chaque ensemble C	 - Choisissez parmi ces étiquettes (0_1)  trouvez (0_2) une étiquette représentant chaque ensemble C	 - Rangez les pièces du tableau dans le diagramme ou vice versa. C	Non introduits à ce niveau	- Choisissez parmi ces étiquettes (0_1) ou trouvez une étiquette (0_2) pour représenter *l'ensemble de tous les triangles de notre boîte* C	- Trouve une étiquette pour représenter *l'ensemble que tu as inventé* (0_2) C
Signes conventionnels E_{2a} Sélection : 0_1 et production : 0^2	Non introduits à ce niveau	Non introduits à ce niveau	Non introduits à ce niveau	Non introduits à ce niveau	Non introduits à ce niveau	Non introduits à ce niveau
Enoncé verbal : E_3 Sélection : 0_1 et production : 0_2	un ensemble de triangles : objets concrets (Dienes) (entourés d'une corde ou sous forme de tableaux) - Décrivez cet ensemble (0_2) - Ce petit carré rouge appartient-il à l'ensemble ? Pourquoi ? (0_1)	 - Décrivez cet ensemble (0_2) - Ce petit carré rouge appartient-il à l'ensemble 0_1 ? Pourquoi	 - Décrivez un objet qui appartienne ou n'appartienne pas à l'ensemble (0_2) - Est-ce qu'un petit carré rouge appartient à l'ensemble 0_1 ? Pourquoi ?	Non introduits à ce niveau	Copie	- *Invente un ensemble* et décris-le (0_2)

A première vue, pareille façon de procéder paraît difficile. L'expérience montre cependant qu'un instituteur bien initié au système finit par le considérer comme indispensable à la pratique éclairée de son action.

Il n'en reste pas moins que, dans le contexte général de cette méthode de travail, la coopération entre les centres de recherche et les écoles s'avère indispensable.

C. Conclusion.

Les démarches très « mécaniques » ou, au moins, très analytiques que nous venons d'envisager peuvent paraître fort éloignées du milieu vivant, d'une pédagogie fonctionnelle. Pour éviter tout malentendu, rappelons que le but poursuivi est surtout de suggérer des instruments *diagnostiques*. Ils ne préjugent en rien de la pédagogie mise en œuvre; quelle qu'elle soit, elle exige un contenu et des contrôles d'acquisition.

Nous aurions pu nous cantonner aux deux premiers exemples simples, parmi ceux qui précèdent. Nous ne l'avons pas voulu. Pourquoi laisserait-on supposer que la pégadogie échappe à la complexité croissante des autres disciplines ?

CHAPITRE 3

LA NOTATION

I. Un préambule indispensable: la courbe de Gauss

Il est impossible de discuter de l'évaluation des résultats sans s'appuyer sur quelques notions mathématiques. Nous n'oublions pas la promesse faite au début de cet ouvrage: l'arithmétique élémentaire nous suffira !

A. La courbe de Gauss, image de la probabilité.

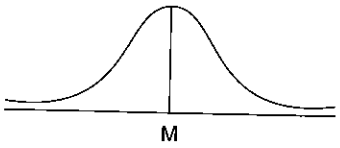
Un sac contient deux boules, en tout point semblables, sinon que l'une est rouge et l'autre blanche.

Dans ces conditions, chaque boule a une chance sur deux de sortir à chaque tirage aveugle.

Si nous nous livrons à ce jeu un très petit nombre de fois, il est possible que nous tirions plusieurs fois consécutivement la même couleur. A mesure que nous augmentons le nombre de tirages, cette probabilité diminue.

Imaginons 100 tirages consécutifs. Il est fort probable que le rouge sortira à peu près autant de fois que le blanc. Il est, par contre, fort peu probable que l'on tirera 100 fois consécutivement la boule rouge ou la boule blanche. Il est déjà un tout petit peu plus probable que l'on obtiendra 99 fois rouge et une fois blanc, plus probable déjà 98 fois rouge et 2 fois blanc, etc. Bref, la probabilité va augmenter jusqu'à 50 fois rouge et 50 fois blanc, puis diminuer progressivement pour arriver à aucune fois rouge et 100 fois blanc.

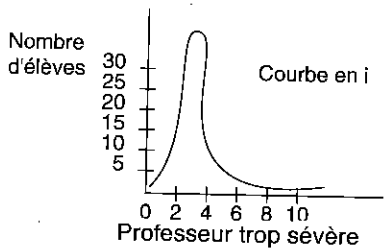
Dans un très grand nombre de tirages, ce mouvement ascendant-descendant correspond à une courbe revêtant la forme d'une cloche: c'est la fameuse courbe de Gauss, dont les deux moitiés sont symétriques par rapport à la moyenne arithmétique et dont les extrémités ne touchent jamais la ligne du zéro, la probabilité nulle n'existant qu'à l'infini.

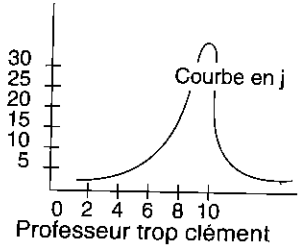


Cette *distribution*, dite *normale*, est à l'image de beaucoup de qualités humaines, telles qu'elles se répartissent dans des groupes nombreux, *pris au hasard*. Ainsi, parmi les habitants d'une grande ville, les hommes de taille moyenne sont les plus nombreux, tandis que les géants et les nains sont très rares; entre ces deux extrêmes, la population se distribue selon la courbe de Gauss.

Même l'erreur se soumet souvent à la loi normale: si l'on fait un très grand nombre de mesures, on verra fort probablement apparaître une erreur de grandeur moyenne, et une distribution allant de l'erreur infime à l'erreur maximum.

Dès que le hasard ne joue plus, la distribution se modifie. Un professeur peut, par exemple, donner une très grande majorité de mauvaises notes et fort peu de bonnes, ou le contraire. Il est possible que l'on obtienne alors une des deux courbes caractéristiques suivantes:





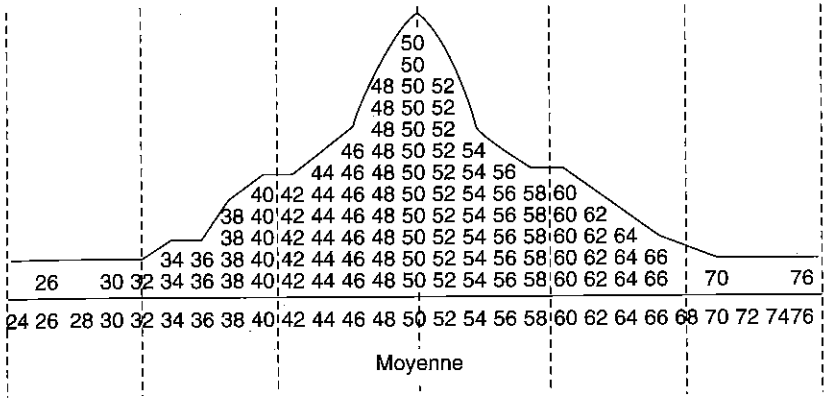
B. La courbe de Gauss, image des résultats de l'enseignement non individualisé.

Un professeur qui enseigne de façon non individualisée dans une classe où les élèves ne sont pas spécialement sélectionnés donne normalement à son cours un degré de difficulté adapté à la majorité du groupe. Si l'ajustement est correct, il y aura donc beaucoup de résultats moyens, peu de très bons et peu de très mauvais. La distribution de ces résultats s'approchera de la courbe gaussienne.

Pareil phénomène se produit, plus spontanément encore, dans les exercices échappant à la quantification rigoureuse, parce qu'ils mettent en jeu un ensemble complexe de facteurs.

Ci-dessous, on a disposé 100 résultats (imaginaires) obtenus à un examen, lui-même noté sur 100. On constate que 12 élèves ont obtenu 50 sur 100 (note moyenne), alors qu'un seul obtient la note la plus basse (26) et un seul la plus haute (76).

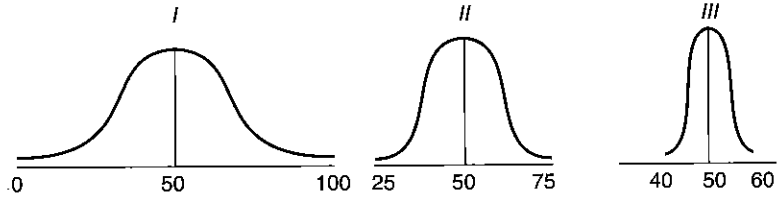
On observe aussi que la ligne correspondant à la répartition des résultats n'est pas une courbe en cloche parfaite, mais seulement son ébauche. Dans pareil cas (nous allons bientôt donner quelques précisions à ce propos), on suppose que, si le nombre de sujets avait été plus élevé, la courbe se serait polie et on considère la répartition comme «normale».



C. L'écart type ou sigma, indice précieux.

1. Signification.

La variation des 100 résultats que nous venons d'examiner aurait pu être soit plus grande, par exemple de 0 à 100, soit plus petite, par exemple de 40 à 60, tout en se distribuant toujours en courbe de Gauss. Nous aurions pu avoir:



Dans les trois cas, la moyenne est 50. Pourtant, il s'agit de situations très différentes.

La marge de variation des résultats est:

Courbe I : 100 — 0 = 100
Courbe II : 75 — 25 = 50
Courbe III: 60 — 40 = 20

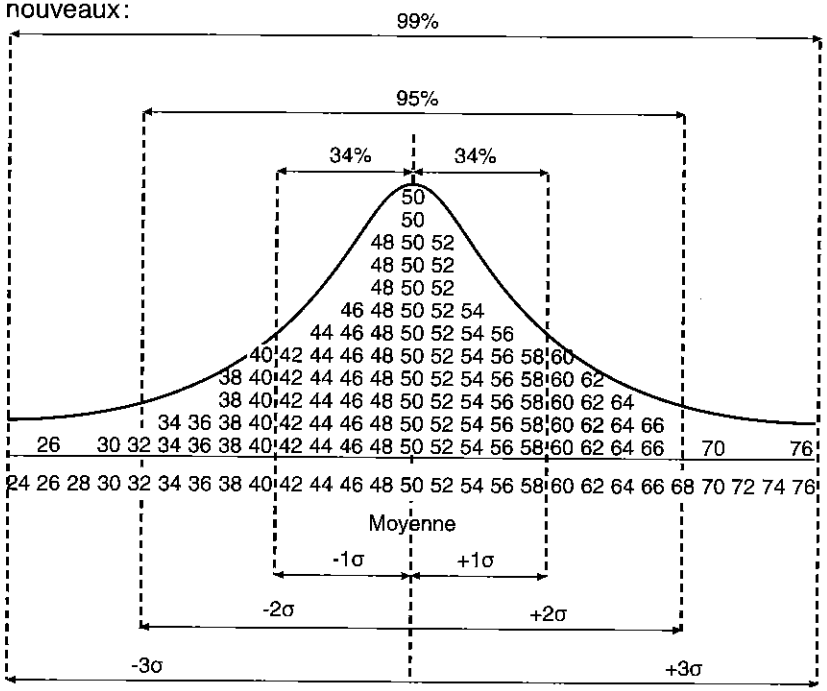
Sigma ou σ (symbole de l'écart type) est un indice facile à calculer (voir plus loin la méthode de calcul), qui nous indique immédiatement comment les résultats varient par rapport à la moyenne arithmétique des notes obtenues.

Pour les 100 notes prises en exemple, sigma = 9. Si l'on prend un sigma avant et après la moyenne, on obtient: 41 et 59. Comptez le nombre de résultats qui tombent entre ces deux limites, vous obtenez 68 notes, soit ici 68% de notes de tout le groupe.

Il en est toujours ainsi, dans une *distribution normale* : le sigma indique toujours la même proportion des résultats par rapport à l'ensemble:

1 sigma de part et d'autre de la moyenne = 68% des notes
2 sigma de part et d'autre de la moyenne = 95% des notes
3 sigma de part et d'autre de la moyenne = 99% des notes.

Voyons maintenant la distribution des résultats avec des yeux nouveaux:



Comme, à un sigma donné, correspond toujours la même surface de l'aire déterminée par la courbe, c'est-à-dire la même proportion des résultats, *cet indice nous fournit le moyen de comparer les résultats attribués par des professeurs différents, pour autant que ces résultats se répartissent normalement.* Nous montrerons bientôt comment.

2. Estimation rapide de la moyenne et du sigma.

a) *Problème.*

Le calcul rigoureux de la moyenne et de l'écart type (sigma) est relativement lourd. De plus, quand la distribution n'est pas normale, la moyenne arithmétique donne une idée fausse de la tendance centrale.

Un procédé simple permet d'éviter de longs calculs. Il ne fournit que des résultats *approximatifs.* Ils suffisent néanmoins dans presque tous les cas de la pratique courante à l'école[1].

Soit deux questions d'examens auxquelles 10 élèves ont répondu. Voici les notes attribuées, les moyennes et les écarts types calculés par la méthode classique.

Elève	*Question 1*	*Question 2*
1	39	32
2	33	28
3	25	32
4	22	28
5	26	27
6	18	31
7	23	33
8	13	27
9	57	35
10	45	36
	301	309
	Moyenne arithmétique: 30	Moyenne arithmétique: 31
	Ecart type:	Ecart type:
	$\sigma = \sqrt{\frac{\sum d^2}{N}} = \sim 13.$	$\sigma = \sim 3.$

Les calculettes peu coûteuses, *a fortiori* des ordinateurs domestiques, résolvent rapidement ces problèmes. Les illustrations qui suivent permettent cependant de mieux comprendre les démarches.

b) *Méthode simple de calcul.*

1° Représentation graphique des résultats

Notes	*Question 1*	*Question 2*
60		
58 56 54 52	.	
50		
48 46 44 42	.	
40		
38 36 34 32	. . .	: . .
30		
28 26 24 22	: :	: :
20		
18 16 14 12	. .	
10		

Marges de variation:

Question 1: 57 - 13 = 44
Question 2: 36 - 27 = 9

Les notes de la question 1 sont beaucoup plus dispersées. Elles influencent plus le classement final que celles de la question 2. Il est souvent utile d'ignorer les deux notes extrêmes: dans ce cas:

Marge Question 1 = 27
Question 2 = 8

Médians

= notes du milieu.
Ici, chiffre pair, donc moyenne entre 5e et 6e.

Question 1:
(25 + 26) : 2 = 25,5

Question 2:
(31 + 32) : 2 = 31,5

Pour la question 1, la différence assez nette entre la moyenne arithmétique et le médian (30 et 25,5) indique une distribution asymétrique des notes (ici: majorité sous le médian). Pour la question 2, la symétrie est bonne (Moy. 31 et Méd. 31,5).

Remarquons que, dans une distribution parfaitement normale, la moyenne et le médian coïncident: voir notre exemple.

2° Estimation de l'écart type (σ).

σ = $^3/_4$ de l'écart interquartile.

Donc:

a. Chercher le quartile supérieur, c'est-à-dire le milieu des notes au-dessus du médian.

Question 1 = 39
Question 2 = 33

b. Chercher le quartile inférieur.

Question 1 = 22
Question 2 = 28

c. Ecart interquartile.

Question 1 = 39 - 22 = 17
Question 2 = 33 - 28 = 5

d. Estimation du σ :

Question 1 = 17 x $^3/_4$ = 12 $^3/_4$
Question 2 = 5 x $^3/_4$ = 3 $^3/_4$

D. La concentration des résultats autour de la moyenne.

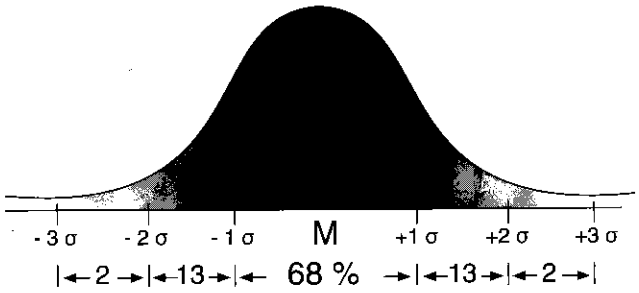
Reportons-nous de nouveau à la distribution des 100 notes prises comme exemple (p. 97). Nous observons que, plus nous nous rapprochons de la moyenne, plus les notes sont nombreuses:

12 notes sur 100 à la moyenne exacte
22 notes sur 100 entre 48 et 52
68 notes sur 100 entre 41 (- 1 σ) et 59 (+ 1 σ)

Imaginons qu'il s'agisse de compositions françaises. Alors que le professeur pouvait noter de 0 à 100, presque un quart des notes se situent entre 48 et 52, soit à 4 points de distance.

Que signifie une différence de 4 points sur 100 en dissertation française? Il est à peu près certain que, si nous faisons juger le même travail par dix professeurs différents, les écarts seront presque toujours supérieurs à 2. Or, si c'est la moyenne qui constitue la limite de l'échec (ici, elle coïncide avec la moitié de l'échelle totale), on peut dire, sans crainte de se tromper que, pour 22% du groupe d'élèves pris comme exemple, l'échec ou la réussite dépendra du pur hasard.

Voici, en dégradé, une image (très approximative) de la densité de population par rapport à la moyenne:



E. Courbe de Gauss voulue par les maîtres.

La courbe de Gauss est un excellent instrument de classement puisqu'elle permet d'identifier les meilleurs (par exemple entre 25 et 35) et les moins bons.

Dans le cas d'un concours où de nombreux candidats postulent un petit nombre de places, on peut, par exemple, choisir parmi ceux qui se situent entre + 1 σ et + 3 σ (partie qui échappe à la forte concentration autour de la moyenne). De même, si l'on souhaite classer les élèves - ce que l'on a fait trop exclusivement dans le passé, mais qui reste tout de même utile dans certains cas, pour permettre à chacun de se situer -, la distribution normale est certainement utile.

Comme il ne s'agit plus, cette fois, de groupes choisis au hasard, mais de sujets spécialement entraînés, le professeur *crée artificiellement* les conditions nécessaires à une répartition gaussienne. Comment ? En dosant des *items* selon leurs indices de difficulté (et d'efficacité).

F. Comment savoir si une distribution est normale ?

Tout ce que nous venons de dire à propos de la courbe de Gauss et de l'écart type n'est valable que si les résultats se distribuent *normalement*. Avant d'engager toute opération, il importe donc de vérifier si cette condition est remplie.

Il existe des procédés mathématiques rigoureux pour atteindre ce but, mais, de nouveau, nous nous contenterons ici d'une simple approximation graphique : l'*histogramme*.

Le procédé est simple :

1. Classer les résultats.

   On peut les ordonner du plus grand au plus petit, ou inversement.

   Voici ce qu'on obtient en classant les 100 résultats que nous avons déjà pris comme exemples :

Notes	*Nombre d'élèves ayant obtenu ces notes (effectifs) (f)*
26 I	1
30 I	1
32 I	1
34 II	2
36 II	2
38 IIII	4
40 𝍸	5
42 𝍸	5
44 𝍸 I	6
46 𝍸 II	7
48 𝍸 𝍸	10
50 𝍸 𝍸 II	12
52 𝍸 𝍸	10
54 𝍸 II	7
56 𝍸 I	6
58 𝍸	5
60 𝍸	5
62 IIII	4
64 III	3
66 II	2
70 I	1
76 I	1
	N = 100

Mais, si les notes sont nombreuses, ce procédé n'est ni rapide, ni pratique. Il est bien plus aisé de constituer des classes.

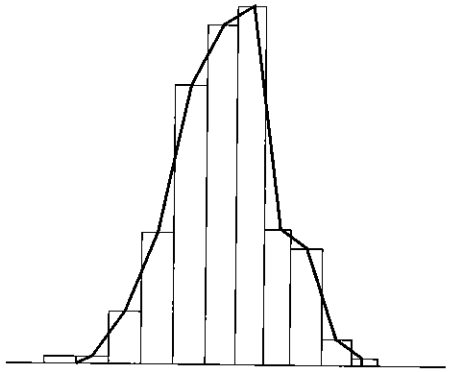
1° Calculer la marge de variation entre les deux notes extrêmes : 76 - 26 = 50.
2° Diviser cette marge par 15[1] : 50 : 15 = 3,33.
3° Choisir comme intervalle de classe un des deux nombres impairs les plus proches : 3 ou 5. On choisit ici 5, vu le petit nombre de notes.
4° Placer la note supérieure au milieu de l'intervalle supérieur : note supérieure : 76 ; la classe supérieure est donc : 74-75-76-77-78.

1 Ce nombre est arbitraire. La pratique montre que, dans la plupart des cas, il conduit à une bonne répartition.

	Classes	Effectifs (f)	
1	74 - 78	I	1
2	69 - 73	I	1
3	64 - 68	𝍸	5
4	59 - 63	𝍸 IIII	9
5	54 - 58	𝍸 𝍸 𝍸 III	18
6	49 - 53	𝍸 𝍸 𝍸 𝍸 II	22
7	44 - 48	𝍸 𝍸 𝍸 𝍸 III	23
8	39 - 43	𝍸 𝍸	10
9	34 - 38	𝍸 III	8
10	29 - 33	II	2
11	24 - 28	I	1
11 classes	intervalle (i) = 5	Nombre de notes (N) = 100	

2. *Dessiner l'histogramme.*

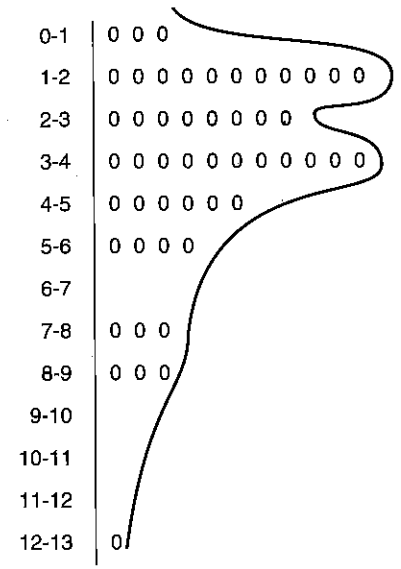
Dans la partie «effectifs» du tableau ci-dessus, le pointage a été fait avec beaucoup de soin. Il suffit de tenir la page horizontalement pour voir apparaître le squelette du graphique appelé histogramme: il reste à tracer les rectangles.



Si l'on joint les milieux des bases supérieures, on obtient une ligne (polygone de fréquences) qui, ici, est une ébauche dont on se contentera pour considérer que la distribution est fort probablement normale. Pourquoi? Parce que le tracé n'a qu'*un sommet*, à peu près *centré*, à partir duquel les quantités *décroissent assez régulièrement de part et d'autre.*

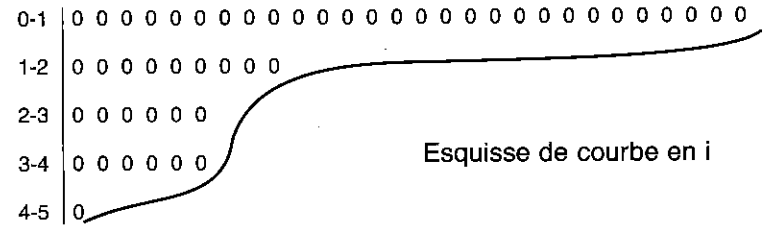
Par contre, les deux distributions ci-dessous ne seraient pas considérées comme *normales.*

(Distribution estimée anormale)



Esquisse de courbe décentrée et à deux sommets (courbe bimodale).

(Distribution estimée anormale)



Esquisse de courbe en i

II. La notation subjective : l'échelle d'évaluation

A. Introduction.

Par notation subjective, nous entendons toute évaluation faite par les maîtres, d'une façon globale, soit immédiatement lors de l'observation de la performance des élèves, soit après avoir eu une vue d'ensemble sur des scores particuliers. De façon plus générale, est subjective toute note à laquelle on arrive par une démarche non automatisable.

Le degré de subjectivité varie considérablement. Tantôt elle joue presque sans limite (par exemple : « c'est bien - c'est très bien... »); tantôt elle cède presque le pas à la démarche objective (exemple : échelle dont les degrés sont définis avec une précision ne laissant pratiquement plus de place à l'interprétation personnelle).

Dans tous les cas, l'évaluation subjective se fait par comparaison à des critères externes ou internes. Le notateur conclut à l'égalité ou à l'inégalité par rapport au critère. Il opère donc un classement.

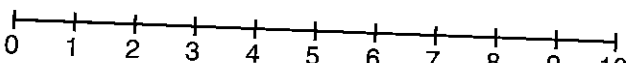
C'est pourquoi nous pouvons limiter cette partie à l'étude des échelles d'évaluation.

L'échelle d'évaluation aide à *ordonner* des individus ou des objets par rapport à une qualité qu'ils possèdent à un degré plus ou moins élevé.

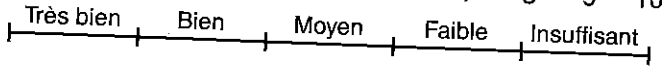
Elle représente un *continuum* qui peut théoriquement aller d'un minimum absolu à un maximum absolu, mais qui, pratiquement, n'est presque toujours qu'une fraction, un segment de cet absolu. De l'excellence ... à ... l'insuffisance, du très petit ... au très grand, etc.

L'échelle d'évaluation, graphique ou non, est :

soit *numérique* :



soit *verbale* :



L'échelle d'évaluation est l'instrument le plus fréquemment employé à l'école... et le plus mal connu. On peut dire, sans crainte de se tromper, qu'il a littéralement vicié une large partie des examens, à tous les niveaux de la scolarité, et probablement fait le malheur d'un nombre considérable d'individus.

C'est pourquoi il importe d'en connaître la véritable nature et surtout les limitations.

B. Nature et faiblesse des échelles d'évaluation.

L'échelle d'évaluation est une *échelle ordinale.*

1) *Elle n'a ni zéro ni maximum naturels ou absolus*; elle commence et finit n'importe où, au choix de l'utilisateur. Par conséquent, même s'ils évaluent un même phénomène, du même point de vue, les juges n'ont jamais des échelles rigoureusement les mêmes; les différences peuvent être considérables, et le sont souvent.

   Exemple : *Connaissance de l'anglais, deuxième langue.*

   Théoriquement, le zéro pourrait être le moment où l'élève ne connaît pas encore un seul mot de la langue. Toutefois, on juge rarement à partir de ce point à l'école.
   Quant à la connaissance totale, absolue, elle n'existe évidemment pas.

   Imaginons deux professeurs d'anglais jugeant, sans se concerter, deux classes où chaque élève a fait un court exposé oral. Fort probablement, les deux extrémités des échelles d'évaluation utilisées indépendamment par les deux professeurs seront plus définies par l'élève jugé le plus fort et l'élève jugé le plus faible, dans son groupe, que par une évaluation abstraite de la quantité d'anglais qui devrait être connue au niveau pédagogique considéré.

2) *Les degrés ne sont pas de même grandeur à l'intérieur d'une même échelle.*

   Il est, par exemple, impossible de démontrer que la distance séparant la bonne de la très bonne connaissance de l'anglais est la même que la distance séparant la connaissance moyenne de la bonne.

   Si les degrés sont : A, B, C, D, E, on ne peut donc pas dire que A - B = B - C = Ø.

3) *Des degrés correspondants d'une échelle à l'autre (pour une même branche) ne sont pas de même grandeur.*

   Bon en anglais, pour un professeur, n'est pas quantitativement égal à bon en anglais pour un autre : rien ne peut établir pareille égalité de facon mathématique.

4) *Des échelles portant sur des branches différentes ne sont pas comparables.*

   Pour les élèves d'une même classe d'école primaire où un seul maître enseigne toutes les branches, être bon en mathématique ne recouvre pas les mêmes phénomènes qu'être bon en français. Les échelons ne sont certainement pas comparables sur le plan quantitatif.

De tout ceci, il résulte que les échelles d'évaluation ne permettent aucune opération arithmétique. Si on note la composition française et l'anglais deuxième langue de 0 à 10, rien ne permet d'additionner les deux résultats pour évaluer un savoir total. C'est pourtant ce que l'on fait, depuis des décennies, dans nos écoles. Il est tout aussi suspect, sinon plus, d'utiliser des échelles numériques différentes pour les différentes branches, en fonction d'une pondération empirique: de 0 à 120 pour le français, de 0 à 80 pour les sciences, etc.

Si le système a pu fonctionner, c'est qu'il a, de toute façon, permis d'identifier les meilleurs (ceux qui se classent en tout dans les premiers) et les plus faibles (derniers en tout). Mais que d'injustices envers ceux qui ont des résultats en dents de scie!

C. **Utilité.**

Malgré ses limitations, l'échelle d'évaluation est, dans bien des cas, le seul instrument dont nous disposons pour concrétiser notre jugement sur des **comportements** humains **complexes**. Comment évaluer l'élégance du langage, la beauté d'un mouvement sinon de façon globale?

En fait, elle permet d'apporter, dans un dossier scolaire, des éléments *qualitatifs*, aussi objectifs que possible, *indispensables* compléments des résultats *quantitatifs* apportés par les examens objectifs ou les tests.

D. **Construction.**

Beaucoup d'éducateurs se laissent abuser par la facilité apparente avec laquelle on construit une échelle.

En réalité, il n'est pas rare que la mise au point de pareil instrument exige de longs mois de travail, nécessaires à la clarification théorique, et appelle l'utilisation de techniques raffinées comme l'analyse factorielle. Celle-ci aide notamment à déterminer dans quelle mesure plusieurs échelles évaluent les mêmes choses sous des appellations différentes.

Toutefois les maîtres peuvent procéder plus simplement dans leur pratique quotidienne.

1. **Combien de degrés?**

Théoriquement, une échelle d'évaluation peut compter une quantité infinie d'échelons. En pratique, les juges ne classent avec quelque sûreté que si l'on n'utilise qu'un petit nombre de degrés: 3, 5, 7 ou 9.

La littérature expérimentale suggère que l'on s'en tienne *le plus souvent* à 5 échelons. Pareilles échelles permettent des classements relativement sûrs et fidèles, à condition que chaque échelon soit clairement défini et les notateurs bien formés[1].

Pratiquement,

1° On classe d'abord en trois catégories: les meilleurs, les plus faibles, les autres;

2° On répartit ensuite les «autres» en bons, moyens et faibles.

Excellent	Très bien	Bien	Assez bien	Faible ou échec
A	B	C	D	E

Remarquons que l'échelon C est à cheval sur le centre de l'échelle, c'est-à-dire sur la moyenne arithmétique. Comme nous l'avons vu, à propos de la courbe de Gauss, une forte concentration se produit souvent autour de ce point. Choisir un nombre pair d'échelons créerait un centre dans l'échelle, devant ou derrière lequel plusieurs élèves basculeraient au hasard.

Dans le cas d'une distribution normale, il n'est pas exclu que ce verdict du hasard affecte 20, 30, voire 40% des élèves.

2. **Définir l'objet de l'évaluation.**

Demander, par exemple, d'apprécier le «courage» des membres d'un groupe donné - sans autre précision - conduirait à des réponses presque entièrement dépourvues d'intérêt. Quel ou quels sens les observateurs auront-ils donnés au mot «courage»: ardeur, volonté, zèle, persévérance, bravoure, fermeté, stoïcisme?

Et même si nous précisons que par «courage», nous entendons la fermeté devant le danger, ferons-nous aisément la distinction entre l'intrépidité et la témérité?

---

1 Il serait cependant dangereux d'accorder une valeur dogmatique à ces recommandations, car, comme l'a souligné Reuchlin, on n'a pas encore pu évaluer avec certitude les dangers respectifs d'une échelle comptant un petit nombre de degrés (c'est-à-dire de catégories larges augmentant le poids de chaque erreur d'évaluation), et d'une échelle divisée plus finement (catégories étroites). Kaufman, cité par Noizet et Caverni (o.c. p. 53), «suggère que le nombre de catégories à utiliser pour évaluer des copies devrait varier selon la discipline concernée, il devrait être moins élevé pour la composition française que pour l'épreuve de mathématique». Le débat fondamental reste ouvert.

Voir M. REUCHLIN, in M. DEBESSE et G. MIALARET, *Traité des Sciences Pédagogiques*, tome 4, Paris, P.U.F., 1974 p. 216.

J. KAUFMANN, Note sur les problèmes de métrique en matière de notation scolaire, in *Le Travail humain*, 38, 1975, 130-148.

Pour pallier les imprécisions de notations telles que très bien, moyen, etc., il est nécessaire d'ajouter à l'échelle une description aussi précise que possible du trait ou de l'objet à apprécier, et d'illustrer la définition par des situations types. Définir non par des formules abstraites, mais par des comportements précis est une condition sine qua non de validité.

Voici comment Schonell[1] présente le trait «Confiance en soi»:

Confiance en soi

Extrêmement confiant en soi. Presque trop sûr de lui.	Très confiant en ses propres forces.	Confiant.	Manque de confiance. Timide.	Manque extrême de confiance. Dépend des autres. Décline les responsabilités.

Description du trait.

«Sous sa forme positive, cette qualité est marquée par les manifestations suivantes: l'individu compte sur lui-même, est capable de faire face aux difficultés, a de l'assurance, est indépendant et prêt à assumer des responsabilités.

«L'enfant qui a confiance en lui-même essaie d'avancer avec le minimum d'assistance; celui qui manque de confiance doit être aidé constamment. Le premier aime de voir ce qu'il est capable de construire et de produire quand il a reçu des instructions claires; le second veut qu'on lui mâche la besogne, qu'on l'aide durant toute la phase de la réalisation.»

Situations types.

1° A-t-il peur de l'obscurité?
2° Peut-il prendre soin de lui-même et de ce qu'il possède ou faut-il que quelqu'un soit tout le temps à ses côtés?
3° Voyage-t-il seul en tram ou en bus? (pour enfants de plus de 9 ans)
4° Parle-t-il librement à des visiteurs inconnus?
5° Est-il bon dans les jeux? Sait-il nager?
6° Est-il à l'aise et répond-il avec assurance aux examens oraux?
7° Lit-il bien, dramatise-t-il bien un texte devant la classe?
8° S'attaque-t-il bien à des tâches nouvelles ou pose-t-il constamment des questions à ses compagnons et à ses maîtres?

1 F.J. SCHONELL, *Backwardness in the Backward Subjects*, cité par F. WARBURNE, Measurement of Personality (*Educational Research*, novembre 1961, p. 9).

Définir les degrés de l'échelle.

Exemple: Organisation de l'enseignement de la lecture.

0	1	2	3	4	5
Médiocre En lecture, tous les élèves suivent la même progression. Pas de travail par groupes.	Assez bien Cf. 1. Mais parfois un élève très lent reçoit un peu moins de travail que les autres.	Moyen Constitution de 2 ou 3 groupes, selon les aptitudes en lecture. Peu de flexibilité dans le groupement.	Très bien Groupement selon les aptitudes. Flexibilité.	Excellent Groupement après étude approfondie des aptitudes et des difficultés rencontrées. Grande flexibilité	

On aura remarqué que ce dernier exemple combine les échelles graphique, numérique et descriptive.

Conséquences d'une définition insuffisante des degrés.

On a pu croire qu'abandonner des échelles numériques (en cent points ou autres) pour des échelles littérales (A, B, C, D, E) ou verbales (très bien, bien, moyen, faible, insuffisant) assurait une meilleure concordance entre les évaluations. Il n'en est rien si les degrés ne sont pas bien définis. En voici, parmi bien d'autres, trois témoignages expérimentaux.

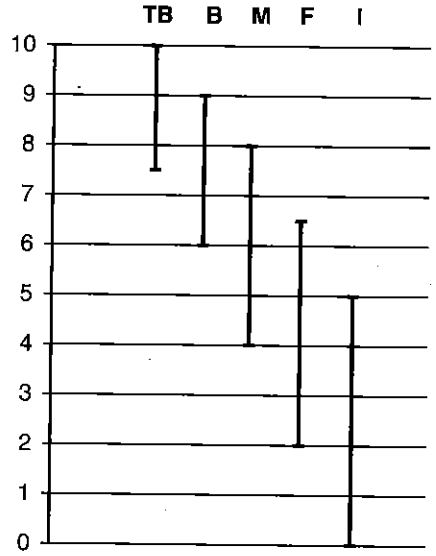
Lors de l'introduction de l'échelle verbale à cinq degrés dans l'enseignement secondaire rénové belge, dans les années 60, les professeurs de français et les professeurs de mathématiques[1] ont indiqué comment ils procédaient pour noter selon le nouveau système. La plupart attribuaient mentalement une note chiffrée sur dix puis convertissaient en évaluation verbale. Il a ainsi été possible de voir dans quelle mesure les notes verbales recouvraient des évaluations concordantes.

1 *Enquête sur les procédures de notation*, Laboratoire de Pédagogie expérimentale de l'Université de Liège, 1971 (document ronéotypé).

Voici d'abord les extrêmes trouvés pour une même appréciation:

Très bien:	limite inférieure:	7,5 - 9
Bien:	limite supérieure:	7,5 - 9
	limite inférieure:	6 - 7,5
Moyen:	limite supérieure:	6 - 8
	limite inférieure:	4 - 7
Faible:	limite supérieure:	4 - 6,5
	limite inférieure:	2 - 5
Insuffisant:	limite supérieure:	2 - 5
	limite inférieure:	0 - 2.

Une représentation graphique de ces observations est éclairante:



On constate donc que, selon les professeurs:
- une note de 7,5 correspond à très bien, bien ou satisfaisant;
- une note de 6 correspond à bien, satisfaisant ou faible;
- une note de 4 correspond à satisfaisant, faible ou insuffisant;
- une note de 2 correspond à faible ou insuffisant.

F. Bacher rapporte une autre expérience tout aussi frappante[1]. Demangeon et Larcebeau ont comparé, à propos des mêmes travaux de français et de calcul, l'accord entre professeurs selon qu'ils utilisaient des notes chiffrées ou quatre appréciations verbales: très bon, bon, moyen, faible.

Les discordances pour un même devoir sont:
- pour les notes chiffrées:
  - d'environ la moitié du maximum pour le calcul;
  - d'environ un quart du maximum pour le français.
- pour les appréciations verbales:
  - *Très bon* est attribué selon les professeurs:
    pour le calcul à 15 à 23% des élèves;
    pour le français à 0 à 30% des élèves.
  - La même copie reçoit la même appréciation:
    pour le calcul par 42% des correcteurs;
    pour le français par 16% des correcteurs.
  - La même copie reçoit:
    trois appréciations différentes
    dans 8% des cas en calcul,
    dans 20% des cas en français.

Les quatre appréciations possibles, en calcul: dans trois cas.

Passer des notes chiffrées aux échelles A, B, C, D, E ou très bien, bien, etc., ne garantit pas en soi une meilleure évaluation.

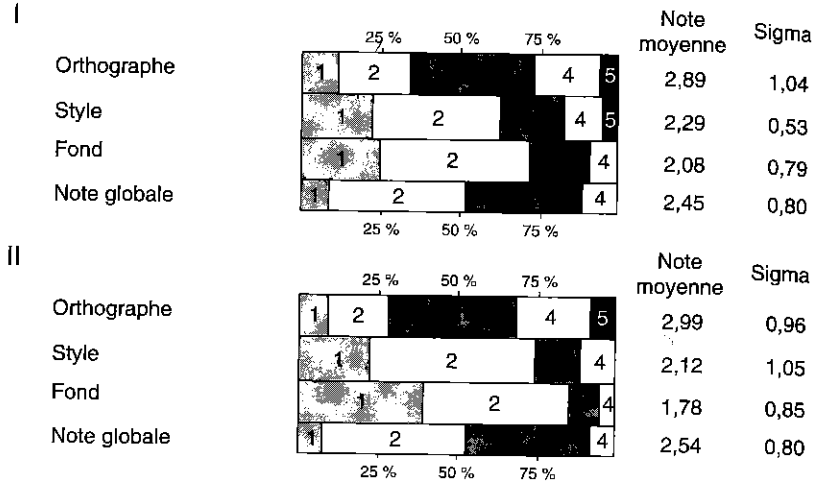
Adoptant les échelles descriptives imposées par le Ministère autrichien de l'Education nationale, Weiss[2] a invité 92 instituteurs à noter deux compositions d'élèves du niveau où ils enseignaient (4e primaire) et 272 instituteurs à noter deux travaux d'arithmétique (problèmes: 4e et 5e années).

Les diagrammes ci-dessous montrent combien les notations sont divergentes, pour un même travail (dispersion des notes sur toute la longueur de l'échelle).

1 M. DEMANGEON et S. LARCEBEAU, Une expérience de correction multiple, in *BINOP*, 1958. 14, 131-156. Cité par F. BACHER, *La docimologie*, o.c., p. 36.
2 R. WEISS, o.c.

NOTATION DE DEUX COMPOSITIONS

Notation de deux rédactions



Notation de deux travaux d'arithmétique

On le voit, il ne suffit pas d'adopter une nouvelle échelle en remplacement du système de notes traditionnel pour que tous les problèmes soient résolus. Loin s'en faut!

E. **Utilisation.**

1) *Combien d'élèves par échelon?*

Premier cas: Elève comparé à lui-même (évaluation critérielle).

Les échelons marquent les degrés à franchir dans l'ascension vers un savoir, un savoir-faire ou un savoir-être. Le maître a défini un objectif à atteindre. Dans un enseignement non individualisé, cet objectif est commun à la majorité, sinon à la totalité des élèves de la classe. Il doit évidemment être choisi de telle sorte que tous les élèves puissent se dépasser. Ce sera donc un objectif large, ce qui risque de nuire à sa précision au point de le rendre inopérant.

Il est clair que nous agissons actuellement comme si, au début de l'année scolaire, tous les élèves se trouvaient approximativement à un même endroit de l'échelle du savoir. On suppose donc que, l'année précédente, tous ont atteint le degré supérieur de l'échelle précédente.

Or, nous le sentons bien, la réalité est différente et la recherche confirme notre sentiment:

a) *Marges de variation de l'âge mental.*

Lors de la révision du test d'intelligence de Binet-Simon par Terman et Merrill (1937), on a constaté, aux Etats-Unis[1]:
- qu'en première année primaire, l'âge mental variait de 4 à 8 ans;
- qu'en sixième année primaire, l'âge mental allait de 8 à 16 ans;
- que, dans l'enseignement secondaire, une marge de variation de 8 à 10 ans n'était pas exceptionnelle.

b) *Marges de variation du rendement scolaire.*

Il suffit d'observer combien les normes des tests de connaissances, étalonnés par année scolaire, se chevauchent, pour prendre conscience de l'ampleur des marges de variation.

S'appuyant sur ses propres recherches et sur celles de Lindquist, Cornell, Learned et Wood, W.W. Cook[2] fournit les indications suivantes:

- Pour la compréhension de la lecture, pour le vocabulaire, les sciences, la géographie et l'histoire, la marge de variation du rendement est:
  - En 1re primaire: de 3 à 4 ans.
  - En 4e primaire: de 5 à 6 ans.
  - En 6e primaire: de 7 à 8 ans.
- Pour le raisonnement arithmétique et le calcul:
  En 6e primaire: de 6 à 7ans.
- En culture générale (sciences, littérature, arts, histoire, géographie et langue maternelle), au niveau universitaire:
  - 28% des étudiants de 4e année sont inférieurs à la moyenne des étudiants de 2e année;
  - 10% des étudiants de 4e année sont inférieurs à la moyenne des élèves de fin d'enseignement secondaire.

1 Cf. O. McNEMAR, *The Revision of the Stanford-Binet Scale*, Boston, Houghton-M., 1942.
2 W.W. COOK, The Functions of Measurement in the Facilitation of Learning, in E.F. LINDQUIST, Ed., *Educational Measurement*, Washington, A.C.E., 1961. 4e éd., pp. 3-47.

Semblables observations - qu'il ne faut pas prendre au pied de la lettre, mais dont il faut retenir la tendance - pourraient être multipliées à volonté.

Elles prouvent qu'aussi longtemps que nos écoles fonctionneront selon le système de classes rigides, l'évaluation continue allant de pair avec le souci de permettre à chaque élève de progresser à son allure propre, de se dépasser toujours, d'aller aussi loin qu'il le peut, sera un leurre.

Toutefois, pour des raisons pédagogiques, technologiques et économiques, l'enseignement intégralement individualisé ne se généralisera pas. Seul l'enseignement semi-individualisé offre une solution réaliste, immédiatement applicable. Dans une école semi-individualisée, des groupes homogènes, par branches principales, se substituent à la classe. Ainsi, un élève doué en langue étrangère peut travailler à un niveau A pour cette branche, mais se joindre au niveau D pour les problèmes d'arithmétique[1].

Afin de ne pas alourdir ce texte, nous avons reporté en annexe une description détaillée d'une école fonctionnant depuis longtemps selon ce principe.

Deuxième cas: Elèves comparés entre eux (évaluation normative).

Même si l'on essaie d'amener chacun à un niveau élevé, des différences suffisantes subsisteront probablement pour faire de nouveau apparaître une distribution gaussienne (ou, du moins, son ébauche), où la moyenne se sera simplement déplacée vers le haut et où la marge de variation sera étroite. Même parmi les champions, une hiérarchie existe. Mais, dans certains cas, elle ne s'exprime qu'en fractions de secondes.

Si l'on veut vraiment classer, il semble donc justifié d'essayer de peupler chaque échelon selon des pourcentages qui s'approchent de la distribution normale.

Dans des petits groupes, il n'est évidemment pas possible de respecter strictement cette proportion.

1 Ce n'est pas par hasard que nous parlons de *problèmes d'arithmétique* et non d'arithmétique, en général. Dans une branche complexe comme celle-ci, un élève peut être bien doué en calcul, mais faible en problèmes. La recherche contemporaine montre que si l'on ne groupe pas assez finement, les marges de variation entre élèves restent considérables.

n aura, par exemple:

			La distribution *normale* serait:
Excellent:	1	5%	2,5%
Très bon:	4	20%	13,5%
Bons (moyens):	10	50%	68 %
Faibles:	4	20%	13,5%
Très faible:	1	5%	2,5%
	20	100%	100 %

Il faut y insister, se forcer à respecter pareille répartition (*distribution forcée*) n'implique pas nécessairement un échec pour un certain nombre d'élèves. Toutes les notes sont *relatives* les unes par rapport aux autres: «très faible» peut, dans certains groupes forts, se situer au-dessus de la note d'échec. C'est d'ailleurs pourquoi *il vaut mieux adopter des lettres de classement* plutôt que les notes chiffrées.

2) *Lutter contre la contamination et la tendance centrale.*

Outre les dangers de stéréotypie et d'effets de halo que nous décrivons, page 47, un autre phénomène de contamination, de nature plus mécanique, guette aussi les juges. S'ils utilisent consécutivement, pour un même élève, une série d'échelles orientées dans le même sens (par exemple du positif au négatif ou du meilleur au plus faible) et peut-être même présentées sur une même page, les notations tendent à se grouper dans un même côté.

On conseille de faire figurer chaque échelle sur une page différente, de tirer au sort le sens de présentation de chacune.

Enfin, il est bon de ménager un intervalle assez long après l'évaluation de chaque qualité d'un même sujet et de faire évaluer une même qualité par autant de juges que possible.

Beaucoup de juges ont aussi tendance à grouper leurs notes vers le centre de l'échelle: crainte de surévaluer ou de sous-évaluer un élève, peur de supporter la responsabilité d'un échec. La règle de la distribution forcée rend ici d'utiles services. Mais, pour être réellement efficace, elle doit pouvoir s'appuyer sur une description précise et nuancée des différents échelons.

On ne perdra pas non plus de vue que la distribution gaussienne est pédagogiquement contre nature (voir dans la cinquième partie: Le mythe de la courbe de Gauss).

Quand les évaluations par plusieurs juges doivent être combinées, les professeurs qui se laissent séduire par la solution de facilité qu'est la note centrale doivent savoir qu'ils abdiquent une large partie de leur influence en faveur de leurs collègues qui osent utiliser l'ensemble de l'échelle.

F. Bacher remarque avec raison que la tendance centrale se manifeste surtout pour les épreuves les plus difficiles à noter « qui sont peut-être aussi celles dans lesquelles les aptitudes du candidat se manifestent le mieux». Pour se convaincre de la pertinence de cette remarque, il suffit de comparer les résultats pour une dictée (où un barème fixe de pénalisation ne laisse guère de place à la subjectivité du notateur) aux résultats attribués à un compte rendu de lecture. Guerbet-Seux et Reuchlin, cités par Bacher, observent un cas où le poids réel de la dictée augmente de plus de 50% et où celui des comptes rendus diminue d'autant, lorsque ces deux notes se combinent dans l'évaluation d'un même élève.

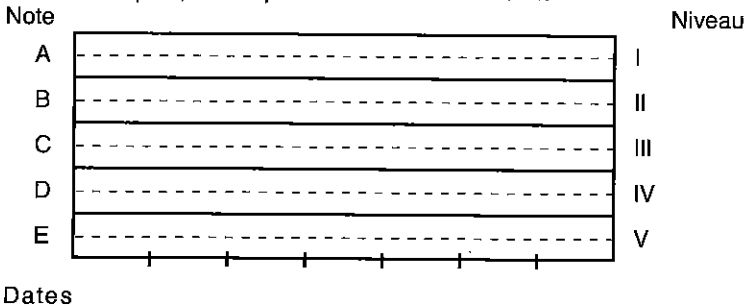
F. **Comment synthétiser les évaluations.**

Au terme d'une période scolaire, il importe de synthétiser les évaluations (que l'on souhaite aussi nombreuses que possible), faute de quoi maîtres et élèves ne sauront pas les interpréter.

En revenir au simple calcul de la moyenne risquerait de ruiner les efforts de rigueur précédents. Car des évaluations ordinales, théoriquement du moins, ne permettent ni addition, ni soustraction.

Une méthode simple permet de situer un élève dans chaque branche, puis pour l'ensemble des branches.

Au départ, on dispose du tableau suivant:



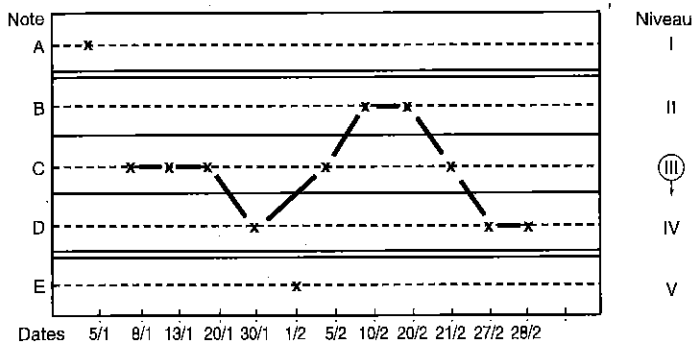
1 F. BACHER, La Docimologie, in M. REUCHLIN, *Traité de psychologie appliquée*, 6. Paris, P.U.F., 1973, p. 33.

Un professeur utilise autant de tableaux semblables qu'il souhaite distinguer d'aspects dans son cours: dictée, grammaire, orthographe, rédaction, ...

Au fur et à mesure des évaluations, on porte une croix dans le tableau, en indiquant la date au bas.

Au moment de faire la synthèse, on commence par barrer (sans l'effacer!) la croix correspondant à la note la plus haute et celle qui représente la note la plus basse, afin d'éliminer des notes qui pourraient être accidentelles.

Supposons que l'on obtienne le tableau ci-dessous.



Interprétation.

1) Les résultats se concentrent sur trois bandes: II, III, IV. Ils ne sont donc ni excellents, ni faibles ou insuffisants. La zone de concentration est marquée par des traits doubles.

2) Le plus grand nombre de notes (le mode) se situe en bande III. C'est la tendance dominante.

3) Comme la bande la plus peuplée, après la III, est la IV, une flèche sera placée, dans cette direction, sous la III.

4) En réunissant les croix, on obtient un profil qui fait apparaître:

   a) Une assez bonne homogénéité des résultats. Si l'élève se promenait anarchiquement sur les 5 bandes, il faudrait voir s'il est seul dans ce cas (cause à déterminer au niveau de l'élève), ou si la majorité de la classe se trouve dans la même situation (voir si les travaux du professeur ne sont pas mal ajustés).

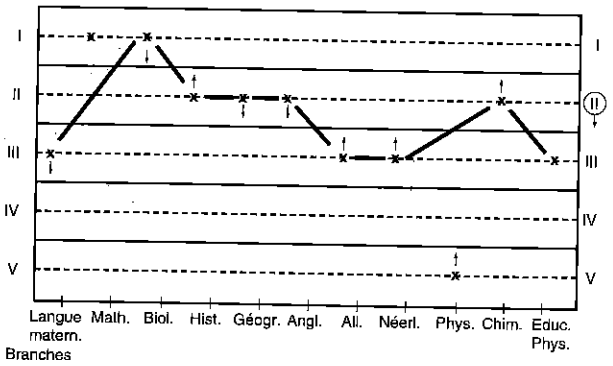
   b) Une descente en fin de période. Elle n'est pas très accusée: elle peut s'expliquer par la fatigue, par une maladie bénigne, ...

On le voit, il ne suffit pas de constater une situation traduite par le profil, il faut tâcher d'en découvrir le pourquoi. Ainsi, les commentaires, oraux ou écrits, des maîtres viennent-ils éclairer et guider l'élève et ses parents.

Remarquons, par ailleurs, que ce système ne peut fonctionner si le professeur se borne à deux ou trois évaluations; encore moins avec une seule! Cette fois, c'est le comportement de certains maîtres obnubilés par la quantité de matière à enseigner ... ou allergiques au long travail de notation des travaux, que l'introduction d'une nouvelle forme d'évaluation devrait heureusement améliorer.

TABLEAU POUR L'ENSEMBLE DES BRANCHES

PERIODE DU ..... AU ..... 19 .



Interprétation.

Quatre notes d'ensemble se trouvent en bande II et quatre en bande III. C'est la tendance du profil (ici vers le haut) qui décidera de la dominante: ici II, avec flèche vers le bas.

En cas d'égalité pratique entre deux niveaux, on peut soit décider que le niveau supérieur l'emporte (surtout s'il s'agit de deux niveaux contigus), soit entourer deux chiffres et réunir les deux cercles par un trait.

Remarques.

1° Ce système de synthèse des évaluations ne détient aucun pouvoir magique! Si l'on commet des erreurs signalées dans le présent chapitre, la note de synthèse sera probablement dépourvue de sens.

Un phénomène de concentration dans les bandes médianes étonne souvent les utilisateurs. Ils oublient que, dans le système traditionnel, aboutissant à un pourcentage, les notes se concentraient aussi dans la même zone: entre 65 et 75 ou 80%, tout le reste de l'échelle (de 100 échelons!) étant bien souvent presque vide.
En réalité, le phénomène de concentration est plus frappant parce que l'échelle à cinq degrés est beaucoup mieux visualisée que l'ancienne.

2° Rien n'empêche de pondérer les notes, c'est-à-dire d'accorder, par exemple, deux fois plus d'importance au travail en classe qu'aux travaux à domicile, à tel aspect d'une branche plutôt qu'à tel autre, etc.

3° *On trouvera, dans la partie consacrée à la «modération», des indications supplémentaires pour exploiter les notes, en particulier pour déterminer les résultats et les classements en fin d'année (voir page 218).*

G. Un cas particulier:

La notation de la composition française.

Ce problème mérite quelques considérations particulières, parce qu'il compte parmi les plus difficiles.

S'il était résolu, la docimologie remporterait peut-être sa plus belle victoire. Hélas! ou, plutôt, heureusement pour l'homme, la plus noble de ses activités, l'évaluation du beau, du vrai et du bien, échappera sans doute toujours à la quantification objective et donc automatisable.

1) *Quatre méthodes d'évaluation*

Les principales méthodes se ramènent à quatre: la méthode de l'impression générale, l'échelle de spécimens, la méthode analytique et les comptages de fréquences.

a) **La méthode de l'impression générale.**

C'est la plus subjective. On lui reproche le manque de fidélité d'un même correcteur, les désaccords entre plusieurs notateurs, spécialistes ou non, et son inefficacité fréquente: maint professeur s'est usé à noter avec soin des centaines de travaux par an sans obtenir des résultats convaincants.

Il y aurait beaucoup à dire sur la méthodologie de la composition française, et surtout sur la nécessité d'une réforme profonde des habi-

tudes scolaires qui traitent chaque élève comme un écrivain en puissance, mais il ne nous appartient pas d'en traiter ici. Retenons simplement que la conjugaison de deux faiblesses souvent accusées, l'une docimologique, l'autre méthodologique, grève l'exercice de rédaction d'une hypothèque tellement lourde que certains déclarent cet exercice inutile. Nous ne sommes pas de cet avis.

Malgré ses faiblesses, la méthode de l'impression générale est la mieux en harmonie avec la complexité essentielle de la composition. En utilisant une échelle courte, à trois ou, au maximum, à cinq degrés, un accord assez élevé peut être obtenu entre des notateurs expérimentés, de même formation et appartenant à un même milieu scolaire. Nous verrons plus loin les précautions à prendre avant de décider d'un échec.

Une composition corrigée par cette méthode ne doit pas, croyons-nous, ambitionner d'améliorer beaucoup le style ou la simple correction du langage écrit : d'autres exercices existent à cet effet.

La composition traditionnelle doit être une prestation d'exception, destinée ou bien à repérer des talents ou à en contrôler la maturation, ou bien à identifier des faiblesses, des carences, des maladresses que l'on combattra dans des exercices systématiques d'enrichissement et de remédiation.

Si l'on décide de s'en tenir à la seule appréciation globale, *la technique de l'évaluation en temps imposé* mérite l'attention.

Elle comporte des variantes mineures. La méthode proposée par S. Wiseman[1] est caractéristique. Quatre notateurs évaluent indépendamment des travaux faits en trente minutes, à une allure d'environ cinquante copies à l'heure (taux imposé pour les obliger à se décider très rapidement). On calcule la moyenne arithmétique des notes.

Cette méthode est probablement la plus efficace pour les travaux narratifs ou descriptifs. Pour les « dissertations », mot recouvrant toutes les compositions où le sujet doit argumenter, la supériorité de la méthode est plus discutée. Elle semble pourtant valide aussi en ce domaine si les notateurs sont de haut niveau intellectuel.

La validité générale de la méthode et son applicabilité à la « dissertation » est confirmée par deux recherches expérimentales de J. Britton.

1 S. WISEMAN, The Marking of English Composition in Grammar School Selection, in *British Journal of Educational Psychology*, XIX, nov. 1949, 208.

1re expérience[1].

Un ensemble de compositions écrites par 168 élèves de 15 ans a d'abord été noté, selon la méthode analytique, par un examinateur expérimenté. Les mêmes travaux ont été ensuite évalués indépendamment, selon la méthode de l'impression générale en temps imposé, par huit notateurs.

Pour disposer d'un critère, cinq autres travaux de chaque élève ont été notés par deux examinateurs.

Les notes analytiques et la note moyenne de trois impressions rapides tirées au hasard parmi les huit dont on disposait pour chaque étudiant, ont alors été comparées aux notes de critère :

corrélation entre notes analytiques et critère : 0,71
corrélation entre notes rapides et critère : 0,76

En outre, la fidélité de la notation rapide a été calculée en comparant les notes de deux groupes de trois examinateurs, choisis au hasard parmi les huit. On obtient un coefficient très élevé : 0,87

2de expérience[2].

Cette seconde expérience, beaucoup plus fine que la précédente, a porté sur un échantillon de 500 élèves de 17 ans, stratifié selon le sexe, le type d'école secondaire et le lieu d'habitation (Londres, centre et faubourgs, grandes villes, petites villes). Tous les types de sujets de compositions, de la simple narration à la dissertation philosophique et à l'amplification poétique, étaient aussi représentés.

De nouveau, la notation multiple selon l'impression rapide s'est montrée plus fidèle et plus valide que la notation analytique.

b) **Les échelles de spécimens.**

Un petit nombre de compositions traitant de sujets imposés, souvent cinq, sont choisies de façon à constituer des modèles représentatifs des différents niveaux de qualité : de la médiocrité à l'excellence.

Le notateur évalue les travaux par comparaison aux cinq textes étalons de l'échelle.

1 J. BRITTON, Experimental Marking of English Composition Written by Fifteen-Year-Olds, in *Educational Review* (Birmingham), vol. 16, 1, 1963, 17-23.
2 J. BRITTON, N. MARTIN et H. ROSEN, *Multiple Marking of English Composition*, Londres, H.M.S.O., 1966.

Après avoir connu une vogue considérable dans certains pays[1], les échelles de spécimens ont été abandonnées pour plusieurs raisons[2]:

1. il est rare que la composition à noter ressemble étroitement aux textes étalons;
2. les notateurs ont tendance à n'identifier que les caractéristiques communes qui les intéressent spécialement et négligent les autres;
3. des échelles différentes sont nécessaires selon le type de sujet et selon le niveau des élèves.

D. Pidgeon nous a signalé qu'une vaste expérience, menée récemment en Grande-Bretagne, avec une échelle de spécimens à 5 degrés, s'est soldée par un écart moyen de 2 degrés entre correcteurs. Le résultat n'est donc guère meilleur qu'en conditions purement subjectives.

Notons cependant que, depuis 1903 au moins, la méthode des échelles de spécimens retrouve périodiquement la faveur des chercheurs.

c) **La méthode analytique.**

Deux ou trois notateurs accordent un certain nombre de points pour chacune des différentes qualités qu'il est convenu d'observer. On totalise et on calcule la moyenne entre notateurs.

Cette méthode est la plus lourde des quatre; elle est souvent critiquée parce qu'elle dissèque un tout qui semble précisément échapper à toute dissection systématique. S. Wiseman remarque d'ailleurs que, dans un ensemble noté analytiquement, le travail estimé unanimement comme le meilleur, par les experts, se classe rarement en tête.

Pourtant, la recherche montre que, si les qualités à observer et l'importance à leur accorder sont définies avec précision, la méthode analytique est la plus sûre, si l'on doit se fier à un seul correcteur. La méthode de l'impression générale donne, nous l'avons déjà vu, d'aussi bons résultats, mais elle exige la participation de plusieurs notateurs[3].

1 Pour une étude d'ensemble, voir: E. HINTON, Study of the Qualities of Style and Rhetoric Found in English Compositions, New York, 1940.

2 R. BRADDOCK et al., *Research in Written Composition*, Champain, Ill, N.C.T.E., 1963, p. 12.

3 On trouvera aussi une bonne étude comparative in B. CAST, The Efficiency of Different Methods of Marking English Composition, in *British Journal of Educational Psychology*, IX, Nov. 1939, 257-259 et X, Feb. 1940, 49-60, cité par R. BRADDOCK, o.c., p. 13.

Pour être efficace et praticable, la méthode analytique semble devoir répondre à deux exigences au moins:

1° les correcteurs doivent accepter le plan d'analyse. L'idéal est qu'ils participent à son élaboration;

2° le plan ne doit pas être trop détaillé. Ainsi, la méthode très fine utilisée par E. Burton[1] n'offre d'intérêt que pour la recherche, en raison du temps très long qu'elle exige.

Quelles qualités observer?

C. Remondino[2] a montré que les attitudes fondamentales des correcteurs cultivés (professeurs ou non) ne diffèrent pas essentiellement: les mêmes qualités retiennent leur attention. Mais ils leur accordent des poids très variables, dans l'appréciation d'ensemble, et là gît réellement la source des divergences.

Elaborer un tableau très détaillé d'analyse du contenu est un leurre, car, même si les nuances distinguées existent, les notateurs les refondent inconsciemment. L'analyse factorielle met ce processus en évidence. Le travail de Remondino apporte d'intéressants résultats à ce propos.

Remondino a d'abord interrogé longuement vingt professeurs, de branches littéraires (enseignement secondaire) et il a aussi dressé la liste des qualités relevées dans des compositions scolaires. Il aboutit à l'ensemble suivant:

1 E. BUXTON, An Experiment to Test the Effects of Writing Frequency ..., in *Alberta Journal of Educational Research*, V, Juin 1959, 91-99.

2 Etude factorielle sur la notation des compositions scolaires portant sur la langue maternelle, in *Le Travail Humain*, XXII, Janv.-Juin 1959, 27-40.

1. Lisibilité	Qualité d'une écriture qui se prête à une lecture facile, rapide, sans équivoque.
2. Esthétique	Ligne harmonieuse et agréable des lettres et bon goût dans la mise en pages.
3. Présentation	Propreté, soin, bonne présentation de la copie.
4. Exactitude de l'orthographe .	Densité des erreurs d'orthographe.
5. Exactitude morphologique ..	Densité des erreurs morphologiques.
6. Exactitude syntaxique	Densité des erreurs de syntaxe.
7. Structure de l'exposé	Qualité d'un exposé fait avec ordre, dans les proportions voulues, et selon un plan.
8. Richesse d'idées	Quantité d'idées; ressources utilisées.
9. Pertinence des idées	Qualité par laquelle les idées exposées sont en juste rapport avec le thème traité.
10. Précision d'information	Véracité et exactitude objective des affirmations et des faits exposés.
11. Exhaustivité	Qualité qui consiste à ne rien laisser de côté de tout ce qui devait se dire.
12. Concision	Qualité par laquelle les choses à dire le sont avec le minimum indispensable de termes sans répétition, redondance ou tortuosité.
13. Propriété du langage	Juste emploi des termes.
14. Style	Facilité, exactitude, maîtrise de la langue au point de vue de la «construction des phrases».
15. Originalité	Qualité par laquelle, à travers le travail, transparaît et s'affirme quelque chose de la personnalité.
16. Maturité	Capacité de jugement, profondeur critique, acuité des raisonnements exposés.
17. Imagination	Capacité de création, de transfiguration, de «projection» révélée par le travail.

Une analyse factorielle a ensuite montré que ces 17 qualités relevaient de quatre groupes seulement:

1. présentation graphique (1, 2 et 3);
2. usage de la langue (4, 5, 6, 13, 14);
3. contenu et organisation de l'exposé (7, 8, 9, 10, 11, 12);
4. aspects personnels du fond (15, 16 et 17).

Exemples.

Les conclusions de Remondino recoupent largement celles des autres chercheurs. Les échelles suivantes, construites par l'*Educational Testing Service* (E.T.S.) de Princeton[1], et légèrement remaniées par E. Page[2], témoignent de la similitude des vues.

Nous traduisons le texte complet en raison de son intérêt méthodologique.

CRITERES POUR NOTER LES COMPOSITIONS.

I. **Définition des traits à évaluer.**

A. Idées ou contenu: la quantité et la qualité du matériel utilisé pour traiter du sujet.
B. Organisation: la relation entre les parties de la composition et l'ensemble.
C. Style: utilisation du langage au-delà de la simple correction grammaticale.
D. Mécanique: orthographe, grammaire, ponctuation.
E. Créativité.

II. **Guide pour l'évaluation de ces cinq traits.**

A. *Idées ou contenu.*

Niveau élevé.

L'étudiant traite de tous les points appelés par le sujet ou le plan de travail. Il comprend bien le sujet et utilise des définitions claires. Il sait considérer le sujet dans une perspective plus large que celle des autres élèves de la classe. Autrement dit, il témoigne d'une expérience plus riche.

Niveau moyen.

Les idées sont appropriées, mais conventionnelles et peu nombreuses. Certains aspects du sujet sont négligés. L'élève ne semble pas avoir un esprit richement meublé.

Niveau bas.

L'étudiant omet beaucoup d'aspects importants du sujet. Il semble ne pas disposer d'une réserve de connaissances relatives au sujet et, par conséquent, répète sans cesse quelques idées simples.

B. *Organisation.*

Niveau élevé.

L'étudiant suit un plan défini. S'il présente le pour et le contre, il avance des raisons pertinentes, dans un ordre efficace. S'il décrit quelque chose, il le fait de façon ordonnée (du sommet à la base, par ordre d'importance, par ordre de complexité, etc.). Si l'étudiant explique un concept ou un processus, il utilise un plan cohérent d'analyse, de définition

1 E.T.S., *Definitions of Ratings on the E.T.S. Compositions Scale*, cité par E. PAGE, o.c., pp. 70-77.
2 O.c., pp. 78-80.

ou d'illustration. L'étudiant sent bien ce qui se rapporte à son plan et évite des répétitions. Il témoigne du sens de la mesure en traitant les différentes parties de son travail.

Niveau moyen.

L'étudiant ne s'en tient pas à son plan ou introduit des idées sans rapport avec le sujet. Il consacre trop de temps à des choses peu importantes ou se répète. Il traite le sujet par association libre (Qu'est-ce qui me vient à l'esprit quand je pense à Hawaii ?) plutôt qu'en poursuivant un but bien défini.

Niveau bas.

L'étudiant ne semble pas s'être demandé ce qu'il allait faire avant de commencer à écrire. Il ne suit pas de plan. Le travail prend une direction, puis en change, en change encore et encore, jusqu'à ce que le lecteur soit perdu. Les points principaux ne sont pas clairement séparés les uns des autres, et leur ordre de présentation est laissé au hasard.

C. *Style.*

(Plusieurs aspects du style peuvent intervenir dans l'évaluation : individualité, vivacité, élégance, etc. Toutefois, nous nous intéressons ici à trois aspects stylistiques seulement : clarté, variation et éventail des ressources linguistiques.)

Niveau élevé.

L'étudiant utilise un langage qui rend aisé la compréhension du travail. Il utilise des mots adéquats, dans leur sens habituel. Les mots sont présentés dans un ordre normal. Les transitions sont bien ménagées. L'élève évite les ambiguïtés et ne trompe pas l'attente du lecteur. En même temps, l'étudiant évite la répétition monotone de mots, de compléments ou de structures de phrases. Finalement, il témoigne de la connaissance d'un large éventail de ressources linguistiques. Son vocabulaire est bon. Il utilise des structures parallèles ou fait un usage subtil de la subordination.

Niveau moyen.

L'étudiant égare parfois le lecteur en utilisant un mot inapproprié ou une tournure bizarre ; ou bien en utilisant une métaphore peu claire, ou en déplaçant de façon inopportune un complément ou une subordonnée, ou encore en pratiquant des transitions abruptes. La répétition de mots, de tournures et de structures de phrases devient monotone. Les ressources linguistiques sont limitées. L'élève utilise volontiers des clichés et des tournures éculés.

Niveau bas.

Les mots sont utilisés de facon vague. Tournures ambiguës, constructions boiteuses, vocabulaire et structures de phrases enfantins.

D. *Mécanique.*

Niveau élevé.

La structure des phrases est habituellement correcte, même lorsqu'il s'agit de modèles variés et compliqués. Les règles de l'orthographe sont respectées, même les mots difficiles sont généralement écrits sans faute. Pas de violation grave des règles de ponctuation, de majuscules, d'abréviations, d'écriture des nombres.

Niveau moyen.

Défauts de syntaxe occasionnels. Les mots difficiles sont parfois mal orthographiés. Quelques violations des règles de ponctuation, etc.

Niveau bas.

Très grand nombre de fautes.

E. *Créativité.*

Niveau élevé.

L'étudiant surprend par des façons neuves et efficaces de considérer le problème. Il introduit des idées nouvelles dans son traitement du sujet. Il trouve des façons fraîches et intéressantes d'utiliser le langage pour faire ressortir ses idées.

Niveau moyen.

L'étudiant pense à ce que l'on s'attendait qu'il penserait. Il traite les choses comme à peu près tout le monde. Il utilise des expressions et des structures de phrases ordinaires.

Niveau bas.

L'étudiant utilise des clichés de pensée et d'expression. Il traite le sujet de façon superficielle. Il répète des formules sans réellement en comprendre la signification.

Un groupe de professeurs, décidant d'évaluer une même composition en se référant aux cinq échelles que nous venons de parcourir, aboutirait certainement à des notes encore fort discordantes. Pourquoi ? Parce que les limites entre les différents degrés restent floues (nous ne considérons pas ici le problème de la synthèse des notes relatives aux cinq échelles).

Pour arriver à un degré de concordance élevé, plus de précision est encore nécessaire. En voici un exemple[1].

En guise de composition, des élèves de douze ans ont été invités à combler une lacune dans un récit d'aventures dont on fournissait le résumé des chapitres précédents, puis deux pages complètes entre lesquelles un « trou » d'une page était ménagé.

Un même travail a été évalué selon cinq dimensions :

- la pertinence sémantique de la réponse ;
- la pertinence syntaxique ;
- le vocabulaire ;
- les structures grammaticales ;
- l'organisation des idées.

Voici comment l'évaluation de la pertinence sémantique a été conduite. On remarquera que la démarche a été en quelque sorte programmée. Un mode d'emploi correct ne suffit pas encore : chacun doit

1 R. DE BAL, G. DE LANDSHEERE, J. PAQUAY-BECKERS, *Construire des échelles d'évaluation descriptive*, Bruxelles, Ministère de l'Education nationale, Organisation des Etudes, 1976.

le comprendre et l'appliquer de la même façon. Or, en cours d'expérience, on a constaté que des divergences d'évaluation subsistaient, non plus à cause de l'imprécision des consignes, mais de leur mauvaise compréhension.

PERTINENCE SEMANTIQUE.

A. DEFINITION DE LA DIMENSION.

Le travail de l'élève est pertinent au point de vue sémantique s'il développe des idées appropriées au contexte, c'est-à-dire le résumé et les paragraphes 1 et 3. Il ne doit pas nécessairement comporter toutes les informations du texte de l'auteur. L'important est qu'il ne les contredise pas.

Deux questions permettront d'évaluer la pertinence sémantique:
- L'élève n'a-t-il pas introduit d'élément contredisant les informations?
- Les intègre-t-il dans un contexte plus large?

Ces deux aspects de l'évaluation sont décrits ci-après.

B. DEFINITION DES ASPECTS QUI INTERVIENNENT DANS LA DIMENSION.

1er aspect: L'introduction d'éléments non pertinents.

Les éléments non pertinents sont des éléments qui entrent en contradiction avec les informations apportées par le texte.

La lecture d'un nombre important de copies d'élèves nous a permis d'établir la liste de ces éléments non pertinents. Elle vous servira à évaluer ce premier aspect de la pertinence sémantique.

1. Le début du paragraphe 2 ne situe pas les faits dans le laboratoire (exemple: pousser la porte, entrer... sont des actions non pertinentes; elles sont déjà exprimées dans le paragraphe 1).
2. Aucune allusion n'est faite au «spectacle insolite». Cette allusion peut se limiter à une exclamation, à un seul mot.
3. Dans le paragraphe 2, le commissaire est présent ou attendu.
4. L'aspect et le contenu du message sont modifiés (par exemple, le bout de papier devient une lettre, longue et soignée).
5. L'enfant évoque des blessures graves, des sinistres, des catastrophes...

2e aspect: L'intégration des informations.

La seule estimation du nombre d'éléments non pertinents ne suffit pas à rendre compte de la valeur du travail de l'élève.

On nuancera cette estimation en appréciant la manière dont l'élève intègre les faits qu'il relate dans un contexte plus large.

a) On cherchera les indices de cette intégration dans
   - l'évocation de la personnalité de l'oncle;
   - les allusions à des personnages secondaires;
   - l'expression des sentiments de Luc...

   (Cette liste n'est pas exhaustive.)

b) On considérera que l'élève n'a pas intégré les informations s'il se contente de relater strictement des faits qui se situent après l'entrée de Luc dans le laboratoire (fin du premier paragraphe) et sa réponse au commissaire (début du troisième paragraphe).

c) Certains travaux développent des idées qui, sans être nécessairement en contradiction avec les paragraphes 1 et 3, n'ont pas de rapport direct avec la situation.

C. DESCRIPTION DE L'ECHELLE: *Pertinence sémantique.*

	Introduction d'éléments non pertinents	*Intégration des informations*
TB	Pas d'éléments non pertinents.	Intégration des faits du § 2 dans un contexte plus large
B	Pas d'éléments non pertinents.	Pas d'intégration.
	1 ou 2 éléments non pertinents.	Intégration des faits du § 2 dans un contexte plus large.
S	1 ou 2 éléments non pertinents.	Pas d'intégration.
F	3 ou 4 éléments non pertinents.	Quelles que soient les caractéristiques d'intégration.
I	Plus de 4 éléments non pertinents.	Quelles que soient les caractéristiques d'intégration.
	Quel que soit le nombre d'éléments non pertinents.	Idées sans rapport direct avec la situation.

D. DEMARCHE POUR L'EVALUATION DE LA DIMENSION.

1. Portez votre appréciation pour le premier aspect:
   - pas d'éléments non pertinents? ☐
   - 1 ou 2 éléments non pertinents? ☐
   - 3 ou 4 éléments non pertinents? ☐

2. Portez votre appréciation pour le second aspect:
   L'élève
   - intègre les informations dans un contexte plus large? ☐
   - n'intègre pas les informations? ☐
   - développe des idées sans rapport avec la situation? ☐

3. En vous reportant au schéma suivant, choisissez le niveau qui correspond aux deux caractéristiques que vous venez de repérer dans le travail de l'élève.

Intégration d'éléments non pertinents	*Intégration*	
Pas d'éléments non pertinents	→ intégration →	TB
	→ pas d'intégration →	B
	→ idées sans rapport direct avec la situation →	I
1 ou 2 éléments non pertinents	→ intégration →	B
	→ pas d'intégration →	S
	→ idées sans rapport direct avec la situation →	I
3 ou 4 éléments non pertinents	→ idées sans rapport direct avec la situation →	F
Plus de 4 éléments non pertinents	→	I

Note pour la PERTINENCE SEMANTIQUE: ☐

d) **La méthode des comptages de fréquences.**

Nous la signalons ici, bien qu'elle n'ait pas de place dans la catégorie des évaluations subjectives, afin de compléter le tableau des méthodes.

Elle consiste à définir des types de fautes (après une étude diagnostique), à compter celles-ci dans la composition et à exprimer le nombre obtenu pour 100 ou pour 1.000 mots.

Au lieu d'adopter un point de vue négatif, on peut aussi faire l'inventaire (relativement exhaustif) des divers processus utilisés dans la composition. Le plus bel exemple de pareille méthode est fourni par les *Eléments pour servir à l'étude d'analyses littéraires* de A. Purves[1].

Une dernière méthode positive consiste à isoler certains aspects qui se révèlent prédictifs, à un seuil donné, de la note normalement attribuée par des correcteurs qualifiés. Un ordinateur, programmé en conséquence, peut alors se charger de la notation.

Les comptages de fréquences offrent probablement le seul espoir de notation objective des compositions. Ils semblent toutefois devoir rester du domaine de la recherche étrangère à la pratique scolaire jusqu'à ce que l'ordinateur et les logiciels adéquats ne deviennent d'un accès tellement facile que l'étude des fréquences se fasse de facon économique et constitue un des aspects de l'évaluation. Cette facilité sera normalement offerte dans un proche avenir.

2) *Plusieurs sujets au choix ?*

J. Britton remarque avec raison que la rédaction d'examen sur un thème unique implique souvent, chez le professeur, l'hypothèse qu'un sujet comme «Nuages» révélera aussi bien l'aptitude d'un étudiant que «Exposez vos vues sur la peine capitale».

On a montré expérimentalement qu'il n'en est rien, nous allons y revenir. Toutefois, il est douteux que la justice soit beaucoup mieux respectée si l'on offre trois sujets au choix quand les points d'examens sont attribués pour une rédaction unique. Il faudrait d'abord prouver qu'avec trois, voire cinq sujets, on recouvre tout le spectre des stimuli nécessaires pour donner à chacun une chance égale; et même si cela était établi, il importerait encore que le professeur sélectionne toujours les sujets à bon escient.

1 A. PURVES et V. RIPPERE, Eléments pour servir à l'étude d'analyses littéraires (traduit de l'anglais par J. Dubois), in *Scientia Paedagogica Experimentalis*, VI, 2, 1969, 229-334.

L'expérience, disions-nous, nous montre que le rendement de l'élève varie selon le thème proposé. Finlayson[1] a établi, pour la fin de l'école primaire, que si, à une semaine d'intervalle, les élèves sont invités à faire une rédaction à partir d'un même choix de quatre sujets, ceux qui changent de sujets (plus de la moitié dans l'expérience) obtiennent des notes significativement différentes.

Wiseman et Wrigley[2] arrivent à la même conclusion (choix parmi le même ensemble de cinq sujets à quatre mois d'intervalle), tout en prouvant que la différence de note n'est pas imputable au manque de fidélité du professeur. Vernon et Mallican[3] confirment la conclusion au niveau universitaire.

Que faire ? Si l'on veut vérifier la capacité de s'exprimer par écrit dans une perspective étroitement définie (par exemple: capacité de faire un rapport scientifique ou d'écrire une lettre commerciale), la sensibilité particulière des candidats n'a guère à être considérée et un sujet suffit.

Par contre, s'il s'agit ou bien de déceler une aptitude à l'expression, où qu'elle soit, ou d'amener les élèves à discuter de problèmes ou la tournure d'esprit et le niveau d'information peuvent jouer un grand rôle, il n'est guère possible de porter un jugement à partir d'une seule composition. L'idéal est alors de se référer aussi au travail de l'année; si, chaque fois, l'élève a pu choisir parmi plusieurs sujets (cinq, par exemple), il aura réagi à plusieurs dizaines de thèmes différents en quelques mois et l'on peut supposer que chacun aura vraiment eu sa chance. Laisser inventer le sujet à traiter serait tentant, si l'on pouvait se prémunir contre la préparation frauduleuse. C'est bien difficile...

3) *Conclusion.*

Le sort à réserver à la composition française dépendra du but poursuivi. Que veut-on ? Corriger un défaut particulier ? Dans ce cas, il vaut mieux n'examiner que celui-là dans les travaux ou l'attaquer dans des exercices spéciaux.

Connaître la capacité des élèves à se corriger, à identifier la prose conforme au bon usage ? Dans ce cas, un test objectif, avec réponses à choix multiple, peut faire l'affaire.

1 D.S. FINLAYSON, The Reliability of the Marking of Essays, in *British Journal of Educational Psychology*, XXI, 2, 1951, cité par Britton (1966).

2 S. WISEMAN et J. WRIGLEY, Essay-Reliability : the Effect of Choice of Essay-Title, in *Educational and Psychological Measurement*, 18, 1, 1958.

3 P. VERNON et G. MILLICAN, A Further Study of the Reliability of English Essays, in *British Journal of Statistical Psychology*, VII, 2, 1954.

Faire acquérir la capacité de s'exprimer clairement ? Dans ce cas, des thèmes très limités permettent de concentrer l'attention sur l'objet du travail. Pourquoi bannir les sujets tels que : « Décrivez un vélo de course » ?

Une indication cruciale est trop souvent oubliée : qui sera le lecteur ? Dans la réalité, on varie l'écriture, de façon parfois considérable, selon le destinataire du message. Pourquoi ne pas apporter chaque fois cette précision lorsque l'on propose un thème de composition ?

Déceler des talents d'écrivains ? Alors, il faut laisser libre cours à l'expression de l'élève, l'invitant souvent à choisir lui-même son sujet.

Mais que l'on ne s'y trompe pas, on testera ainsi des capacités différentes. Dans la seconde expérience de J. Britton, décrite plus haut, l'auteur a montré :

1° que la corrélation entre les résultats à la rédaction créative et un résumé de texte varie de 0,30 à 0,40 ;

2° que la corrélation de la rédaction avec un test de compréhension de texte est un peu plus élevée : de 0,35 à 0,45.

Les écarts sont donc considérables.

Enfin, J. Foley[1] remarque que chacun semble convaincu intuitivement que la meilleure façon d'évaluer la capacité d'écrire est de faire écrire. Or, on dispose de preuves de plus en plus nombreuses « qu'un test d'aptitude verbale permettrait de prédire les notes de composition mieux qu'un test exigeant que l'étudiant écrive réellement ».

Nous l'avons déjà dit, nous ne pensons pas cependant qu'il faille bannir la composition traditionnelle. Dans certains pays où elle avait été supprimée à cause de la difficulté d'évaluation rigoureuse, on l'a d'ailleurs réintroduite, notamment aux Etats-Unis où la Commission des Examens d'Admission dans l'Enseignement Supérieur a renoncé au « Test objectif de composition ».

Toutefois, si la rédaction continue à jouer un rôle important dans les examens, la façon encore trop répandue de noter doit être modifiée. Nous faisons nôtre la proposition de R. Braddock :

« Si l'on doit évaluer un grand nombre d'étudiants pour décider de leur réussite ou de leur échec, il importe de permettre à ceux qui échouent de faire un second travail (...). Si l'on entend attribuer des grades, trois travaux au moins sont nécessaires ; on retiendra les deux

1 J. FOLEY, in B. BLOOM, J. HASTINGS et G. MADAUS, *Handbook..., o.c.*, pp. 800-801.

meilleurs dont on fera la moyenne (...). Les travaux doivent être notés par au moins deux correcteurs, utilisant un système de notation, bien compris et accepté, auquel on les a bien entraînés. »[1]

III. La notation objective

La notation objective ne fait pas intervenir l'avis personnel des correcteurs. L'exemple le plus simple est fourni par une question comportant dix multiplications à raison d'un point par résultat exact. De même, la notation d'un test composé d'*items* à choix multiple s'opère par simple comptage du nombre de choix corrects.

De façon plus générale, on appelle notation objective « l'assignation de valeurs numériques à des échantillons comportementaux suffisamment limités, définis et contrôlés pour permettre un accord général parmi les juges ou notateurs »[2].

Comme nous l'avons remarqué déjà, l'objectivité de la notation ne garantit en rien l'objectivité de l'examen dans son ensemble. Le choix des questions reste, en dernière analyse, toujours subjectif et l'on imagine aisément un examen dont aucune question n'échantillonnerait réellement la capacité que l'on prétend évaluer. Tout objective que puisse être la correction, l'épreuve n'aurait, dans pareil cas, aucune validité. Les exemples de semblable mésaventure fourmillent.

IV. L'étalonnage

L'interprétation de toute note exige un point de référence, un critère, une norme.

Etalonner consiste à définir des normes.

Traditionnellement, on prend pour norme la distribution des performances individuelles dans une tâche ou dans un ensemble de tâches. A partir du score obtenu, on constate qu'un étudiant se situe, par exemple, à la 70e place dans une population de référence dont la distribution des résultats est ramenée à cent échelons. Par conséquent, *cette 70e place n'indique pas ce que l'élève considéré connaît de la matière du test (il en connaissait peut-être huit dixièmes), mais bien le rang qu'il occupe dans une course comptant cent participants appartenant à sa catégorie* (par exemple, les élèves de la région de Paris fréquentant le CM 2). Le critère d'appréciation, la norme est, dans

1 R. BRADDOCK et al., o.c., p. 45.

2 J.C. FLANAGAN, Units, Scores and Norms, in E. LINDQUIST, Ed., *Educational Measurement*, Washington, A.C.E., 1961, 4e éd.

ce contexte, soit le rang dans un classement, soit, dans des procédures plus évoluées, la situation par rapport à la moyenne du groupe de référence. Les tests ainsi étalonnés sont appelés de plus en plus souvent *tests normatifs (norm referenced tests)*, appellation qui n'est certes pas la plus heureuse, mais qui est déjà passée dans l'usage.

Depuis quelques années, une nouvelle tendance est apparue et s'affirme de plus en plus dans la pratique scolaire.

La première préoccupation de l'éducation de base n'est pas compétitive ou sélective, mais formative. Pour aider un élève, il importe bien moins de lui indiquer quelle place il occupe, par rapport aux autres, dans la course au savoir, que de lui apprendre jusqu'où ses efforts l'ont conduit sur le chemin qui mène à la maîtrise désirée d'un apprentissage. A quelle distance se trouve-t-il encore de l'objectif à atteindre et quels obstacles doit-il encore surmonter?

Dans cette perspective, le critère, la norme, c'est la tâche assignée et l'étalonnage se fera selon les composantes maîtrisées. On appelle les tests destinés à fournir ce type d'information les tests centrés sur les objectifs (*criterion referenced tests*)[1].

IV-1. **Etalonnage des tests normatifs**

Nous venons de le voir, la norme est ici essentiellement fournie par les performances d'un groupe d'individus pris pour référence.

Selon le but poursuivi, les dimensions de ce groupe varient: tous les élèves ou un échantillon des élèves de même niveau scolaire ou de même âge, dans une ville, un canton, une région, un pays... En toute rigueur, la validité des normes se limite aux populations qui les ont produites. Il importe donc de savoir lesquelles.

Nous avons déjà vu qu'une des façons de procéder consiste à informer l'élève de la place qu'il occuperait si son groupe comptait cent individus (centilage). Pour des raisons sur lesquelles nous allons revenir, on préfère souvent situer une performance par rapport à la moyenne ou au médian[2] d'un groupe de référence.

A partir de ces repères, on construit des échelles qui permettent d'exprimer des scores bruts en unités comparables.

1 «J'appelle mesures à référence critérielle celles qui se rapportent à une valeur absolue de qualité, alors que j'appelle mesures à référence normative celles qui se rapportent à une valeur standard relative.»
GLASER, cité par R. HORN, *Lernziele und Schülerleistung*, Weinheim, Beltz, 1972, p. 14.

2 Rappelons que le médian est le point qui divise une série de notes ordonnées en deux parties égales, 1-2- **3** 4-5 ou 1-2- **3.** 4-5-6.

Les systèmes d'étalonnage normatifs les plus utilisés sont les suivants:

1. le centilage;
2. les notes types ou notes z;
3. l'échelle normalisée à 5 classes;
4. l'échelle normalisée à 9 classes (Stanines).

Nous les étudions successivement.

A. **Le centilage**

1) *Définition.*

Le centile, ou rang occupé sur une échelle à cent degrés, jouit encore d'une certaine faveur à cause de sa ressemblance extérieure avec le pourcentage, dont il diffère pourtant de façon fondamentale. Le pourcentage indique quelle proportion du total des points attribués à un examen l'élève a obtenu. Le centile indique combien d'élèves se classeraient après un élève donné si la classe comptait cent élèves. Donc un score équivalant au 90e centile est supérieur à 90% de la population considérée.

Le médian correspond au 50e centile.

2) *Calcul.*

En principe, on ne calcule pas les centiles à partir de moins de cent notes. L'exemple suivant utilise les cent notes qui nous ont déjà servi lors de l'étude de la courbe de Gauss.

On peut déterminer les centiles à l'aide d'une formule qui entraîne des calculs assez longs. Le plus souvent, on se contente d'approximations, aisément lues sur un abaque construit de la façon suivante sur du papier millimétré:

1° porter horizontalement la valeur des classes;

2° pointer les effectifs cumulés aux limites supérieures exactes des classes;

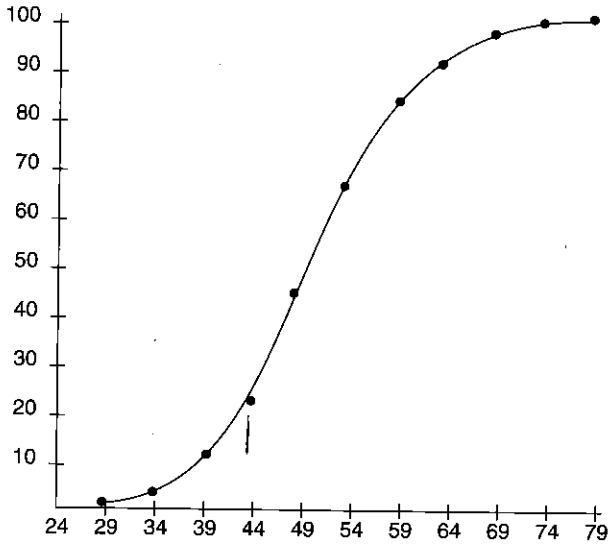
3° rejoindre les points en une courbe qui, si les scores se distribuent assez normalement, prend la forme d'un S (ogive de Galton). Si l'ogive s'ébauche bien, on peut se permettre de régulariser, de «polir» son tracé;

4° il suffit alors de partir de l'échelle verticale pour venir lire, à partir du point rencontré sur l'ogive, la valeur approximative du centile cherché.

Exemple

Classes	Effectifs (f)	Effectifs cumulés (f_c)
74 - 78	1	100
69 - 73	1	99
64 - 68	5	98
59 - 63	9	93
54 - 58	18	84
49 - 53	22	66
44 - 48	23	44
39 - 43	10	21
34 - 38	8	11
29 - 33	2	3
24 - 28	1	1

Abaque.



Comparaison de quelques valeurs obtenues:

	par lecture sur l'abaque[1]	*par calcul*
16e centile	*40,5*	*41*
50e centile	*49,5*	*49,9*
84e centile	*58,5*	*58,5*

1 Dans la pratique, l'ogive est dessinée sur du papier millimétré. Plus le dessin est grand, plus la lecture est aisée et précise.

Remarques.

1. Le 50e centile correspond à la médiane.
2. On préfère parfois étalonner en déciles: le 10e centile = le 1er décile, etc.
3. Ecart interquartile = 75e C - 25e C.
4. Valeur approximative de l'écart type = 3/4 de l'écart interquartile.

3) *Critique.*

L'échelle en centiles n'est, nous l'avons déjà dit, qu'une vaste échelle d'évaluation à 100 degrés. Elle informe sur le rang occupé, mais non sur la distance entre rangs. Or, si la distribution est normale, cette distance rétrécit à mesure que l'on se rapproche du médian (égal à la moyenne dans le cas de normalité parfaite).

Le tableau suivant montre clairement ce phénomène.

	-2σ	-1σ	0	$+1\sigma$	$+2\sigma$
Centiles	2e	16e	50e	84e	98e

On constate que 68 centiles sont agglomérés autour de la moyenne, ce qui donne un classement trop grossier pour la majorité du groupe. On risque, par exemple, de considérer qu'une différence importante sépare le 75e du 25e centile. L'ancienne norme du pourcentage subsiste tenacement dans les esprits et incline à penser que le 75e C vaut trois fois le 25e. Le tableau ci-dessus nous montre qu'il s'agit, en réalité, de deux résultats relativement proches et moyens.

En fait, si les normes en centiles restent utilisées, notamment pour la sélection dans certaines universités américaines, c'est parce que celles-ci ne recrutent que dans la bande des dix ou vingt centiles supérieurs, zone où la discrimination est satisfaisante.

B. Les notes types ou notes z (écart réduit).

Un élève a obtenu les notes brutes suivantes[1]. Calcul: 22 sur 25; lecture: 72 sur 100; sciences: 26 sur 50.

L'examen de ces notes n'apprend pas grand-chose. Si le professeur est sévère en lecture et beaucoup moins en calcul, le 72/100 en lecture est peut-être plus méritoire que le 22/25 en calcul. De plus, des notes brutes ne permettent pas la comparaison avec d'autres élèves de même niveau pédagogique, par exemple.

1 Adapté d'après R. THOMAS, *o.c.*

Pour rendre la comparaison possible, on exprime ces notes en fonction des écarts types, ce qui permet de les situer sur une même courbe (notes types ou scores standard).

Supposons que l'on obtienne les résultats suivants:

	Moyenne	*Ecart type*
Calcul	15	2,5
Lecture	50	10
Sciences	29	5

Calcul des scores standard: $z = (x - M) / \sigma$

Exemple: transformation d'une note 22, obtenue en calcul.

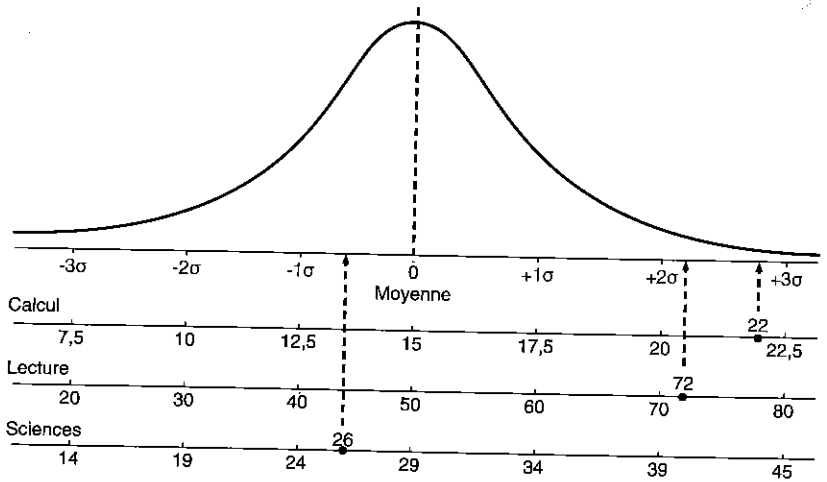
$$(22 - 15) : 2{,}5 = 2{,}8\ \sigma$$

Connaissant z, on peut retrouver la note brute:

$$x = M + z\,\sigma$$

$$x = 15 + (2{,}8 \times 2{,}5) = 22.$$

Représentation graphique[1]:



1 Il importe de souligner qu'on étalonne les notes de chaque branche sans en changer la distribution (qui peut donc ne pas être tout à fait normale). Par contre, dans les *échelles normalisées* (voir plus loin), on répartit le nombre total d'observations dans des classes, selon la distribution normale.

Application.

Les étudiants I et II ont passé une série de tests.

Comparer les résultats[1].

TEST	Moyenne M	Ecart type σ	X Scores bruts		*x Déviations (M-X)*		*Notes étalonnées z*	
			I	*II*	*I*	*II*	*I*	*II*
Anglais	155,7	26,4	195	162	+ 39,3	+ 6,3	+ 1,49	+ 0,24
Lecture	33,7	8,6	20	54	- 13,7	+ 20,3	- 1,67	+ 2,48
Information	54,5	9,2	39	72	- 15,5	+ 17,5	- 1,67	+ 1,88
Aptitudes scolaires	87,1	25,8	139	84	+ 51,9	- 3,1	+ 2,01	- 0,12
Attention	24,8	6,8	41	25	+ 16,2	+ 0,2	+ 2,38	+ 0,03
Totaux			434	397			+ 2,54	+ 4,51
Moyennes							+ 0,51	+ 0,90

Si l'on s'en tenait aux simples totaux (scores bruts I et II), l'étudiant I serait considéré comme supérieur à l'étudiant II. Le calcul des pourcentages confirmerait cette impression.

Or si, par le calcul des notes z, on ramène tous les résultats à la même moyenne et à la même unité de déviation par rapport à celle-ci, on constate:

1° qu'en moyenne, l'étudiant II (0,90) obtient un résultat presque deux fois aussi favorable que l'étudiant I (0,51);

2° que l'étudiant II obtient les résultats les plus homogènes.

Critique.

La transformation en notes z repose sur l'hypothèse que, pour le groupe considéré, les aptitudes sur lesquelles portent les tests se distribuent de la même façon et conduisent aux mêmes moyennes et dispersions. Cette hypothèse est pratiquement impossible à vérifier, mais, comme le remarque J.-P. Guilford, on peut être assuré que ce système est de toute façon meilleur que le calcul du pourcentage.

Par ailleurs, le centrage de l'échelle sur la moyenne donne de ce point une image discriminative fausse. En raison des imperfections inévitables de la notation, le pur hasard fera basculer un nombre plus ou moins élevé de sujets d'un côté ou de l'autre.

1 Chiffres empruntés à J.-P. GUILFORD, *Fundamental Statistics*, o.c., p. 513.

Une transformation.

Les notes z prennent zéro pour moyenne et un écart type égal à un comme unité.

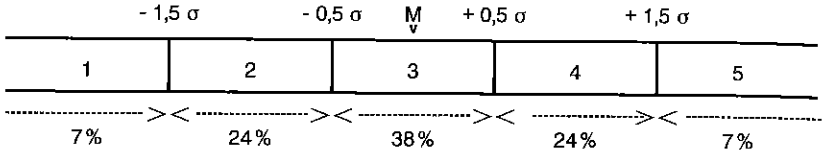
Certains considèrent ce système comme peu parlant et préfèrent exprimer les résultats à la façon des quotients intellectuels où 100 est traditionnellement pris comme moyenne et 145 considéré comme un quotient de «génie». La transformation est aisée. Cent est donc pris comme moyenne et chaque note z multipliée par 15[1].

C. **L'échelle normalisée à 5 classes** (N entre 50 et 100).

Dans les 2e, 3e et 4e classes, les notes sont réparties par écart type, selon la distribution normale.

La 3e classe est centrée sur la moyenne.

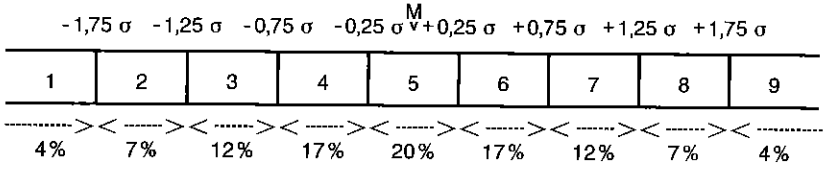
Les 1re et 5e classes sont d'étendue illimitée.



D. **L'échelle normalisée à 9 classes** (Stanines)

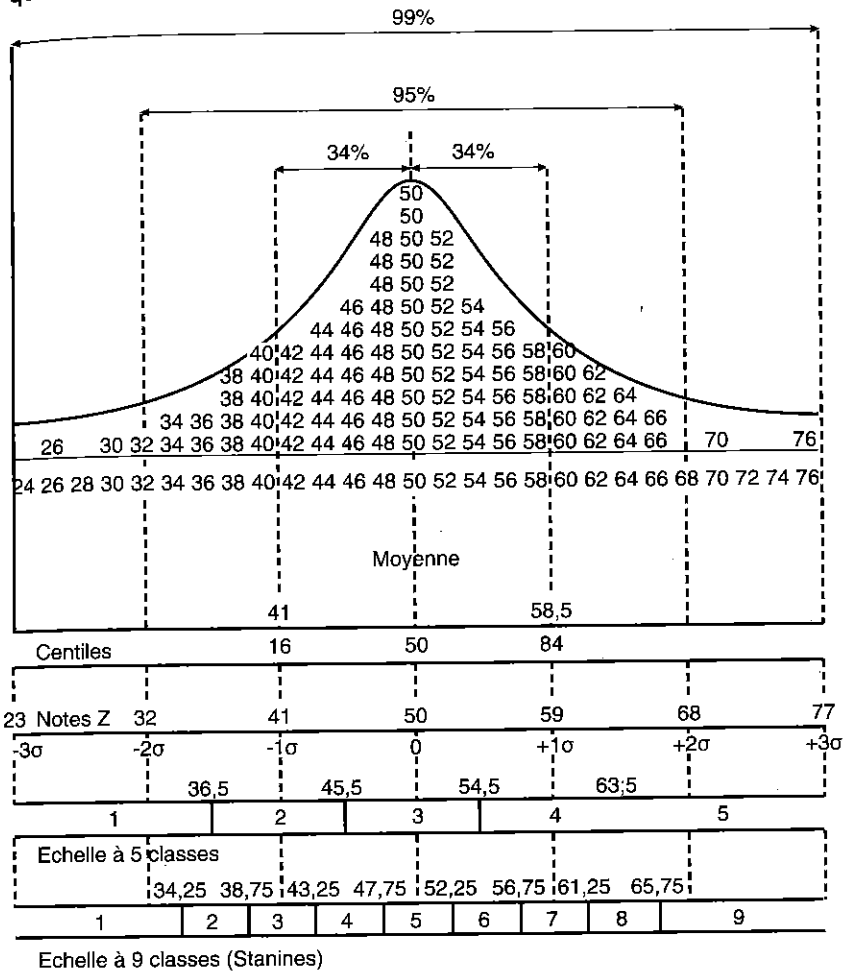
Les 1re et 9e classes sont d'étendue illimitée.

Les autres classes comportent un demi-écart type, la 5e étant centrée sur la moyenne.



1 Il existe bien d'autres échelles obtenues à partir de la note z. Voir G. DE LANDSHEERE, *Dictionnaire de l'évaluation et de la recherche en éducation*, Paris, P.U.F., 1992, 2e éd.

Le graphique ci-dessous permet une comparaison aisée des quatre types de normes.



Echelle à 9 classes (Stanines)

IV-2. **Etalonnage par rapport à l'objectif. Les tests critériels.**

En passant de l'étalonnage traditionnel à l'étalonnage centré sur les objectifs, on abandonne un critère externe (les performances du groupe de référence) pour un critère interne (le degré de maîtrise d'un apprentissage).

On l'a déjà souligné à plusieurs reprises, informer un élève que le résultat obtenu à un examen le classe premier, cinquième ou dernier de sa classe n'informe pas sur ce qu'il a effectivement appris. Cette

constatation n'est évidemment pas neuve. Il faut toutefois attendre 1962 pour que soit formulée par R. Glaser[1] une théorie rigoureuse relative à l'évaluation destinée à informer sur le point atteint dans un apprentissage.

Un apprentissage, écrit en substance Glaser, peut aller d'une ignorance, d'une incapacité totales à une maîtrise parfaite. Le niveau de performance d'un individu se situe à un point entre ces deux extrêmes. C'est ce point qu'élève et éducateur ont intérêt à connaître. Dans quelle mesure la performance observée ressemble-t-elle, à un moment donné, à la performance souhaitée ? Cette performance souhaitée constitue le *critère d'appréciation*. D'où l'appellation *mesure critérielle, test critériel*.

Même si elles sont passées dans l'usage, ces expressions ne sont pas heureuses, car prendre pour point de repère la performance d'autrui, comme on le fait dans les tests normatifs, est aussi un critère. Il faut donc tenir en mémoire que, par convention, l'appellation *mesure critérielle* a le sens que Glaser lui a donné[2].

La définition du test critériel, actuellement la plus utilisée, est celle de Popham[3] :

« Test permettant d'interpréter les performances d'un élève par rapport à un ensemble de compétences bien définies. »

Les résultats d'une telle évaluation peuvent être utilisés à diverses fins :

- Décrire, à un moment donné, la performance d'un élève dans un domaine particulier. Dans quelle mesure le « connaît »-il ? Une telle évaluation est généralement assez pointue, car l'aspect diagnostique est capital. Il s'agit donc d'une évaluation formative.
- Etablir si un élève a fait ou non la preuve d'un apprentissage relevant des principaux objectifs de l'éducation (par exemple, être capable d'utiliser les principales sources de référence).
- Etablir si, en fin de cycle, l'élève a effectivement acquis les principales compétences visées par le curriculum. La démarche adoptée par A. Inizan[4] pour décider si un élève sait lire offre un bel exemple de macro-évaluation. Possède cette habileté, écrit-il en substance, tout élève capable de déchiffrer le texte de contrôle qui figure dans mon livre.
- Etablir, à des fins de certification d'aptitudes professionnelles, si l'individu est capable (en termes de savoirs, d'habiletés et de savoir-être) d'assumer les responsabilités générales ou spéciales inhérentes à la profession qu'il veut exercer ou à la fonction qu'il ambitionne de remplir.
- Estimer dans quelle mesure un programme d'éducation ou de formation est adéquat, efficace.

Remarquons, au passage, que, dans la littérature relative à l'évaluation critérielle, les termes *objectifs, compétence et habileté* sont assez interchangeables.

1 R. GLASER, Instructional technology and the measurement of learning outcomes. Some questions. *American Psychologist*, 1963, 18, 519-521.

2 Preuve que cette appellation fait problème, R.K. Hambleton relève, en 1985, 57 définitions différentes du test critériel. Voir R.K. HAMBLETON, Criterion-referenced measurement. *In* T. HUSEN et T.N. POSTLETHWAITE, *International Encyclopedia of Education*, Oxford, Pergamon, 1985, 1108-1113.

3 W.J. POPHAM, *Criterion-referenced measurement*, Englewood Cliffs, Prentice-Hall, 1978.

4 A. INIZAN, *Le temps d'apprendre à lire*, Paris, Bourrelier, 1964.

Définir la compétence

Sans une définition précise de la connaissance ou de l'habileté à propos de laquelle on s'interroge, le test ne peut être valide. La distinction entre objectifs de maîtrise, de transfert et d'expression aide à baliser le terrain.

1. *Les objectifs de maîtrise*

Ils portent sur un univers entièrement circonscrit et qui, par là même, peut être totalement connu et *a fortiori* prévu (exemples : tables de multiplication des dix premiers nombres entiers, faits, dates, lieux, règles de grammaire, ...).

2. *Les objectifs de transfert*

Ici, on ne peut prédire toutes les situations. Des comportements appris dans des conditions données devront s'appliquer dans d'autres. Parfois, les conditions sont suffisamment proches pour que le transfert soit direct. Dans d'autres cas, la situation sera beaucoup plus éloignée du connu et, pour résoudre le problème, des éléments pertinents devront être extraits de plusieurs expériences antérieures (analyse), puis recombinés (synthèse) et transférés à la situation nouvelle. Exemples : être capable d'atterrir sur tous les terrains ouverts aux vols internationaux (même sur ceux que l'on n'a jamais vu antérieurement) ; pouvoir déterminer à quelle école littéraire un texte en prose appartient.

3. *Les objectifs d'expression*

L'appellation est d'Eisner[1] qui reconnaît à l'école son rôle d'initiatrice à la culture, mais rappelle qu'elle doit aussi aider à modifier et à développer les outils culturels existants. Guilford parlerait, dans ce contexte, de convergence et de divergence.

Les objectifs d'expression, tels qu'Eisner les conçoit, répondent aux caractéristiques suivantes:

- Ils ne décrivent pas le comportement final à acquérir, mais une situation éducative dans laquelle les élèves doivent résoudre un problème, accomplir une tâche en explorant des possibilités, en imaginant des solutions par toutes les voies qu'ils peuvent imaginer. De multiples réponses sont donc possibles.
- L'évaluation ne se fait pas selon un critère unique, mais par une réflexion sur ce qui a été produit, afin d'en apercevoir l'originalité et la signification.

Exemples: Interpréter un rôle dans une pièce de théâtre; créer une forme à trois dimensions à l'aide de fil de fer et de bois.

Il s'agit donc de faire montre de créativité.

4. *Relations entre ces trois types d'objectifs*

La parenté entre les objectifs de transfert et les objectifs d'expression est indéniable, mais, à mesure que le degré de divergence augmente, la relation entre les situations antérieurement vécues et les comportements nouveaux devient de plus en plus ténue.

Définir le domaine

Un ensemble de compétences tel qu'il est évoqué dans la définition que Popham donne du test critériel (voir *supra*) est appelé domaine.

La distinction entre les objectifs de maîtrise, de transfert et d'expression permet de comprendre aisément qu'on a affaire tantôt à des domaines finis, tantôt à des domaines qui ne le sont pas; ces derniers sont plus nombreux. En effet plus on s'élève dans la hiérarchie qui fonde les taxonomies d'objectifs, plus les objets de l'apprentissage se complexifient.

1 E.W. EISNER, Instructional and expressive educational objectives. *In* J. POPHAM, éd., *Instructional Objectives, op. cit.*, p. 14.

Dans le domaine cognitif, un monde sépare la connaissance de mémoire de faits isolés des comportements d'analyse, de synthèse, de créativité. De même, dans le domaine affectif, on sait la distance entre le respect aveugle d'une règle de conduite et l'exercice du libre arbitre. De même encore, dans le domaine psychomoteur, il y a loin entre marcher en faisant rebondir un ballon et traduire des sentiments et des émotions en mouvements et déplacements gracieux dans l'espace.

Plus haut un apprentissage se situe dans la hiérarchie des comportements, plus il est difficile de définir de façon opérationnelle le domaine dont il relève. Ainsi s'explique que les domaines finis sont relativement rares.

Pour identifier les domaines, on commence généralement par définir les objectifs du programme éducatif à propos duquel il s'agit d'évaluer. On détermine ensuite les contenus, les matières, les comportements correspondant aux objectifs.

A propos de chacun de ces contenus ou ensembles de contenus apparentés, l'idéal serait de formuler toutes les questions possibles, de toutes les manières possibles: demander de calculer l'aire d'un triangle en indiquant la longueur de la base et de la hauteur n'est pas la même chose qu'inviter l'élève à prendre lui-même les mesures nécessaires; la difficulté du langage employé pour poser une même question peut varier considérablement.

La totalité des questions possibles étant posées (réellement ou approximativement), il reste à en tirer au hasard un ou plusieurs échantillons représentatifs. En bonnes conditions, les différents ensembles ainsi tirés constituent des formes parallèles d'un même examen ou test.

On trouvera, dans nos considérations sur le thème « *Objectivité - Subjectivité dans la rédaction des questions d'examens* » (p. 124 sq.) des exemples de techniques de génération de questions.

Plus généralement, pour définir un domaine, on s'efforce de déterminer[1]:

- A quels contenus l'élève aura affaire.
- Dans quelles conditions.

1 E.L. BAKER, Beyond objectives: Domain referenced achievement. *In* W. HIVELY, éd., *Domain referenced testing*, Englewood Cliffs, N.J., Educational Technology Publication, 1974.

- Quelles seront les règles d'échantillonnage du domaine.
- Quelle sera la grandeur du domaine (elle variera considérablement selon le but de l'évaluation: diagnostic fin, bilan général de fin d'année, ...).

Imaginer que ces démarches vont toujours se dérouler de façon rigoureuse, «scientifique», est une illusion. Dans la pratique, le jugement, l'expérience acquise sur le terrain jouent considérablement dans la plupart des cas.

Sur le pouvoir discriminatif des questions

Pour les épreuves normatives, on s'efforce d'établir dans quelle mesure chaque question aide bien à classer les élèves entre eux: on rejette donc les questions auxquelles la grande majorité des élèves, sinon tous, sont capables ou incapables de répondre.

La perspective est très différente pour les tests critériels: le souci majeur est de poser des questions qui permettent de distinguer les élèves qui savent de ceux qui ne savent pas. L'idéal de l'éducateur n'est-il pas de conduire tous les élèves à la maîtrise?

Dans cette perspective resurgit la question de l'égalité des chances. Elle n'est pas assurée si les questions sont biaisées, c'est-à-dire si, par exemple, elles avantagent ou désavantagent les garçons ou les filles, les minorités sociales ou ethniques.

CHAPITRE 4

FIXATION DE LA NOTE DE RÉUSSITE[1]

Un examen ou un ensemble d'examens sont réussis ou non. La note de succès correspond au minimum de performance jugé acceptable, autrement dit à la compétence estimée minimale. Dans une perspective de la pédagogie de la maîtrise, on rencontre un problème similaire quand on se demande à partir de quel moment l'élève a atteint cet état.

Plusieurs appellations désignent ce point critique: *note de réussite, note de coupure, seuil de réussite, niveau d'exigence minimale, ...* La fixation de cette note soulève presque toujours un problème redoutable. Comment, par exemple, déterminer à partir de quel moment précis quelqu'un connaît une langue étrangère?

La question n'est simple que si l'on juge - là où c'est possible - en termes de tout ou rien. Ainsi, on peut décréter qu'un élève n'aura pas réussi aussi longtemps qu'il commet une ou plusieurs erreurs dans la multiplication des dix premiers nombres. On n'est bon pilote d'avion qu'à partir du moment où l'on réussit *tous* les atterrissages.

Dès qu'une part d'erreur est jugée admissible et si l'acquisition à évaluer est complexe, la note de coupure devrait être nuancée sous forme de fourchette à l'intérieur de laquelle on réussit. On parle alors de *zone de coupure*.

La note de réussite est tantôt *absolue* et déterminée *a priori* (on fixe la note avant l'épreuve par rapport à un objectif à atteindre) ou *a posteriori* (on prend pour référence les résultats d'un groupe extérieur ou ceux du groupe même).

1 Pour un traitement approfondi de cette question, voir V. DE LANDSHEERE, *Faire réussir - Faire échouer. La compétence minimale et son évaluation*, Paris, Presses Universitaires de France, 1988.

I. LES DECISIONS EMPIRIQUES

1. NOTE DE REUSSITE AUX EXAMENS TRADITIONNELS

L'une des règles les plus fréquentes dans notre pays était (et est parfois encore) d'obtenir au moins 50 % dans chaque branche et 60 % pour l'ensemble. Aucune justification objective de ces seuils n'est fournie, pas plus d'ailleurs que les raisons de placer ou non toutes les branches sur le même pied.

Parmi les démarches plus systématiques pour fixer la note de coupure, Ebel[1] propose les suivantes:

a) *Pour les tests à choix multiple*

Partant du principe que, dans un test à choix multiple bien construit, un élève ignorant peut obtenir un score correspondant à des réponses au pur hasard (50 % pour des réponses juste-faux; 25 % pour des réponses à quatre choix, etc.) et qu'un élève très fort devrait arriver à peu près au maximum, Ebel estime que la note de coupure peut, dans ce cas, être calculée de la façon suivante:

	Juste - Faux	*4 choix*
Nombre d'items	100	100
Score attendu si on répond au hasard	50	25
Score moyen idéal	(100 + 50) : 2 = 75	(100 + 25) : 2 = 62,50
Score de coupure	(50 + 75) : 2 = 62,50	(25 + 62,50) : 2 = 43,75

Le *score moyen idéal* se situe donc à mi-chemin entre le score maximum possible et le score attendu si l'on répond au pur hasard.

Le *score de coupure* se situe à mi-chemin entre le score moyen idéal et le score attendu si l'on répond au pur hasard.

b) *Pour les examens traditionnels*

Le score de réussite est fixé arbitrairement à un pourcentage donné; ce score est ensuite ajusté en fonction des performances observées.

Exemple: on estime que 3/4 des élèves devraient fournir des réponses correctes, à condition qu'au moins 60 % et qu'au plus 80 % d'entre eux dépassent ce score.

1 R.L. EBEL, *Essentials of Educational Measurement*, Englewood Cliffs, Prentice Hall, 1989, 3e éd. Les données suivantes sont empruntées à V. DE LANDSHEERE, 1989, o.c.

Si au moins 60 % des élèves arrivent à 75 % de réussite, le score de réussite est fixé exactement à la moitié de la distance entre le score correspondant à 75 % et le score qu'au moins 60 % des élèves dépassent.

Si plus de 80 % des élèves dépassent le score de réussite, ce score est alors situé à mi-chemin entre le score correspondant à 75 % et le score que 80 % des élèves dépassent.

Ebel précise (p. 341): «Les chiffres de 60 et 80 % des sujets sont pris purement à titre d'exemple. Selon les circonstances, on peut souhaiter les augmenter ou les diminuer. Le but poursuivi en fixant les valeurs est de maintenir le niveau de connaissances et le pourcentage de réussite dans ce qui semble des limites raisonnables, tout en tenant aussi raisonnablement compte de l'erreur sur la mesure. Cette méthode n'est pas souvent utilisée et, pourtant, elle offre le mérite d'apporter une solution rationnelle à un problème difficile.»

2. NOTE DE REUSSITE AUX EPREUVES DE COMPETENCE MINIMALE

Dans la pédagogie de la maîtrise, le seuil de réussite exigé se situe, en principe, à 100 %. Cette exigence est souvent modérée, ne fût-ce que pour tenir compte des erreurs fortuites, accidentelles. En général, on accepte 90, voire 80 %, mais pas moins.

En évaluation normative, il arrive que le seuil fixé corresponde à la performance moyenne d'un groupe de référence. Par exemple, V. De Landsheere (1988, p. 124) rapporte que, dans plusieurs Etats américains, l'attribution du diplôme de fin d'études secondaires est subordonné à l'obtention, à un test de capacité en lecture, d'une note au moins égale au niveau moyen des classes de 3e année.

V. De Landsheere écrit très justement (p. 130): «Fondamentalement, un seuil d'exigence unique est impossible à fixer parce que le degré de perfection voulu varie en pratique selon l'objectif poursuivi et la nature de l'apprentissage à réaliser. Rares sont les cas simples comme: on sait rouler à bicyclette lorsque, sauf accident, on n'en tombe jamais et lorsqu'on est capable de démarrer et de s'arrêter à volonté. Dans ce cas, le niveau d'exigence est pratiquement de 100 %. Comment appliquer une règle similaire à «Savoir jouer du piano»? Comment justifier que, dans les bulletins scolaires, on demandait une note d'au moins 50 % aussi bien en langue maternelle, en mathématiques, ... qu'en gymnastique et en dessin? Que signifie d'ailleurs savoir à moitié dessiner?»

Bref, aucune des démarches qui viennent d'être décrites n'est pleinement satisfaisante. Tout au plus apportent-elles un peu moins d'anarchie dans les décisions.

II. VERS PLUS D'OBJECTIVITE

Depuis les années 70, de nombreuses recherches ont porté sur les méthodes scientifiques de détermination de la note de coupure et sur le taux d'erreurs que l'on commet quand on les applique. En 1986, Berk[1] relève plus de 35 de ces méthodes. On trouve, dans l'ouvrage que V. De Landsheere (1988) a consacré à la compétence minimale, la présentation et une analyse critique des quinze méthodes qui ont le plus retenu l'attention. Phénomène qui doit beaucoup faire réfléchir les éducateurs: selon la méthode choisie, le taux d'échec varie de 9 à 78 %!...

Nous ne décrivons que la méthode d'Angoff[2], parce qu'elle réunit un ensemble de qualités largement reconnues et qu'elle semble la plus sage.

1. On demande à un groupe d'éducateurs expérimentés d'indiquer, pour chaque question d'un examen ou d'un test, quel pourcentage de chances de répondre correctement a un élève minimalement compétent, dans le groupe à examiner.

10 - 20 - 30 - 40 - 50 - 60 - 70 - 80 - 90 - 100 %

2. Chacun des juges note ses appréciations sur le tableau suivant:

Juge n° 1

Pour 100 élèves, à la limite de la compétence, combien, à votre avis, répondaient correctement aux questions posées?

Questions	*Pourcentage*	*Pourcentage / 100*
1	80 %	0,8
2	40 %	0,4
3	50 %	0,5
4	0 %	0
5	20 %	0,2
	Note minima de réussite:	1,9

1 R.A. BERK, A Consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, 1986, 56, 1, 137-172.

2 W.H. ANGOFF, Scales, Norms and Equivalent Scores. *In* R.L. THORNDIKE, Ed., *Educational Measurement*, Washington, American Council on Education, 1971, 2e éd.

Pour obtenir la note de coupure pour l'ensemble des juges, on calcule la moyenne des notes minimales obtenues.

Dans cette démarche, la subjectivité ne cherche pas à se cacher, ce qui incite en permanence à la pondération dans les décisions.

Saunders[1] propose un perfectionnement intéressant de cette méthode.

Après avoir fait les calculs selon Angoff, les juges se voient offrir trois occasions de réviser leurs appréciations:

1. Après avoir pris connaissance des notes de coupure auxquelles les autres juges ont abouti.
2. Après avoir, en plus, pris connaissance de la distribution des scores obtenus par les élèves qui ont subi le test.
3. Après avoir étudié les statistiques descriptives (moyenne, médian, marge de variation, écart type), élaborées à partir des notes de coupure fixées au stade précédent (principalement pourcentage d'élèves qui réussiraient).

La note de réussite est finalement fixée sur la base des positions prises au stade 3, lors d'une discussion entre tous les juges.

CONCLUSION

Même si elles présentent un indéniable progrès par rapport à l'exigence traditionnelle d'au moins 50 ou 60 % dans chaque branche, les démarches qui viennent d'être décrites restent tâtonnantes.

Une préoccupation majeure doit toujours rester à l'esprit: la fonction première de l'évaluation n'est pas de classer les élèves, mais bien d'informer sur ce qu'ils ont appris effectivement et sur ce qui leur reste à faire pour être au moins minimalement compétents.

Cette notion de «compétence minimale» est, à la limite, aussi insaisissable que celle de la «note vraie», notamment parce qu'elle varie en fonction des contextes. Etre minimalement compétent pour piloter un petit avion de tourisme ne qualifie pas pour prendre les commandes d'un gros avion de ligne... De surcroît, même s'il n'est pas sans utilité, un constat d'incapacité n'offre pas d'intérêt réel pour l'éducateur s'il ne s'accompagne pas d'un diagnostic clair ouvrant la voie à la remédiation.

1 J.C. SAUNDERS et L.L. MAPUIS, Accuracy and consistency of expert judges in setting passing scores on criterion-referenced tests, *Communication à la Conférence annuelle de l'AERA*, La Nouvelle-Orléans, 1984.

CHAPITRE 5

LE CONTROLE DE LA FIDELITE DE L'EXAMEN

Beaucoup de mesures que nous avons envisagées jusqu'à présent tendent à assurer une meilleure justice scolaire.

Le contrôle de la fidélité relève de la même préoccupation : idéalement, un même examen (écrit ou oral), passé plusieurs fois, sans que l'élève ait le temps d'apprendre des choses nouvelles, devrait toujours conduire au même résultat. On souhaite donc qu'il soit aussi *fidèle* qu'un mètre qui, bien utilisé, mesure toujours la même longueur, à une erreur minime près.

Evidemment, faire subir la même épreuve à diverses reprises ne permet pas un bon contrôle de la fidélité. Le premier passage provoque une réflexion génératrice d'apprentissage. Par la suite, l'élève continue à penser au problème, vérifie l'exactitude de ses réponses, prend des informations complémentaires, etc. Bref, il faut recourir à un autre moyen ou, plus exactement, à un ensemble de mesures favorables à la fidélité.

En supposant que soit acquise la fidélité de la correction, il reste à garantir la fidélité des réponses.

1. Eviter toute ambiguïté dans les questions.

Une question qui manque de clarté au point de se prêter à plusieurs interprétations différentes prive, d'avance, l'examen de sa fidélité. Rien ne permet, en effet, de prévoir avec quelque certitude lequel des sens aurait été donné par un même individu ou par un même groupe, à diverses occasions.

Nous avons envisagé, page 100, un certain nombre de moyens d'assurer la clarté des questions.

2. Des questions en nombre suffisant.

Si l'on ne pose qu'un très petit nombre de questions, on laisse de vastes zones de matière inexplorées. Or, il se peut que l'élève n'ait pas

compris toutes les matières ou ait négligé des parties du cours en spéculant sur la chance ou sur la préférence accusée du professeur. Selon l'endroit où tomberont les questions, on enregistrera un brillant résultat ou un échec.

Le seul moyen d'éviter pareille aventure est de couvrir *toute* la matière. En d'autres termes, l'exemple doit bien échantillonner l'ensemble. Ceci concerne à la fois la validité du contenu et la fidélité.

En principe, plus on ajoute de questions, plus la fidélité augmente.

3. Un contrôle mathématique.

Pour des examens importants, on souhaitera acquérir plus de garanties encore.

a) *La méthode pairs-impairs.*

Si le nombre de questions est assez élevé, la démarche suivante simule une situation où un même élève subirait, pratiquement en même temps, deux fois le même examen.

1) Les questions étant numérotées, on les divise en deux groupes: paires et impaires.

2) On dresse deux tableaux parallèles des résultats et l'on calcule leur corrélation (r); autrement dit, on évalue numériquement la relation qui existe entre les deux groupes[1].

3) Les deux examens artificiellement créés sont de moitié plus courts que l'original. En vertu de ce que nous avons vu au paragraphe 2, ils sont donc moins fidèles et la corrélation sous-estime donc la situation réelle.

   Une formule simple permet d'opérer la correction nécessaire:

   l'indice de fidélité est finalement obtenu par $\frac{2r}{1+r}$.

   Plus le résultat sera proche de 1, plus la fidélité sera élevée.

   On exige généralement au moins 0,80.

Quand un contrôle de fidélité se fait après l'examen effectif, une fidélité trop faible incitera les maîtres à un supplément de prudence dans l'interprétation des résultats et à plus de soin dans la construction des examens futurs.

1 La méthode du calcul de la corrélation est expliquée dans tous les manuels de statistique élémentaire.

b) *Deux formes parallèles.*

Le contrôle mathématique est le même que dans la méthode *pairs-impairs*, mais ici les examinateurs préparent deux séries complètes de questions, unanimement jugées équivalentes, à partir de la même grille d'objectifs.

Si la population est assez nombreuse, une moitié, choisie au hasard (par exemple, selon l'ordre alphabétique des noms) reçoit la première forme et l'autre, la seconde. Pour des populations peu nombreuses, on invite les élèves à répondre aux deux séries de questions après quelques jours. Cette seconde solution est évidemment beaucoup moins rigoureuse.

4. **Répétition de la notation.**

La fidélité des réponses ne suffit pas. Elle doit être accompagnée de la fidélité de la notation. On contrôle celle-ci, soit en invitant un même professeur à noter deux fois les mêmes travaux à un intervalle de quelques jours ou de quelques semaines, soit en comparant les notes accordées par plusieurs correcteurs à un même travail.

5. **La théorie de la généralisabilité.**

Sans être inexact, ce qui vient d'être dit à propos de la fidélité n'est, en réalité, qu'une première approche, relativement grossière, de la problématique. En outre - on va y revenir dans le chapitre suivant -, la fidélité n'est pas garante de la validité.

Les résultats observés à un examen varient notamment:
- selon les questions posées (facette questions);
- selon le moment où l'épreuve est subie (facette occasions);
- selon les points de vue et la qualité des examinateurs (facette juges).

Par généralisabilité d'une mesure, on entend le degré auquel on peut, à partir d'une observation ou épreuve particulière, tirer des conclusions sur la valeur théorique recherchée.

Par exemple, à propos de la facette questions, il importe de considérer:

1) Le langage utilisé

- Influence du vocabulaire utilisé et de la syntaxe (lisibilité psychologique).

- Mode de présentation matérielle et psychologique. Par exemple, bien que portant exactement sur le même problème, les questions suivantes peuvent donner des résultats différents:
  - 3 + 5 = ?
  - J'achète un crayon à cinq francs et un bonbon à trois francs, combien dois-je payer?

2) La nature de la question

J. Cardinet aime à donner l'exemple suivant à propos du calcul de l'aire du triangle. Dans un groupe qui a étudié cette question pendant l'année scolaire, on obtient des résultats différents selon que l'on demande:

- Quelle est la formule de calcul de l'aire du triangle?
- Quelle est la surface d'un triangle de 8 m de base et de 5 m de hauteur?
- Voici un triangle et une règle graduée. Calculez la surface de ce triangle.
- Même question que la précédente, mais la forme du triangle est telle que la perpendiculaire abaissée à partir du sommet tombe en dehors de la base.

Bref, connaître une formule ne suffit pas à témoigner d'une capacité réelle. Seules des performances concrètes, suffisamment variées dans leurs conditions pour constituer un échantillon représentatif des cas qui peuvent se présenter, permet une évaluation fidèle et valide.

A propos de cette facette «occasions», on peut notamment poser les questions suivantes:

- L'élève a-t-il eu effectivement l'occasion d'apprendre ce dont il va être question?
- Cet apprentissage est-il récent ou déjà ancien? Dans ce second cas, la connaissance a-t-elle été réactivée ou est-elle restée dormante?
- Quel pouvait être le degré de fatigue de l'élève quand il a subi cette épreuve?
- Une même question est-elle posée sous des formes diverses?

Quant à la facette «juges», elle a déjà été évoquée: inconsistance des appréciations, effet de halo, effet de contraste, stéréotypie, effet de fatigue du correcteur ou de l'interrogateur, ...

CHAPITRE 6

CONTROLE DE LA VALIDITE

La fidélité d'un examen ne garantit en rien sa validité. Un mètre mal construit, qui mesurerait cinq centimètres de trop, indiquerait la même longueur à chaque mesure, mais conduirait néanmoins à une conclusion fausse.

En bref, valider un examen, c'est prouver qu'il mesure effectivement ce pour quoi il est proposé. Acquérir cette certitude est capital: car si la conclusion est négative, tous les efforts déployés pour l'organisation et le passage des examens sont vains ou, au moins, détournés de leur objet.

Or, la validation n'est pas chose aisée. Faisant le total de notre savoir sur la validité, R. Cox conclut, en 1969, que la recherche a à peine touché aux questions fondamentales[1]. C'est encore partiellement vrai aujourd'hui.

Selon la nature rétrospective ou prospective de l'examen, on distingue la validité du contenu et la validité prédictive.

I. La validité du contenu.

Elle intéresse principalement les examens destinés à dresser le bilan des acquisitions et donc aussi d'un enseignement. Tel but avait été assigné, tels objectifs choisis. Ont-ils été atteints ? C'est la question à laquelle l'évaluation critérielle s'efforce de répondre prioritairement.

Dans les examens portant sur de longues étapes d'apprentissage, on ne peut évidemment interroger en détail sur toute la matière. Non seulement il importe donc de localiser les points principaux, mais encore faut-il tenir compte de leur importance relative. Si l'on satisfait à ces deux critères, les questions pourront susciter des réponses *représentatives* de l'ensemble des acquisitions, de la compétence totale.

1 R. COX, Reliability and Validity of Examinations, in J. LAUWERYS et D. SCANLON, *Examinations, o.c.*, p. 43.

En d'autres mots, la validité du contenu dépend de la qualité et de l'adéquation de l'échantillonnage des connaissances et des capacités.

Sans oublier que, pour un point de connaissance donné, il importe aussi d'échantillonner les principaux comportements qui les mettent en œuvre. Un examen qui ne porterait que sur la seule mémoire, en négligeant donc les processus intellectuels supérieurs, ne serait pas valide au sens général de ce terme.

Nous avons vu que, dans un enseignement bien conçu, les objectifs à atteindre ont été définis, au moins de façon provisoire, au début de l'année.

Mais, à ce moment, il s'agissait d'intentions que les circonstances peuvent avoir modifiées. Avant d'essayer de dresser le bilan du travail de ses élèves, le professeur fera donc d'abord le sien. Faute de ce retour sur soi-même, les examens n'ont qu'une validité illusoire ou, pire encore, une validité de façade, destinée à dédouaner le professeur aux yeux de ses supérieurs. Ainsi s'explique que des questions portent sur des matières à peine effleurées, mais figurant au programme officiel, ou sur des habiletés cognitives supérieures que l'on a omis d'installer patiemment, systématiquement pendant l'année. Combien d'échecs scolaires ne sont-ils pas dus aux belles «questions d'intelligence», posées au terme d'un enseignement qui ne l'a pas cultivée ?

Si la définition précise des objectifs permet une validation de contenu bien meilleure que dans le passé, celle-ci n'en reste pas moins encore très limitée dans l'état actuel de nos connaissances. D'abord, parce qu'il est bien difficile d'acquérir la certitude que tel comportement réel est bien représentatif de tel trait que l'on ambitionnait de cultiver. Ensuite, parce que, même si la vision est correcte au départ, les vicissitudes de la notation peuvent toujours trahir les intentions.

Un professeur d'histoire ou de géographie, écrit P. Vernon[1], qui, par un examen écrit, essaie de juger la compréhension profonde de ces disciplines, attribue souvent une bonne partie des points pour la reproduction exacte de faits détaillés, la restitution de ses théories favorites, la longueur des réponses, leur intelligibilité et l'élégance du style dans lequel elles sont exprimées. Dans un examen oral, combien les points ne dépendent-ils pas de l'intelligence sociale, du contrôle de soi-même et, spécialement, de l'aptitude de l'élève à créer, par la parole et l'attitude, une bonne relation avec l'examinateur ?

1 P. VERNON, Types of Examination, in J. LAUWERYS et D. SCANLON, *Examinations*, p. 43.

Que faire, en pratique ?

D'abord, suggérer que, pour chaque question d'examen qu'il rédige, le professeur s'interroge sur l'objectif poursuivi.

Ensuite, soumettre les questions et le barème de notation à plusieurs éducateurs expérimentés (dont un au moins appartiendra à un groupe de disciplines différent du domaine concerné) qui jugeront indépendamment de la validité du contenu des questions. En cas de désaccord grave, la question doit être rejetée[1].

L'expérience montre qu'il faut actuellement s'en tenir à un petit nombre d'objectifs généraux conçus de façon assez large. Sinon, les juges ne réussissent plus à se mettre d'accord.

Dans certains cas, aucun accord ne sera obtenu sur les questions fondamentales. Ce n'est sans doute pas un motif suffisant pour cesser l'enseignement de la matière ou de la branche incriminée, mais l'impossibilité d'accord doit alors être clairement indiquée.

On risque, par exemple, de rencontrer pareille difficulté à propos du contrôle systématique de la réalisation des objectifs assignés aux langues anciennes, à certaines matières de la mathématique, etc.

De nouveau, nous constatons qu'en tentant d'accroître la rigueur scientifique des examens, des enseignements peuvent être remis en cause.

II. **La validité prédictive.**

Elle semble, de loin, la plus simple à contrôler, car elle n'exige pas de compréhension profonde des phénomènes. Même si l'on ne s'explique pas le rapport existant entre le succès à tel test et une réussite brillante dans un domaine déterminé, le fait peut être aisément observé.

Qu'il s'agisse d'un examen donnant accès à un nouveau cycle d'études ou autorisant l'exercice d'une profession, il suffit de suivre les individus pendant un certain temps pour savoir si le pronostic formulé à partir des notes attribuées se vérifie ou non.

On constate avec étonnement que semblable vérification est rarement faite. J.-C. Passeron y voit le signe que l'examen est moins destiné à mesurer objectivement la capacité qu'à servir la reproduction sociale au profit des classes privilégiées[2].

1 Un coefficient de concordance entre les avis peut être calculé par la formule de Kendall et un seuil d'acceptation choisi de commun accord.

2 J.-C. PASSERON, *Sociologie des examens*, o.c., p. 7.

Pour n'être pas un leurre, la vérification systématique de la validité prédictive exige le contrôle rigoureux des variables, faute de quoi les variables cachées fausseront les conclusions. Par exemple, on tirait encore récemment argument en faveur du latin en faisant la statistique du nombre d'universitaires brillants qui avaient fait des études secondaires classiques. Or, cette observation ne prouve rien, sinon que, traditionnellement dans nos pays, les élèves qui obtenaient les meilleurs résultats à l'école primaire ou qui étaient socialement favorisés étaient orientés vers ce type d'études.

Nous venons de voir que la prédiction se valide par la réalisation ultérieure du prédit qui constitue donc le critère. Dans le monde physique, on dispose en général de critères sans ambiguïté: il n'est pas difficile de savoir si le beau temps prévu existe réellement. En éducation, des critères aussi clairs sont rares, du moins si l'on ne se contente pas de signes finalement peu représentatifs de la réalité. Quand saura-t-on, par exemple, que tel individu dont on a prédit qu'il deviendrait un « bon » enseignant l'est effectivement devenu ? Il n'existe manifestement pas un seul modèle de bon enseignant. Il ne pourra être réputé bon que lorsqu'il travaille dans certaines circonstances, avec un certain type d'élève. Et puis, la nature même du critère fondamental varie : on peut être un enseignant de qualité, soit parce qu'on suscite un maximum d'apprentissages cognitifs, soit parce qu'on crée un climat affectif propice au développement de la personnalité, à l'apparition d'attitudes positives, d'intérêt vif pour le savoir, ...

Enfin, pour les tests critériels, la maîtrise observée à un moment donné subsistera-t-elle à moyen et à long termes ? On sait que la résistance des apprentissages dépend, principalement, d'une part, de la façon dont ils ont été réalisés (on retient mieux ce que l'on a découvert, ce que l'on a construit, au sens piagétien) et, d'autre part, de la fréquence d'utilisation des acquis (on peut avoir maîtrisé la technique de résolution des équations du second degré à un moment de la scolarité et ne jamais plus avoir l'occasion d'utiliser cette habileté dans la vie adulte).

Bref, établir la validité prédictive n'est souvent simple que dans son principe...

En guise de synthèse, nous empruntons la récapitulation suivante à l'*Examination Bulletin n° 3*[1].

1 Londres, H.M.S.O., 1964, pp. 19-20.

FACTEURS DE VALIDITE D'UN EXAMEN

1. Identification adéquate des objectifs dont la réalisation doit être vérifiée par l'examen écrit, l'évaluation du travail de l'année, l'épreuve pratique, etc.
2. Parmi toute la gamme des objectifs ainsi identifiés, sélection de ceux sur lesquels la vérification se concentrera.
3. Evaluation efficace de l'adéquation du contenu et de la structure de l'examen au but poursuivi.
4. Relation claire entre chaque question et les objectifs de l'enseignement.
5. Elaboration d'un barème de notation et rédaction de directives aux correcteurs en fonction des objectifs. Un niveau suffisant de fidélité doit aussi être assuré.
6. Bonne connaissance des capacités des candidats.
7. Disposition des notateurs à tenir compte des jugements indépendants du leur et des données objectives qui leur seraient fournies.
8. Comparaison avec des épreuves antérieures dont la validité a été prouvée.

A propos du point 4, signalons que l'on accorde de plus en plus d'importance à un autre aspect: l'élève a-t-il eu effectivement *l'occasion d'apprendre*? En effet, il ne suffit pas que la matière figure au programme, il faut encore qu'elle ait été enseignée.

D.F. WALKER[1] estime que cinq conditions doivent être remplies pour que l'occasion d'apprendre ait vraiment existé.

1. La matière en question doit avoir été abordée de façon répétée.
2. Son importance doit avoir été signalée aux élèves; ils doivent être prévenus qu'elle fera probablement l'objet d'une question d'examen, et il doit en être effectivement ainsi.
3. La matière doit être traitée de façon à pouvoir être apprise par les élèves.
4. Un temps suffisant doit lui être consacré pour que l'apprentissage puisse se produire.
5. Des révisions périodiques, des occasions de réapprentissage et d'exercice des habiletés importantes doivent être offertes, y compris des exercices de remédiation.

1 D.F. WALKER, What constitutes Curricular Validity. *In* G.F. MADAUS, Ed., *The courts validity and minimum competency testing*, Boston, Kluwer-Nijhoff, 1983.

QUATRIEME PARTIE

LES PROCEDURES DE MODERATION

L'HARMONISATION OU L'EQUILIBRATION DES ECHELLES DE NOTATION

CHAPITRE 1

POSITION DU PROBLEME

1. Définition.

L'harmonisation ou l'équilibration des échelles de notation (*modération*) a d'abord eu pour objet de tempérer les excès de sévérité ou de générosité chez certains examinateurs.

Au sens large, le terme modération désigne l'ensemble des mesures prises pour rendre comparables les notes d'examens internes et donc pour unifier leur signification au niveau des différentes classes de même type dans une école, dans un groupe d'écoles, dans des établissements similaires d'une région ou d'un pays.

Surtout dans les systèmes scolaires où le redoublement est imposé si les résultats d'examens sont jugés gravement insuffisants, des mesures doivent être prises pour éviter des injustices graves.

On sait, en effet, que les professeurs interprètent les programmes officiels notamment en fonction de la qualité des élèves qu'ils reçoivent. Le niveau de difficulté de l'enseignement et des exigences varie selon que le groupe-classe paraît « fort » ou « faible ». Et c'est louable. N'importe-t-il pas de stimuler chacun en tenant compte de ce qu'il peut réellement ?

Toutefois, cette louable préoccupation peut conduire à des aberrations. Si, dans une classe de 3e primaire, le professeur réussit à travailler, dans certains domaines, au niveau de la 5e, il risque d'oublier qu'un élève de 3e ne devrait échouer que s'il n'a pas satisfait aux exigences minimales du programme de 3e année.

A. Grisay[1] a montré expérimentalement que, faute de tenir compte de ce principe, un même niveau de connaissances fait échouer un élève dans une classe forte et le classerait premier dans une classe faible. C'est notamment pour éviter cette injustice que les procédures d'harmonisation des notes (*modération*) sont conçues.

1 A. GRISAY, *Rendement en français, notes et échecs à l'école primaire : les mirages de l'évaluation scolaire*, Liège, Service de Pédagogie expérimentale de l'Université de Liège, 1982.

Avec Noizet et Caverni[1], on peut distinguer la *modération a priori*, c'est-à-dire les mesures d'harmonisation prises avant l'évaluation, et la *modération a posteriori*, où les mesures interviennent correction faite.

Les procédures *a priori* sont les plus anciennes. La plus spontanée est la *concertation* préalable entre correcteurs qui peuvent s'entendre sur les objectifs de l'épreuve, sur les aspects sur lesquels ils feront porter leur jugement, sur l'importance relative des différents critères et, éventuellement, sur le système de compte ou de décompte des points.

Si ces deux dernières conditions sont présentes, on aboutit ainsi à un *barème de correction*. En cas d'examen externe, notamment, le barème est élaboré par les responsables de la rédaction des questions et est fourni aux correcteurs. On peut considérer que la batterie d'échelles d'évaluation descriptives évoquées à propos de la notation de la composition française constitue un barème d'évaluation sophistiqué.

Les recherches les plus fines sur la *modération a posteriori* sont certainement dues aux docimologistes britanniques. Dans un système complètement décentralisé comme celui de la Grande-Bretagne, les programmes et les méthodes des écoles primaires ou secondaires peuvent différer considérablement. Aussi, pendant très longtemps, des examens externes, c'est-à-dire organisés en dehors des écoles par des commissions spéciales, ont dû sanctionner les études aux moments cruciaux de la scolarité. Jusqu'à un passé récent, le fameux 11 + *Examination*, examen d'entrée dans le secondaire, décidait, en particulier, de l'admission dans l'enseignement général classique. Le C.S.E. (*Certificate of Secondary Education*) est requis pour entrer dans l'enseignement supérieur. Ce dernier examen pourvoit chaque élève d'un document signalant le niveau dans chacune des branches par rapport à des normes nationales.

Pour admettre un élève dans une section déterminée, les universités annoncent leurs exigences particulières : ici, des notes d'excellence (A) sont demandées en mathématiques, en biologie et en langue maternelle ; là, les notes « très bien » (B) sont aussi acceptées, etc.

Ce système d'examens externes est critiqué, parfois très sévèrement, parce qu'il limite gravement la liberté pédagogique des maîtres et des autorités locales.

2 G. NOIZET et J.-P. CAVERNI, o.c., pp. 47 sqq.

Aussi, depuis quelques années, on tente de les remplacer par des examens internes dont les résultats sont rendus comparables par des procédures de modération locale, régionale, puis nationale, expérimentalement mises au point.

Dans les systèmes éducatifs centralisés, où un programme d'études unique est imposé, il est beaucoup plus facile, si on le veut, de modérer les examens. Avant de voir comment, il n'est peut-être pas inutile de rappeler pourquoi une modération est souhaitable.

En gros, la raison est double. La première concerne les individus, élèves et parents, qui ont le droit de connaître le niveau «réel» des performances scolaires, avant de décider de l'orientation ultérieure dans les domaines des études ou de la profession. Qui se croirait qualifié pour les jeux Olympiques parce qu'il a gagné une course organisée entre quelques amis? Un constat défavorable n'est d'ailleurs pas plus une condamnation sans appel qu'un diagnostic de faiblesse ou de maladie. Il faut savoir à temps qu'un problème se pose pour en chercher les causes, puis les remèdes, s'ils existent.

La seconde raison concerne la communauté. De même qu'un consommateur ne peut être trompé sur la marchandise, de même la société ne peut être tenue ou à payer fort cher les échecs d'un étudiant entré à l'université sur la foi d'un certificat invalide, ou à confier une fonction à quelqu'un qui n'est pas capable de la remplir de façon satisfaisante

2. **Modérer n'est pas caporaliser.**

On pourrait craindre que la volonté de rendre les résultats scolaires comparables ne provoque une résurgence des vieilles contraintes. On se souvient de cet inspecteur, de tradition napoléonienne, qui, consultant sa montre, croyait pouvoir dire: «A cet instant, on enseigne telle leçon dans toutes les cinquièmes années primaires de France.» On sait les faiblesses d'un enseignement caporalisé.

Mais la nécessité d'une plasticité pédagogique n'empêche pas qu'au-delà des aménagements circonstanciels, chaque cycle d'études poursuit des objectifs fondamentaux, communs à tous: acquisition d'habiletés intellectuelles de base ou de connaissances et de capacités jugées essentielles. Maîtres et modérateurs doivent s'interroger et s'entendre sur ces apprentissages cruciaux.

3. **Modération volontaire ou imposée?**

Normaliser tous les examens à tous les moments des études ne semble ni souhaitable, ni d'ailleurs possible. Peu souhaitable, parce

que les élèves progressent à des rythmes parfois très différents et s'accommodent mal du découpage rigoureux en périodes et en années scolaires. Impossible parce que la modération, même par les méthodes économiques qui vont être présentées, reste un travail lourd, dans sa préparation et dans son exécution.

La modération doit être imposée aux moments décisifs, avant tout avant d'imposer un redoublement ou lors de l'attribution de diplômes ou de certificats de fin de cycle.

Pour le reste, la décision devrait être laissée soit aux maîtres individuellement, soit aux chefs d'établissement ou aux autorités locales.

4. **La modération commence au début de l'année scolaire.**

Nous venons d'y faire allusion, il importe que professeurs et modérateurs aient pu trouver un accord sur les objectifs de l'enseignement et sur quelques grands principes de notation pour que les examens soient comparables.

Des affirmations vagues telles que: «L'élève doit pouvoir s'exprimer correctement par écrit», sont sans utilité. La capacité désirée doit être traduite en termes de comportements concrets que l'on aura d'ailleurs intérêt à incorporer dans les échelles d'évaluation descriptives. S'il s'agit de l'expression écrite, bien des questions se posent:

Comment graduer les exigences en fonction de l'âge des élèves et du type d'école?

L'originalité sera-t-elle récompensée? Comment? Comment sera-t-elle identifiée?

Quelle importance attribuer à l'orthographe?

La longueur des travaux de composition sera-t-elle considérée? Exige-t-on une longueur minima, en dessous de laquelle des points seront décomptés? Combien?

Quelle importance accorder à la richesse du vocabulaire? Comment la définir? Quelles seront les exigences syntaxiques?

Les mathématiciens peuvent aussi accorder leurs violons. Si, pour nous limiter à un seul exemple, certains professeurs pénalisent gravement les fautes d'opérations, voire les fautes d'orthographe dans la solution de problèmes, alors que leurs collègues estiment que ces aspects sont secondaires, des notes égales risquent de recouvrir des réalités très différentes.

Les programmes officiels peuvent aussi jouer un rôle important en précisant les exigences minimales pour chaque niveau d'études. L'efficacité de ces données de base est d'autant plus grande que le système scolaire est centralisé.

Inutile de se faire des illusions: les objectifs comportementaux des différentes branches ne se définissent pas en quelques heures de méditation. Des commissions, où enseignants, inspecteurs et chercheurs uniront leurs efforts, devront travailler longtemps avant d'arriver à un résultat satisfaisant. Beaucoup de questions risquent d'ailleurs de rester provisoirement sans réponse.

La définition des objectifs fera l'objet d'une recherche permanente, non seulement parce qu'ils pourront se préciser en fonction de l'avancement de la psychologie de l'apprentissage, mais aussi parce que les objectifs mêmes évoluent en fonction des transformations de la société et des progrès de la psychologie.

Les instructions relatives aux examens se préciseront parallèlement.

5. **Pas de comparabilité sans fidélité élevée.**

Il n'est pas concevable de comparer entre eux des résultats d'examens qui, pris isolément, seraient éminemment fluctuants.

Le problème de la fidélité a été discuté dans la partie consacrée à la construction de l'examen. Nous n'y revenons pas.

6. **Peut-on se fier aux tests?**

Dans les systèmes de modération que nous allons examiner, les tests d'intelligence ou de connaissances occupent une place importante. En certaines occasions, ils apportent les points de repère qui servent à ajuster les résultats; en d'autres cas, on leur accorde, à eux seuls, autant de valeur qu'au travail de l'année et qu'à l'examen final.

Les observations suivantes expliquent pareille décision. On admet avec raison qu'en général, les maîtres jugent bien leurs élèves. Pourtant, un simple test verbal, administré en moins d'une heure, permet une prédiction presque aussi sûre, pour certains types d'études générales au moins.

	Corrélation avec les résultats de l'élève:	
	2 ans après	*3 ans après*
Pronostic du maître	*0,821*	*0,748*
Test verbal	*0,796*	*0,722*

Voici encore quelques résultats de recherches confirmant ces observations.

A moyen terme, un test d'intelligence ou, mieux encore, une combinaison de scores à différents tests d'intelligence sont meilleurs prédicteurs du succès scolaire que les résultats d'examens. Cette supériorité a été démontrée à plusieurs reprises, surtout par des chercheurs anglo-saxons:

1) Emmet[1] montre qu'un test d'intelligence verbal permet de mieux prévoir les résultats, après deux ou trois ans d'enseignement secondaire général, que les examens de langue maternelle et d'arithmétique évalués par les maîtres.

2) Emmet et Wilmut[2] ont, par la suite, fait une démonstration tout aussi convaincante de la valeur prédictive du test d'intelligence, cinq ans après.

3) Wrighley[3] a confirmé ces résultats et montré que la prédiction peut être meilleure encore si les résultats à un test d'intelligence et à des tests de connaissances standardisés peuvent être combinés.

4) En Belgique, les psychologues scolaires ont connu, jusqu'à ces derniers temps, la valeur prédictive des scores verbaux (V) et de raisonnement (R) du Test d'Aptitudes Mentales Primaires (P.M.A.) de Thurstone. On observait, en effet, qu'en doublant le score obtenu au test verbal et en ajoutant le score obtenu au test de raisonnement, on obtenait un score global (2 V + R) fortement corrélé avec les résultats obtenus par les élèves dans l'enseignement secondaire général.

Nous ne disposons pas de recherches similaires pour l'enseignement technique ou pour des structures d'enseignement peut-être mieux adaptées à la civilisation de l'an 2000. Il est très possible que les scores d'autres types de tests devraient être utilisés.

1 W. EMMET, *An Inquiry into the Prediction of Secondary School Success*, London, Univ. of London Press, 1942.

2 W. EMMET and F. WILMUT, The Prediction of School Certificates Performance in Specific Subjects, in *British Journal of Educ. Psychol.*, 22, 1952, 52-62.

3 J. WRIGHLEY, The Relative Efficiency of Intelligence and Attainment Tests as Predictors of Success in Grammar Schools, in *British Journal of Educ. Psychol.*, 25, 1955, 107-116.

CHAPITRE 2

QUELQUES SYSTEMES DE MODERATION DES EXAMENS

I. Par référence à un ou plusieurs tests

A. La formule la plus libérale: le système suédois de modération par branche à partir de tests de connaissances.

Dès la fin de la Seconde Guerre mondiale, la Suède a adopté un système de modération, simple et facultatif, applicable par chaque professeur, dans sa classe, tout au long de la scolarité.

Toutes les notes des professeurs sont attribuées selon une échelle de 7 degrés[1].

On suggère que le pourcentage d'élèves recevant une note déterminée soit, en gros, le suivant[2], le professeur restant libre de tenir compte des caractéristiques de sa classe (par exemple: tête très forte, ou presque tous élèves moyens, etc.):

Notes	1	2	3	4	5	6	7
Pourcentage	1	6	24	38	24	6	1

Par ailleurs, une batterie de tests de connaissances portant sur les branches principales du programme est construite chaque année par un office central où collaborent des spécialistes et des professeurs expérimentés. Cette batterie est étalonnée sur un échantillon national représentatif; les résultats sont répartis en 7 classes, dans la proportion conseillée pour les notes des professeurs.

Si le professeur soumet sa classe au test national, il dispose donc de normes directement comparables aux notes qu'il a attribuées. Il lui est donc facile d'ajuster ces dernières.

1 Nous nous référons à l'exposé très clair de S. HENRYSSON, The Swedish System of Equalising Marks, in *Educational Research*, VI, 2, Feb. 1964, 156-160.

2 Distribution normale (voir courbe de Gauss).

Voici un exemple concret proposé par S. Henrysson:

24 élèves			
Notes	*Distribution des notes préliminaires attribuées par le professeur*	*Distribution notes obtenues au test*	*Distribution des notes après ajustement*
7	–	1	–
6	–	4	4
5	4	4	4
4	4	9	10
3	14	5	4
2	2	1	2
1	–	–	–
Moyenne	*3,42*	*4,33*	*4,17*

Le professeur inscrit d'abord, dans la 2e colonne, les notes qu'il a attribuées. La moyenne (3,42) fournit une première indication, si la moyenne nationale est connue. (En Suède, elle était de 4 pour la 6e primaire, au moment où le présent ouvrage a été écrit. Il semble donc que le professeur soit ici trop sévère).

Voici maintenant les résultats obtenus au test par les mêmes élèves:

Score brut	*Note correspondante*	*Nombre d'élèves obtenant cette note*
94 - 100	7	1
84 - 93	6	4
67 - 83	5	4
46 - 66	4	9
30 - 45	3	5
22 - 29	2	1
0 - 21	1	–

Ces résultats sont reportés dans le premier tableau, colonne 3.

La moyenne est 4,33, ce qui confirme la première impression de sévérité.

Sans changer l'ordre du classement initial, le professeur rectifie ses notes (colonne 4).

Le système suédois séduit pour plusieurs raisons:

1° Les maîtres gardent leur entière liberté:
   a) de faire passer ou non le test (presque tous le font);
   b) de tenir compte des résultats;
   c) de communiquer les résultats des tests aux élèves, aux autres professeurs, au directeur, aux parents.

Ainsi, le test devient simplement un outil mis à la disposition des praticiens. La tentation de «bachotage» est donc réduite au minimum.

2° Les constructeurs des tests veillent à couvrir une large gamme d'objectifs échantillonnant bien tout le programme.

Il est évident que les professeurs prêtent une attention particulière aux matières traitées dans les tests (effet de reflux - *backwash effect*). On dispose donc ainsi d'un moyen efficace pour sensibiliser le maître à certaines innovations.

Remarquons que, depuis 1965, un système fort semblable est expérimenté dans des classes genevoises de 5e et de 6e primaire (10-12 ans) pour l'orthographe et l'arithmétique[1].

Parmi les critiques formulées contre ce système d'ajustement[2], on retiendra:

1. Si un professeur a jugé un des élèves de sa classe avec une sévérité excessive, la référence à l'épreuve nationale ne réparera pas l'injustice.
2. L'épreuve de référence peut ne pas mesurer les mêmes aspects que les notes.
3. La construction de l'épreuve de référence peut aussi être défectueuse: fidélité trop basse, ...

B. Système imposé de modération par branche à partir d'un test de connaissances.

A partir des mêmes données de base que celles de la Suède (examens internes et tests de connaissances, pour les branches principales, étalonnés nationalement ou régionalement), le système directif suivant peut donner de bons résultats.

1 Voir S. ROLLER, Le problème de l'attribution des notes scolaires. Essai de solution, Genève, in *Docimologie et éducation*, numéro spécial de la revue *Les Sciences de l'Education*, avril-septembre 1969, pp. 66 sqq.

2 W. ANGOFF, Can Usual General Purpose Equivalency Tables be Prepared for Different College Admission Tests, in A. ANASTASI, Ed., *Testing Problems in Perspective*, Washington, ACE, 1966, pp. 251-264.

F. BACHER, *La normalisation des notes, o.c.*, p. 63.

Une commission nationale ou régionale de modération prend un certain nombre d'écoles en charge.

Ces écoles lui envoient les résultats aux épreuves préparées librement par les professeurs et les scores obtenus aux tests de connaissances.

En cas de différence, en plus ou en moins, égale ou supérieure à deux écarts types (par exemple) entre les moyennes aux examens et aux tests, l'école reçoit la visite des modérateurs.

Leur mission n'est pas de dire au directeur comment il doit conduire son école, ni au professeur comment faire son cours, mais bien d'attirer l'attention sur un fait et de tâcher de trouver, en pleine collaboration avec l'école, l'explication et, si possible, le remède.

Si le désaccord subsiste, le droit d'ajuster d'office les notes peut être donné aux autorités régionales, l'école ayant, de son côté, le droit d'interjeter appel.

C. Un système de sélection à partir d'un test d'intelligence.

Nous avons déjà observé qu'en général, les maîtres jugent bien leurs élèves, mais qu'ils tendent à relativer leurs jugements par rapport au niveau global de la classe.

Le système suivant, mis au point par la Fondation nationale anglaise pour la Recherche en Education[1] pour l'admission dans l'enseignement secondaire général classique, permet de *sélectionner* en tenant compte des jugements portés indépendamment par les maîtres, dans leur classe.

1. L'instituteur classe ses élèves par ordre de mérite, c'est-à-dire selon son évaluation globale de la chance de réussite dans l'enseignement secondaire général. Plusieurs élèves peuvent être classés *ex aequo* (Classement 1).
2. Les élèves subissent un test d'intelligence verbal (2).
3. Les scores au test sont classés par ordre décroissant (Classement 3).
4. Les classements (1) et (3) sont placés côte à côte.
5. Le score d'intelligence qui tombe en face du nom de l'élève est considéré comme score d'évaluation par l'instituteur (Jugement ajusté).

1 Voir A. YATES et D. PIDGEON, *Admission to Grammar Schools, o.c.*

Classement par l'instituteur (1)	*Score obtenu par l'élève au test verbal (2)*	*Classement des scores par ordre décroissant (3)*	*Jugement ajusté (4)*
A	121	132	132
B	120	128	128
C	132	121	121
D	128	120	120
E	100	106	106
F	106	100	100
G } *ex aequo*	94	100 } Total	97 }
H } *ex aequo*	82	96 } Total 290	97 }
I } *ex aequo*	96	94 } Total 290	97 }
J	100	86	86
K } *ex aequo*	78	82 } Total	78 }
L } *ex aequo*	79	79 } Total 312	78 }
M } *ex aequo*	86	78 } Total 312	78 }
N } *ex aequo*	73	73 } Total 312	78 }
O	65	65	65

On voit que le classement auquel on aboutit finalement ne modifie en rien l'ordre initialement choisi par l'instituteur. Mais, cette fois, nous disposons d'un moyen de comparaison entre écoles.

Utilisation des jugements ajustés.

Exemple: dans un lycée, on souhaite recruter, pour la 6e latine, 25 élèves provenant de trois écoles primaires différentes.

Jugements ajustés

	Ecole primaire I	*Ecole primaire II*	*Ecole primaire III*
	132		
130			130 x 2
	128		126
125		125	125
		124	123
	121	121	
120	120		120
		119	
		117	116
115	115 x 3	115	
			114
	112		113
110		110	110
	106	107	
105			
		104	
100	100		

Pour sélectionner 25 élèves, on coupe, dans ce cas, à 110. Si, à la coupure de 110, des *ex aequo* avaient placé plus de 25 élèves dans la catégorie des sélectionnés, la justice la plus élémentaire aurait exigé qu'on les admette tous.

Tracer une ligne de démarcation comme on vient de le faire est une source d'injustice pour une autre raison: le hasard seul n'est-il pas responsable du placement d'un certain nombre de sujets juste au-dessus ou en dessous de la limite?

Le psychométricien est armé pour surmonter cette difficulté: il prend une marge de sécurité de trois fois l'erreur type qu'une formule simple permet de calculer à partir du coefficient de fidélité du test utilisé[1].

II. Modération par appel à une banque d'*items*

Dans le cas où les maîtres utilisent des questions sous forme d'*items* à choix multiple, la démarche suivante, actuellement expérimentée en Grande-Bretagne, semble pleine de promesses[2]. Toutefois, elle requiert l'existence d'un service de recherche, nécessité dont un éducateur averti ne peut d'ailleurs plus douter aujourd'hui.

1. Les maîtres indiquent sur une grille les objectifs poursuivis.
2. Ils envoient cette grille au service de recherche, en même temps que les *items* qu'ils ont rédigés et, éventuellement, déjà prétestés localement.
3. Le service examine ces *items* et, selon les possibilités, met certains d'entre eux à l'épreuve; leur difficulté et leur pouvoir discriminatif sont calculés.
4. Le service renvoie au maître:
   a) les *items* examinés;
   b) des *items* complémentaires dont les indices de facilité et d'efficacité sont connus pour une population donnée: ville, canton, pays, ...

   Ces *items* fourniront un point de comparaison à partir duquel le résultat enregistré pour les autres pourra être ajusté.

1 Erreur type = $\sigma \sqrt{1 - r}$.
où σ = écart type des scores.
r = coefficient de fidélité. Ce coefficient est généralement indiqué dans le manuel qui accompagne le test.
2 Voir D. PIDGEON et A. YATES, o.c.

Il est intéressant de noter qu'après s'être étroitement cantonnés dans la rédaction d'*items* à réponse fermée, par choix multiple, les Britanniques commencent à proposer des questions semi-ouvertes. Il s'agit, en fait, de questions ouvertes, très soigneusement formulées, pour lesquelles on a identifié expérimentalement les grands types de réponses, en fonction desquelles des barèmes de correction sont proposés.

L'organisation d'une banque d'*items* est lourde pendant les premières années. Par la suite, la provision d'*items* étalonnés devient telle que le travail s'allège et que le système acquiert une grande souplesse et une grande rapidité de fonctionnement.

Selon les fluctuations des programmes et de l'enseignement, la difficulté des questions peut varier en relativement peu de temps. Il importe donc que l'indice de difficulté soit réajusté, continuellement, en fonction des résultats observés lors de l'utilisation de chaque *item*.

III. Procédure d'équilibration

En Angleterre, un système de modération complet.

Le système que nous allons décrire maintenant est, lui aussi, dû à la Grande-Bretagne. A notre connaissance, il n'en existe pas de plus complet: il porte sur la préparation de l'examen en collaboration avec les écoles, la notation, l'ajustement des notes et des grades finals.

Le but poursuivi est de perfectionner les *examens internes*, au point de pouvoir leur confier le rôle joué jusqu'à présent par les grandes épreuves externes.

Pour en arriver à une modération nationale, on procède par paliers: à l'échelon local où un certain nombre d'écoles poursuivant une même finalité se groupent; à l'échelon régional ensuite, selon une technique qui ne diffère pas fondamentalement de celle appliquée localement; pour le passage au niveau national s'ajoute, à la technique de modération proprement dite, un échantillonnage fin dont l'étude technique serait ici hors de propos.

Pour notre pays, nous suggérons que quelques écoles de même esprit commencent par faire une expérience volontaire du système. C'est pourquoi nous concentrons notre présentation sur le processus de base.

La modération interécoles ne serait faite qu'aux moments cruciaux de la scolarité et pour quelques branches principales[1]. Rien n'empêche toutefois les professeurs d'une même école, enseignant les mêmes cours, d'utiliser spontanément la même méthode de correction.

A. Préliminaires.

Nous avons déjà fait allusion aux accords à prendre sur les objectifs et sur la construction de l'examen. Nous n'y revenons plus.

Chaque école choisit en son sein un professeur qui jouera deux rôles:

1° Coordonnateur des examens dans son école;
2° Membre de la commission de modération interécoles.

Un portrait idéal du modérateur n'existe pas. Les qualités suivantes paraissent souhaitables[2]:

1. maturité générale et bonne expérience pédagogique;
2. contacts fréquents avec des élèves du niveau et du type d'école concernés par l'examen;
3. capacité de formuler clairement ses critères et ses jugements;
4. bienveillance mais fermeté;
5. capacité de discuter sans passion, sans créer de tension;
6. aptitude à comprendre quelques techniques d'analyse statistique.

Dans chaque école, les questions d'examens sont rédigées avec une grande liberté: seuls la grille des objectifs principaux, le nombre et le volume des questions sont à respecter selon la convention prise.

Au début au moins, il est souhaitable de soumettre les questions à la commission de modération qui s'assure de l'unité générale.

B. Les professeurs notent leurs examens.

1 Cette proposition a pris une actualité nouvelle au cours des dernières années. En effet, on accorde de plus en plus d'attention à la vérification des compétences minimales, à différents niveaux de la scolarité et à la fin des cycles. Voir à ce propos V. DE LANDSHEERE, *Faire réussir - Faire échouer. La compétence minimale et son évaluation*, Paris, P.U.F., 1988.

2 MATHER, o.c., p. 67.

C. **Correction par les modérateurs.**

Trois principes dominent le travail :

1° L'intervention dans les examens des écoles doit être aussi discrète que possible. Seuls les cas de divergences importantes méritent l'attention.

2° Les correcteurs doivent, en gros, être de même sévérité, apprécier les mêmes qualités et être d'accord sur la signification des grades finals.

3° Les échantillons à recorriger doivent être petits, et le travail statistique doit être aussi simple que possible.

Dans chaque école et pour un même type d'examen, on prélève, au hasard, 20 copies corrigées.

Soit le cas de 12 lycées[1]. Chacun a délégué un modérateur à la commission, laquelle reçoit donc 12 paquets de 20 copies.

La première opération vise à déterminer l'équivalence des modérateurs à trois points de vue :

1° *sévérité* : son degré est révélé par la moyenne;

2° *discrimination* : notation trop désinvolte ou trop prudente. Révélée par la dispersion ou marge de variation des notes;

3° *conformité* : un même élève est-il classé de la même façon par tous les correcteurs ? Révélée par la corrélation entre deux séries de notes.

Pour vérifier l'accord entre modérateurs, à ces trois points de vue, les douze recorrigent d'abord un même paquet de 20 copies.

Les 20 copies sont réparties, toujours au hasard, en cinq groupes de quatre. Dans chaque groupe, les élèves sont classés par ordre alphabétique.

1 Exemple emprunté à l'*Examinations Bulletin n° 5*, Londres, HMSO, 1965.

On fait ensuite les simples opérations suivantes, dans l'ordre où elles figurent dans le tableau :

ETUDE DE L'ACCORD ENTRE LES MODERATEURS

Notes accordées par 12 modérateurs

Nom du candidat	MODERATEURS a	b	c	d	e	f	g	h	i	j	k	l	Moyenne
A	4	4	4	3	4	4	3	1	4	4	4	4	4
B	2	2	2	2	3	2	1	1	2	2	2	2	2
C	6	5	5	5	5	5	5	5	5	5	5	5	5
D	2	1	1	2	2	3	3	2	2	3	2	3	2
Total	14	12	12	12	14	14	12	9	13	14	13	14	13
Marge de variat.	4	4	4	3	3	3	4	4	3	3	3	3	3
E	3	3	3	4	3	5	3	3	3	3	3	3	3
F	1	1	1	1	1	1	1	1	1	1	1	1	1
G	5	6	6	6	5	3	6	6	6	6	6	6	6
H	3	4	4	4	3	4	3	4	4	2	4	4	4
Total	12	14	14	15	12	15	13	14	14	12	14	14	14
Marge de variat.	4	5	5	5	4	4	5	5	5	5	5	5	5
I	2	2	2	2	4	3	2	3	3	3	3	3	3
J	2	1	2	1	3	2	2	2	2	2	2	2	2
K	5	5	5	6	5	5	5	6	5	5	5	4	5
L	1	2	2	1	2	2	2	3	2	2	3	2	2
Total	10	10	11	10	14	12	11	14	12	12	13	11	12
Marge de variat.	4	4	3	5	3	3	3	4	3	3	3	2	3
M	4	2	3	4	3	4	5	5	5	4	5	4	4
N	4	3	4	4	4	5	4	4	4	4	4	4	4
O	5	4	4	3	4	3	4	4	4	5	4	5	4
P	5	6	5	5	5	5	5	5	5	5	5	5	5
Total	18	15	16	16	16	17	18	18	18	18	18	18	17
Marge de variat.	1	4	2	2	2	2	1	1	1	1	1	1	1
Q	1	2	1	1	1	1	1	1	1	1	2	1	1
R	1	1	2	1	1	2	1	1	1	1	1	1	1
S	3	3	2	3	3	2	2	3	3	3	3	3	3
T	2	1	2	1	2	2	1	2	2	2	2	2	2
Total	7	7	7	6	7	7	5	7	7	7	8	7	7
Marge de variat.	2	2	1	2	2	1	1	2	2	2	2	2	2
Total général	61[2]	58	60	59	63	65	59	62	64	63	66	64	63[1]
Somme des marges	15[4]	19	15	17	14	13	14	16	14	14	14	13	14[3]

Différences entre les notes accordées par chaque modérateur et la moyenne des notes accordées par tous les modérateurs à un même candidat.

Nom du candidat	MODERATEURS											
	a	b	c	d	e	f	g	h	i	j	k	l
A	0	0	0	-1	0	0	-1	-3	0	0	0	0
B	0	0	0	0	1	0	-1	-1	0	0	0	0
C	1	0	0	0	0	0	0	0	0	0	0	0
D	0	-1	-1	0	0	1	1	0	0	1	0	1
Marge de variat.	1	1	1	1	1	1	2	3	0	1	0	1
E	0	0	0	1	0	2	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0
G	-1	0	0	0	-1	-1	0	0	0	0	0	0
H	-1	0	0	0	-1	0	-1	0	0	-2	0	0
Marge de variat.	1	0	0	1	1	3	1	0	0	2	0	0
I	-1	-1	-1	-1	1	0	-1	0	0	0	0	0
J	0	-1	0	-1	1	0	0	0	0	0	0	0
K	0	0	0	1	0	0	0	1	0	0	0	-1
L	-1	0	0	-1	0	0	0	1	0	0	1	0
Marge de variat.	1	1	1	2	1	0	1	1	0	0	1	1
M	0	-2	-1	0	-1	0	1	1	1	0	1	0
N	0	-1	0	0	0	1	0	0	0	0	0	0
O	1	0	0	-1	0	-1	0	0	0	1	0	1
P	0	1	0	0	0	0	0	0	0	0	0	0
Marge de variat.	0	1	0	0	0	0	0	0	0	0	1	0
Q	0	1	0	0	0	0	0	0	0	0	1	0
R	0	0	1	0	0	1	0	0	0	0	0	0
S	0	0	-1	0	0	-1	-1	0	0	0	0	0
T	0	-1	0	-1	0	0	-1	0	0	0	0	0
Marge de variat.	0	2	2	1	0	2	1	0	0	0	1	0
Somme des marges[5]	4	7	5	6	4	8	6	5	1	4	3	3

Contrôles.

Les règles suivantes n'ont rien de magique. Pour les formuler, G. Peaker s'est inspiré des techniques de contrôle de qualité utilisées dans l'industrie. Des expériences très poussées dans le domaine des examens ont permis de les adapter.

Le contrôle ainsi réalisé est raisonnable, expéditif, efficace, mais non très fin. Aussi, si malgré la grossièreté des critères, un aspect de la correction paraît inacceptable, il y a certainement un problème !

1° *Sévérité.*

L'expérience révèle que les efforts d'harmonisation doivent porter surtout sur ce point. On va le voir, en cas de problème, la solution est heureusement facile.

Règle :

a) Partir de la moyenne[1] des totaux généraux (ici 63);
b) Ne pas tolérer des écarts au-delà d'une marge centrale de 10 points[1] par rapport à cette moyenne, soit 5 points en plus ou en moins (ici : 58-68).

Constatation :

Tous les totaux généraux sont compris dans cette marge.
Tous les correcteurs sont donc d'une sévérité acceptable.

La solution en cas d'excès de sévérité ou de générosité peut être la suivante.

Si l'on constate que des correcteurs ne tombent pas dans cette marge, on fait revoir les copies de chacun par deux autres correcteurs qui eux satisfont au critère.

La note est obtenue en faisant la moyenne des trois évaluations.

2° *Discrimination.*

Règle :

Le total des marges moyennes (3) ne peut pas être supérieur au double du total des marges d'un correcteur (4) et réciproquement.

Constatation : aucun problème.

3° *Conformité.*

Règle :

Le total des marges des différences (5) entre les notes attribuées par un modérateur et la moyenne des notes attribuées par l'ensemble ne doit pas être supérieur à 12[1].

Constatation : aucun problème.

1 Les correcteurs notent 20 compositions de 1 à 5. Dix points équivalent à une différence moyenne d'un demi-point par copie pour l'ensemble des 20 copies.

Conclusion.

Dans le cas présent, tous les modérateurs ont surmonté les trois épreuves de contrôle. Ils pourront donc travailler seuls.

Où en sommes-nous ? Des douze échantillons de 20 copies, un échantillon est maintenant corrigé définitivement (douze notateurs l'ont vu !).

Comme chaque modérateur peut travailler seul, la suite des opérations ira vite.

D. **Nouvelle correction des échantillons restants et contrôle.**

Les opérations sont pratiquement les mêmes que pour le contrôle des modérateurs. Nous donnons, toutefois, un exemple détaillé parce que la présentation plus concise (un seul modérateur et un seul professeur) donne une meilleure vue d'ensemble.

1 Nous montrons p. 224 que l'on obtient avec ce repère empirique une bonne estimation de la corrélation entre deux séries de notes.

CALCULS POUR LA COMPARAISON
ENTRE UN PROFESSEUR ET UN MODERATEUR
(1 est la meilleure note ; 5 la moins bonne !)

5 groupes de 4	Elèves choisis au hasard	Note attribuée par Modérateur		Note attribuée par Professeur		Différence (Modérateur-Professeur)
1er groupe	Henri	Min.	5	5	Min.	0
	Jean		4	4		0 Min.
	Paul	Max.	2	1	Max.	1
	Pierre		3	2		1 Max.
	Marge : Min-Max	3	14	12	4	Marge des différences = 1
2e groupe	André		2	1	Max.	1 Max.
	Edouard	Max.	1	1		0 Min.
	Jules		1	1	Min.	0
	René	Min.	4	4		0
	Marge : Min-Max	3	8	7	3	Marge des différences = 1
3e groupe	Antoine		3	3	Max.	0 Max.
	Camille		5	5		0
	Eugène	Max.	2	3		- 1[1] Min.
	Jérôme	Min.	6	6	Min.	0
	Marge : Min-Max	4	16	17	3	Marge des différences = 1
4e groupe	Jacques	Min.	5	5	Min.	0
	Laurent	Max.	1	1	Max.	0 Min.
	Martin		4	3		1 Max
	Victor		4	3		1
	Marge : Min-Max	4	14	12	4	Marge des différences = 1
5e groupe	Bruno	Max.	3	2	Max.	1
	Hugues	Min.	4	2		2 Max.
	Léon		4	5	Min.	- 1 Min.
	Simon		3	2		1
	Marge : Min-Max	4	14	12	4	Marge des différences = 3
Les cinq groupes réunis	Total des grades		66 (1)	59 (2)		Total des marges de différences = 7 (5)
	Marges réunies	15	(3)	(4)	17	
		15	(3)	(4)	17	

* On considère que (0) est plus grand que (- 1).

Contrôle.

1° *Sévérité des correcteurs.*

La différence entre total (1) et total (2) ne doit pas être supérieure à 10.

Ici, 66 - 59 = 7: acceptable.

2° *Discrimination.*

Le total (3) ne peut pas être supérieur au double du total (4) et réciproquement.

Ici, 15 et 17: acceptable.

3° *Conformité.*

Le total des marges des différences (5) ne doit pas être supérieur à 12.

Ici, = 7: acceptable[1].

4° *Conclusion.*

Dans ce cas, aucun problème ne semble se poser. Les notes du professeur sont donc acceptées et les opérations sont terminées.

Si les notes du professeur ne semblent pas acceptables, l'idéal est d'engager une discussion avec lui afin de rechercher un accord.

Ce n'est pas toujours possible, en particulier faute de temps.

La solution suivante est plus expéditive. Si un modérateur constate un problème, il invite quatre de ses collègues à recorriger indépendamment le même échantillon de vingt copies. On calcule la moyenne entre les cinq modérateurs, et le professeur est invité à ajuster toutes ses notes en fonction de celles des modérateurs.

Selon G. Peaker, les cas nécessitant pareille correction deviennent de plus en plus rares, à mesure que le système de modération fonctionne, en raison des échanges de vues fructueux qu'ils provoquent chaque fois.

1 Cette façon simple de procéder évite le long calcul de la corrélation entre les deux séries de notes.

On verra d'ailleurs que l'on obtient un coefficient de corrélation très proche du coefficient calculé par la méthode classique en faisant l'opération:

$$\frac{15^2 + 17^2 - 7^2}{2 \times 15 \times 17} = 90$$

E. Comment ajuster des notes discordantes?

Il s'agit, soit d'un simple ajustement de la moyenne (ou du médian) par relèvement ou abaissement des notes de tous les élèves (sévérité ou générosité excessives), soit aussi d'une modification de la distribution générale, pour améliorer la discrimination.

1° *Ajustement du médian.*

Pour un même examen, la situation est la suivante:

	Quartile supérieur	*Médian*	*Quartile inférieur*	*Ecart type approximatif*
Modérateur	16	13	11	$\sigma = 4$
Professeur	11	8	6	$\sigma = 4$

Le professeur est plus sévère que le modérateur, mais il discrimine aussi bien. Si le professeur ajoute 5 points à chaque élève, le parallélisme est rétabli.

2° *Ajustement du médian et de l'écart type.*

Situation de départ:

	Quartile supérieur	*Médian*	*Quartile inférieur*	*Ecart type approximatif*
Modérateur	16	13	11	$\sigma = 4$
Professeur	16	12	9	$\sigma = 5$

Pour aligner sur le modérateur, on corrige d'abord pour le médian; un point est ajouté à tous les élèves.

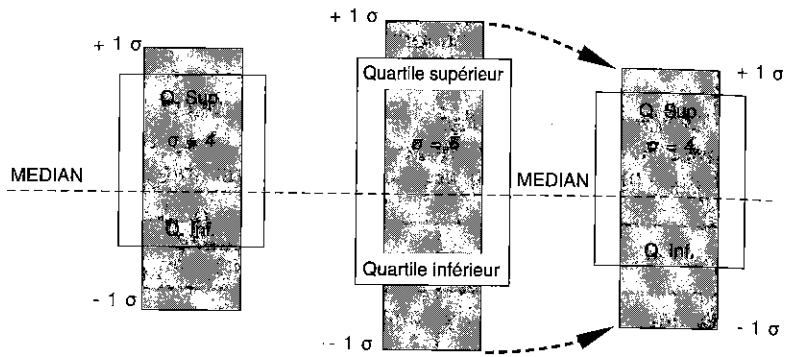
Supposons que les notes de départ du professeur sont:

... 19 18 17 16 15 14 13 **12** 11 10 9 8 7 6 5 ...

médian

1 Adapté d'après D. MATHER *et al., o.c.*, p. 183.

Nous allons d'abord montrer par le dessin en quoi consiste l'ajustement de l'écart type.



I. Notes du modérateur

II. Notes du professeur, augmentées d'un point. Le médian est devenu le même que celui du modérateur. Reste à ajuster σ.

III. Notes du professeur diminuées une seconde fois pour aligner l'écart type sur celui du modérateur.

Pour l'opération III, l'ajustement en fonction du nouvel écart type est un peu plus compliqué que pour le médian. La confection d'une table de conversion facilitera les opérations.

Pour trouver la nouvelle note correspondant à + 1 σ, il suffit d'ajouter le nouveau sigma au nouveau médian, soit 13 + 4 = 17. La nouvelle note correspondant à -1 σ = 13 - 4 = 9, etc.

Pour les notes correspondant à des fractions de sigma, on calcule d'abord l'écart par rapport au médian, on multiplie par 4/5 et on ajoute le résultat au médian. L'exemple suivant va éclairer cette phase.

	Notes de départ du professeur	*Notes ajustées en fonction du nouveau médian*	*Second ajustement en fonction du nouvel écart type (σ)*
	.	.	.
	.	.	.
	19	.	.
	18	.	.
+ 1 σ	17	18	17 *
	16	17	16 **
	15	16	15 ***
	14	15	15 ****
	13	14	14
Médian	12	13	13
	11	12	12
	10	11	11
	9	10	11
	8	9	10
- 1 σ	7	8	9
	6	.	.
	5	.	.
	.	.	.
	.	.	.

Comment a-t-on trouvé ces nombres ?

* Nouveau médian : 13
Nouvel écart type : 4 $\qquad$ 13 + 4 = 17

** Dans la 2e colonne, 17 est à 4 points du médian.

Multiplier 4 par $\frac{4}{5}$ = 3,2.

Médian + 3,2 = 16, 2 arrondi à 16.

*** 16 - 13 = 3; $3 \times \frac{4}{5} = 2{,}4$; 13 + 2,4 = 15,4 arrondi à 15.

**** 15 - 13 = 2; $2 \times \frac{4}{5} = 1{,}6$; 13 + 1,6 = 14,6 arrondi à 15.

F. **La note de fin d'année.**
Travail de l'année + travaux pratiques + test.

Problème[1]

On veut classer les élèves en cinq groupes de mérite à la fin de leurs études. A sera le grade supérieur et vaudra 1, *E* sera le grade inférieur et vaudra 5.

On veut tenir compte de trois éléments, évalués chacun selon les mêmes échelles à cinq degrés :

- travail de l'année = T.A.
- travaux pratiques = T.P.
- test régional = T.R.

Le test régional se voit attribuer autant d'importance que les deux autres éléments. D'où la pondération :

T.A. = 1/4
T.P. = 1/4
T.R. = 1/2

1 D'après D. MATHER et *al., o.c.*, pp. 149-154.

Voici le tableau général des grades et le tableau des résultats après pondération et ajustement final.

ATTRIBUTION DU GRADE FINAL

Elève N°	*Test régional T.R.*		*Evaluation par l'école*		*Pondération*			*Total*	*Grade final*
	Score	Note	T.A.	T.P.	T.R. x 1/2	T.A. x 1/4	T.P. x 1/4		
1	59	2	3	3	1	1	3/4	2 3/4	3
2	77	1	1	1	1/2	1/4	1/4	1	1
3	51	3	4	3	1 1/2	1	3/4	3 1/4	3
4	12	6	5	5	3	1 1/4	1 1/4	5 1/2	6
5	53	3	3	4	1 1/2	1	1	3 1/2	4
6	40	4	2	5	2	3/4	1 1/4	4	4
7	66	1	2	2	1/2	1/2	1/2	1 1/2	1
8	60	2	3	4	1	3/4	1	2 3/4	3
9	38	5	4	5	2 1/2	1 1/4	1 1/4	5	5
10	70	1	2	2	1/2	1/2	1/2	1 1/2	1
11	56	2	2	3	1	1/2	3/4	2 1/4	2
12	69	1	2	2	1/2	1/2	1/2	1 1/2	1
13	44	4	4	5	2	1	1 1/4	4 1/4	4
14	64	2	1	2	1	1/4	1/2	1 3/4	2
15	19	6	5	5	3	1 1/4	1	5 1/2	6
16	49	3	2	4	1 1/2	3/4	1	3 1/4	3
17	54	3	2	1	1 1/2	1/2	1/4	2 1/4	2
18	47	4	1	2	2	1/4	1/2	2 3/4	3
19	52	3	3	1	1 1/2	3/4	1/4	2 1/2	2
20	48	3	3	3	1 1/2	3/4	3/4	3	3
21	50	3	1	2	1 1/2	1/4	1/2	2 1/4	2
22	24	6	3	4	3	3/4	1	4 3/4	5
23	61	2	4	2	1	1	1/2	2 1/2	2
24	57	2	1	3	1	1/4	3/4	2	2
25	42	4	3	5	2	3/4	1 1/4	4	4
26	35	5	5	2	2 1/2	1 1/4	3/4	4 1/2	4
27	45	4	4	1	2	1	1/4	3 1/4	3
28	41	4	2	4	2	3/4	1	3 3/4	4
29	27	5	5	5	2 1/2	1 1/4	1 1/4	5	5
30	43	4	3	2	2	1	1/2	3 1/2	4
31	67	1	1	1	1/2	1/4	1/4	1	1
32	31	5	5	3	2 1/2	1 1/4	1	4 3/4	5
33	72	1	2	2	1/2	1/2	1/2	1 1/2	1

Les notes pondérées seront ajustées en fonction du test régional (T.R.).

La lecture du tableau nous montre que l'*élève 1* obtient les trois notes pondérées: 1, 1 et $^3/_4$. Soit au total $2\ ^3/_4$; grade final 3*. *L'élève 3*: $1\ ^1/_2 + 1 + ^3/_4 = 3\ ^1/_4$; grade final 3.

Pourquoi un dernier ajustement au moment d'attribuer le grade final? Parce que l'addition des grades provoque une nouvelle concentration vers la moyenne; le rétrécissement est tel que, dans une échelle finale à 5 degrés, toute la population risque de se retrouver au milieu.

L'exemple fictif suivant illustre ce phénomène.

ELEVES	*GRADE*		
	Test R.	Trav. année	Moyenne
Pierre	1	5	3
Paul	2	4	3
Jean	3	3	3
Roger	4	2	3
Henri	5	1	3
Dispersion	4	4	0

Si l'on se reporte au tableau p. 172, on observe que, parmi les grades pondérés, on rencontre un grand nombre de 1 (17 en tout), alors qu'on n'en retrouve plus que deux dans la colonne *Total*. Les notes d'excellence ont été noyées par l'addition des grades.

Le procédé suivant assure une meilleure justice distributive entre les élèves d'une même classe et un meilleur alignement sur le niveau régional révélé par le test régional, celui-ci étant reconnu comme l'étalon le plus sûr.

* Les chiffres en *italiques* dans le tableau indiquent que le modérateur a, avec l'accord de l'école, ajusté la note du T.A. afin de l'harmoniser avec le niveau moyen dans la région.

On part du tableau suivant.

TEST REGIONAL		TOTAL DES GRADES PONDERES		GRADE FINAL	
Grade (1)	*N. d'élèves ayant obtenu ce grade (2)*	*Grade (3)*	*N. d'élèves ayant obtenu ce grade (4)*	*N. d'élèves auxquels il est attribué (5)*	*Grade (9)*
1		1			1
		$1\ ^1/_4$ $1\ ^1/_2$ $1\ ^3/_4$			
2		2			2
		$2\ ^1/_4$ $2\ ^1/_2$ $2\ ^3/_4$			
3		3			3
		$3\ ^1/_4$ $3\ ^1/_2$ $3\ ^3/_4$			
4		4			4
		$4\ ^1/_4$ $4\ ^1/_2$ $4\ ^3/_4$			
5		5			5
Au-delà	Au-delà				Non classé

Opérations.

1. Pointer dans la colonne (2) le nombre d'élèves et faire le total.
2. Commencer à pointer dans la colonne (4) et s'arrêter quand le total égale celui de la colonne (2). On pratique la première coupure à cet endroit. Afin de ne pas désavantager certains élèves, on dépasse, au besoin, le nombre de la colonne (3), de façon à épuiser le niveau où l'on s'est arrêté.
   On verra, par exemple, dans le tableau ci-dessous qu'on accorde le grade 2 à 7 élèves et non à 6, parce que 2 d'entre eux ont obtenu 2 1/2. On ne pouvait évidemment attribuer un grade 2 à l'un et un grade 3 à l'autre.

TEST REGIONAL			TOTAL DES GRADES PONDERES		GRADE FINAL	
Grade (1)	*N. d'élèves ayant obtenu ce grade (2)*		*Grade (3)*	*N. d'élèves ayant obtenu ce grade (4)*	*N. d'élèves auxquels il est attribué (5)*	*Grade (9)*
1	IIII I	6	1	// 2	6	1
			1 1/4 1 1/2 1 3/4	//// 4		
2	IIII I	6	2	/ 1 / 1 /// 3 // 2	7	2
			2 1/4 2 1/2 2 3/4	/// 3		
3	IIII II	7	3	/ 1 /// 3	7	3
			3 1/4 3 1/2 3 3/4	// 2 / 1		
4	IIII II	7	4	// 2 / 1 / 1	7	4
			4 1/4 4 1/2 4 3/4	// 2		
5	IIII	4	5	// 2	4	5
Au-delà	III	3	Au-delà		2	Non classé

Conclusion.

Aucun des systèmes décrits n'est parfait, mais tous permettraient d'améliorer notre système traditionnel de notation.

Un premier choix s'opérera en fonction du but poursuivi : sélection ou comparaison. Ce second aspect a surtout retenu notre attention.

Nous ne nous prononçons pas en faveur d'un système particulier. La décision appartient aux autorités politiques et pédagogiques et aux enseignants.

Un effort intense d'expérimentation doit être fait à tous les niveaux : petits groupes de professeurs, recherches régionales et nationales en collaboration avec les centres psycho-médico-sociaux et les laboratoires de pédagogie expérimentale des universités.

A mesure que ces travaux avanceront, on verra probablement émerger un nouveau système tenant compte de nos traditions et aussi des objectifs particuliers à notre pays. De puissants centres de recherche pédagogique régionaux faciliteraient évidemment des travaux de ce genre.

CINQUIEME PARTIE

UNE PEDAGOGIE DE LA MAITRISE

LE DANGEREUX MYTHE DE LA COURBE DE GAUSS

Dans les sciences humaines, la courbe en cloche de Gauss joue un rôle considérable, parce qu'elle est l'image même de la répartition de bien des aptitudes et des qualités: les individus moyens abondent, mais les génies et les idiots, les géants et les nains sont rares.

La courbe de Gauss est, soit le reflet de la loi du hasard qui préside à notre naissance, soit la résultante de l'influence d'un grand nombre de facteurs agissant de façon plus ou moins indépendante sur un individu ou un objet.

Comme les tests mesurent souvent des aptitudes, des traits de personnalité ou des performances de vastes populations, et servent à classer les individus en les comparant les uns aux autres, il est naturel que ces épreuves soient étalonnées selon la répartition gaussienne: en gros, 70% de moyens, 13% de bons, 13% de médiocres, 2% d'excellents, 2% de très mauvais.

Au cours de la construction de tels tests, on élimine notamment les questions qui seraient réussies par trop ou trop peu de sujets. Le but poursuivi est de classer chacun, de lui attribuer la place qui lui revient dans un groupe nombreux. Bref, il s'agit d'organiser une sorte de concours, où le plus fort occupera nécessairement la première place.

C'est pourquoi beaucoup de tests d'aptitudes ou d'inventaire de connaissances sont d'excellents *instruments de sélection.*

Dans sa classe, l'enseignant poursuit un objectif totalement différent. Son idéal n'est-il pas que tous les élèves apprennent à lire, à calculer et, de façon générale, à maîtriser parfaitement les savoirs, les savoir-faire et les savoir-être nécessaires à leur développement personnel et utiles pour la vie en société? *Eduquer, instruire, n'est pas sélectionner.* Au contraire! C'est s'efforcer que *tous* réussissent. *C'est donc lutter contre la courbe de Gauss* prise comme modèle de sélection.

Les conséquences pédagogiques de ces observations sont particulièrement importantes.

CHAPITRE 1

EVOLUTION DE LA COURBE DES CONNAISSANCES

Quand, dans notre système de classes rigides, un maître reçoit, le jour de la rentrée scolaire, un groupe d'élèves qu'il ne connaît pas, il se trouve normalement devant... deux courbes: l'une représente la distribution des aptitudes et l'autre celle du savoir.

I. La courbe des aptitudes.

Dans son acception habituelle - que nous retenons provisoirement -, le mot *aptitude* désigne des caractéristiques, innées ou acquises, considérées comme symptomatiques de la capacité d'un individu à acquérir un niveau de compétence plus ou moins élevé, dans un domaine déterminé.

Dans l'enseignement non étroitement spécialisé (il va jusqu'à passé vingt ans pour bon nombre de nos élèves), la largeur même de l'éventail des connaissances et des habiletés à faire acquérir (mathématiques, langues, sciences naturelles, arts, ...) rend impossible la sélection très rigoureuse selon une aptitude particulière.

Aussi, jusqu'à un niveau fort avancé de la scolarité, les aptitudes des élèves restent-elles distribuées au hasard.

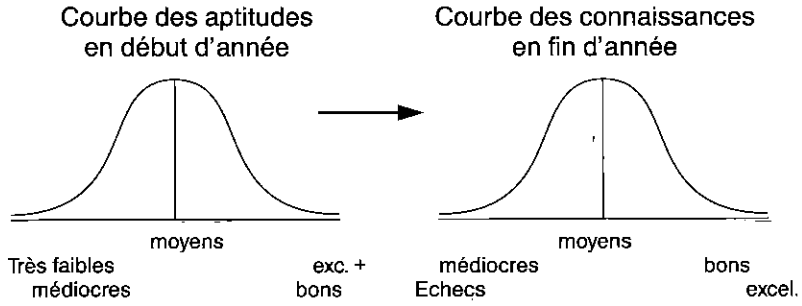
Dans ces conditions, un professeur de mathématiques, par exemple, qui mesurerait les aptitudes spéciales pour sa branche obtiendrait normalement une distribution gaussienne, ou - vu le nombre probablement peu élevé d'élèves - une ébauche de celle-ci.

Les professeurs n'ont d'ailleurs guère besoin de recourir à des tests pour connaître cette situation; ils savent d'expérience que les moyens seront majorité et que les excellents sont rares...

Toutefois, une singulière distorsion se produit généralement dans les esprits. On considère cette répartition des aptitudes comme pronostic des résultats scolaires en fin d'année et on fixe le niveau de l'enseignement de telle façon que ce pronostic se vérifie: il sera «moyennement» difficile tout en permettant aux meilleurs de s'épanouir et en laissant une mince chance aux «médiocres».

Insistons-y, la difficulté «moyenne» est déterminée par la moyenne des aptitudes du groupe considéré et non par une moyenne de difficulté *objective* des notions à enseigner[1].

Dans ces conditions, si le professeur fait le *même cours* à toute la classe, il est normal que la *courbe des connaissances* acquises en fin d'année respecte, à son tour, la distribution gaussienne.



La vocation de l'enseignement est-elle ainsi respectée?

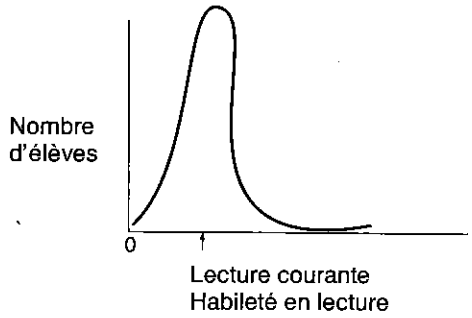
II. La courbe des connaissances.

Revenons au premier jour de l'année scolaire et, au lieu de considérer les aptitudes des élèves, examinons leurs connaissances.

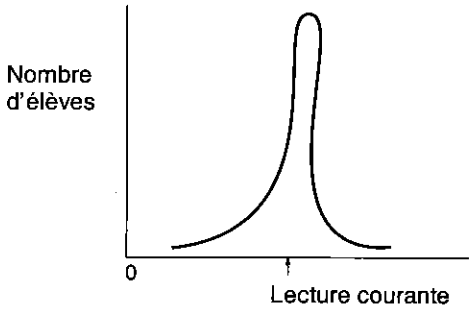
En toute logique, le rôle du professeur est de susciter l'apprentissage de connaissances *nouvelles*. Certes, imaginer que tous les individus formant une classe possèdent exactement la même quantité et qualité de connaissances est utopique. Néanmoins, le système de classes fixes que nous pratiquons repose sur l'hypothèse que *tous* se trouvent approximativement au même niveau. Sinon, comment oserions-nous encore dispenser le même enseignement à chacun?

Et en réalité? Prenons le cas de l'entrée en première primaire. La majorité des enfants ne savent pas lire; quelques-uns sont en bonne voie; deux ou trois lisent déjà couramment. A ce moment, la courbe de la capacité en lecture épouse, en gros, la forme de la lettre *i*.

1 Ainsi s'expliquent les différences considérables du niveau moyen selon les classes et les régions, dont nous avons parlé dans la première partie de ce livre.

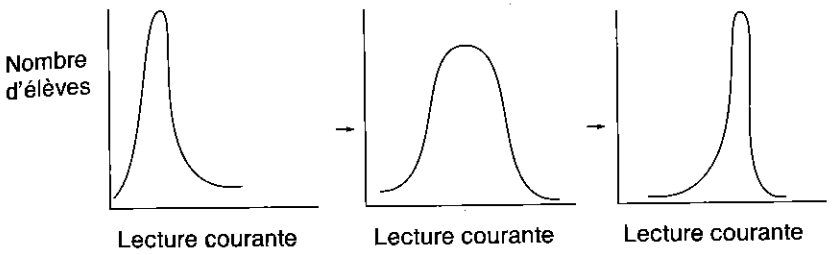


Or, bien que la courbe des aptitudes spécifiques à la lecture soit fort probablement gaussienne, l'instituteur n'admet pas d'emblée que, seule, une partie de la classe pourra apprendre à lire, au contraire. Pour autant que l'on ne verse pas dans le perfectionnisme, on peut dire qu'après un an ou deux, la grande majorité des élèves sauront lire couramment. La courbe des connaissances aura complètement changé de forme: elle ressemblera à un *j*.



Entre la courbe en *i* et la courbe en *j*, il est probable qu'un moment a existé où les mieux doués ont avancé le plus vite, où les moins doués ont traîné et où les moyens se sont situés entre les deux.

L'évolution est donc schématiquement la suivante:



Habileté en lecture.

Dans le cas de la lecture, on ne conçoit pas que les parents ou les responsables de l'enseignement puissent se contenter d'une évolution différente.

Mais à partir de quel niveau de la scolarité pareille exigence n'est-elle donc plus de mise?

CHAPITRE 2

UNE PEDAGOGIE DE LA COURBE EN J

Un examen dont les résultats se distribuent selon la courbe en cloche de Gauss se prête bien à la sélection, au concours.

Or, pour trois raisons historiques principales, l'esprit de concours a imprégné notre enseignement pendant des siècles:

1° Pour des raisons socio-économiques, une partie seulement de la population scolarisable avait accès à l'école. Il y a cent ans, les familles modestes faisaient la première sélection en ne laissant finir l'école primaire qu'aux plus doués de leurs enfants, c'est-à-dire à ceux qui réussissaient le mieux leurs examens. Les bourses d'études, parcimonieusement distribuées, se gagnaient en concours.

2° Jusqu'à ces derniers temps - disons avant l'ère de l'ordinateur -, tous les pays industrialisés disposaient d'un énorme surplus de matière grise. On exploitait donc celle qui s'offrait à meilleur marché et avec le plus de facilité, c'est-à-dire que l'on se souciait surtout d'identifier les plus doués. Même pour les enfants fortunés, l'enseignement secondaire jouait un rôle sélectif.

3° Les connaissances psychologiques et pédagogiques étaient rudimentaires. Les maîtres n'étaient donc pas en mesure d'appliquer des traitements fins aux élèves éprouvant des difficultés d'apprentissage.
D'ailleurs, aujourd'hui encore, on fait souvent répéter une année à l'élève qui n'a pas été capable des performances minimales exigées pour le passage. Autrement dit, au lieu d'appliquer des remèdes particuliers, on se contente de placer de nouveau l'élève dans les conditions (même professeur, même méthode) dans lesquelles l'échec s'est produit.

Chaque année scolaire étant considérée comme un filtre pour la suivante, une certaine quantité d'échecs paraissait donc normale. Bref, la répartition gaussienne semblait satisfaisante.

Mais le souci constant de la promotion des plus aptes a causé, progressivement, une déformation pédagogique plus subtile. On en est arrivé à penser que les connaissances «qui comptent vraiment», les connaissances «approfondies», le «véritable jeu des idées abstraites», ne sont assimilables que par une minorité possédant des aptitudes spécifiques à un degré élevé. Et l'on a forgé les méthodes d'enseignement et d'évaluation en conséquence.

Les élèves et leurs parents acceptent d'ailleurs cet état de choses sans grande discussion. On s'inscrit rarement en section «latin-mathématiques» sans se sentir spécialement apte dans cette direction.

On fera le rapprochement entre ces considérations et le passage suivant de Piaget[1]:

> Au nom de quel critère un enseignement élémentaire est-il jugé plus facile qu'un enseignement dans les classes primaires supérieures et celui-ci plus facile qu'un enseignement secondaire? La seule considération qui justifie cette hiérarchie est, bien entendu, celle des matières à enseigner, mais envisagées du seul point de vue du niveau des connaissances elles-mêmes, indépendamment de leur plus ou moins grande facilité d'assimilation par les élèves. Deux problèmes préalables se posent alors aussitôt. Le premier est d'établir s'il est effectivement plus aisé de faire saisir une structure élémentaire, mettons de calcul ou de langage, à un jeune enfant de 7 à 9 ans que de faire assimiler une structure plus compliquée à un adolescent? Or, rien ne prouve que la seconde structure, qui, du point de vue de la science ou de l'adulte lui-même est effectivement plus complexe, soit plus difficile à transmettre, ne serait-ce que parce que l'adolescent est précisément plus proche, quant à son organisation mentale, des habitudes de penser et de parler de l'adulte. Le second problème est de savoir si pour la suite du développement intellectuel de l'élève une bonne assimilation de la structure en jeu (par opposition à une assimilation approximative ou plus ou moins verbale) est plus importante s'il s'agit de structures d'un niveau supérieur ou d'un niveau élémentaire, celles-ci conditionnant en fait toute la vie scolaire ultérieure, tandis que celles-là peuvent donner lieu à des suppléances ou des autocorrections selon le niveau de l'élève.
>
> De ce double point de vue des difficultés d'assimilation et de l'importance extérieure des notions, il est, en fait, permis de penser, si l'on se place à un point de vue psychologique et même épistémologique plus qu'à celui du sens commun administratif, que plus l'écolier est jeune, et plus l'enseignement est difficile ainsi que gros de conséquences pour l'avenir.

Assurément, on ne peut ambitionner de faire indifféremment de chacun un virtuose de la mathématique, du piano ou de la littérature. Mais où se situe la limite? A partir de quel degré l'accès à un savoir est-il impossible à ceux qui possèdent des aptitudes estimées moyennes, voire médiocres?

1 J. PIAGET, *Psychologie et pédagogie*, Paris, Denoël, 1969, pp. 186-187.

Par nature, chacun de nous semble médiocrement doué en bien des domaines. Il ne manque, par exemple, pas d'intellectuels plus ou moins réfractaires à la mathématique. Pourtant, on observe fréquemment qu'à force de vouloir, de persévérer, de faire redire les explications, de changer de manuel ou de maître pour trouver une forme d'enseignement qui convienne, des notions de mathématiques d'abord considérées comme trop ardues sont bel et bien maîtrisées par certains.

Pour autant qu'ils y consacrent le temps nécessaire, les élèves moyens (c'est-à-dire, vu la sélection déjà opérée par les études antérieures, probablement plus de 80 % de la population scolaire d'un niveau donné) peuvent aller beaucoup plus loin qu'on ne l'imagine.

Dans cette perspective, la formule lapidaire par laquelle J. Carroll caractérise l'aptitude prend toute sa valeur : « L'aptitude est la quantité de temps demandée par l'apprenant pour dominer une matière[1]. »

L'implication de cette conception est considérable : si on leur alloue le temps nécessaire, tous les élèves se trouvant normalement dans une classe pourraient arriver à un bon, voire à un très bon résultat.

La loi de Posthumus.

Formulée dès 1947, la loi de Posthumus[2] peut s'exprimer de la façon suivante : « Un enseignant tend à ajuster le niveau de son enseignement et ses appréciations des performances des élèves de façon à conserver, d'année en année, approximativement la même distribution (gaussienne) de notes. »

Dans son remarquable ouvrage de 1947 (qui n'eut guère de retentissement parce qu'il fut écrit en néerlandais), K. Posthumus a déjà dénoncé les phénomènes qui viennent d'être décrits. Voici deux passages de son analyse remarquable.

- *« La conception selon laquelle les résultats scolaires devraient se distribuer selon une courbe de Gauss est indéfendable et repose sur la méconnaissance des lois de la probabilité.[3] »*
- *« Il est remarquable que la courbe des « progrès » reste bien symétrique (...). On ne peut trouver l'explication de ce phénomène que dans la manière dont les évaluateurs ajustent leurs exigences de façon à toujours retrouver la même distribution de notes. »*

1 J. CARROLL, A Model of School Learning, *Teachers College Record*, 1963, 64 : 723-733.
2 K. POSTHUMUS, *Levensgeheel en School*, La Haye, s. éd., 1947.
3 Posthumus fait ici allusion au fait que les populations scolaires sont de plus en plus sélectionnées ; leurs aptitudes ne se distribuent donc plus selon la loi du pur hasard.

CHAPITRE 3

LA THEORIE DE L'EVALUATION FORMATIVE

L'expression *évaluation formative* - il l'oppose à *évaluation sommative* (traitée au chapitre 4) - a été forgée par Michael Scriven[1].

L'*évaluation normative* nous est maintenant familière. Pour interpréter le score obtenu à un test classique d'inventaire de connaissances ou d'intelligence, on le situe dans une distribution statistique : la performance d'un individu est jugée par référence à celles d'autrui. De même, on classe souvent encore les élèves entre eux selon l'ordre croissant ou décroissant de leurs résultats scolaires, et c'est d'après la place ainsi occupée que bien des parents apprécient le travail de leurs enfants.

Or, dans les deux cas, le résultat est essentiellement relatif. Que le groupe de référence varie de composition et le résultat apparaît sous un autre jour.

Une simple différence d'âge peut aussi changer considérablement la face des choses. Dans bien des normes de tests de connaissances d'usage courant, un an d'âge ou une année scolaire en plus ou en moins suffisent pour qu'une même performance soit considérée comme médiocre ou bonne.

Que la connaissance soit acquise ou non n'a donc pas été la préoccupation première des constructeurs de ces tests, mais bien à quelle vitesse cette acquisition s'est réalisée. En nous reportant à la définition de J. Carroll, on évalue donc l'aptitude au lieu d'évaluer le contenu de l'apprentissage.

Les partisans de l'*évaluation formative* prennent le contre-pied de cette conception.

1 M. SCRIVEN, The Methodology of Evaluation, in R. TYLER, (Ed.), *Perspectives of Curriculum Evaluation*, Chicago Rand McNally, 1967.
Voir aussi : B.S. BLOOM, Learning for Mastery, in *Evaluation Comment*, 1968, 2. Notre exposé de la méthode de l'évaluation formative s'appuie directement sur cette publication. Plusieurs résultats d'expériences où la théorie du *mastery learning* a été appliquée sont publiés par J. BLOCK, B. BLOOM et J.T. HASTINGS (1970).

Puisque l'élève vient à l'école pour apprendre, l'important n'est-il pas de le situer dans l'ascension du savoir ?

Imaginons qu'une analyse scientifique rigoureuse nous révèle qu'avec toutes ses nuances et ses complications, l'accord du participe passé, employé avec l'auxiliaire avoir, présente cent cas différents. Pour chacun, on peut définir des critères de maîtrise.

Selon le niveau scolaire, le nombre de cas à dominer peut alors être fixé. Dans ce contexte, l'évaluation scolaire change de nature.

L'élève est noté en fonction d'un critère objectif : le chemin parcouru dans l'acquisition.

L'évaluation formative consiste fondamentalement à déterminer pour chaque unité ou tâche d'apprentissage dans quelle mesure un élève est maître de la difficulté. Il s'agit donc d'une démarche diagnostique.

L'unité est, dans ce contexte, une portion précise d'un apprentissage à effectuer ; elle est souvent définie sous forme d'objectifs, voire de micro-objectifs à atteindre. Les unités peuvent être hiérarchisées entre elles, dans la mesure où la maîtrise de l'une est nécessaire pour aborder une ou plusieurs autres.

Dans la maîtrise d'une unité donnée, on peut également distinguer une hiérarchie de processus mentaux (par exemple, selon la taxonomie de Bloom).

Quoi qu'il en soit, l'*évaluation formative* a pour but de dresser un état d'avancement, de reconnaître où et en quoi un élève éprouve une difficulté et de l'aider à la surmonter. Cette évaluation ne se traduit pas en notes, et encore moins en scores. Il s'agit d'une information en retour (feed-back) pour l'élève et pour le maître.

En raison de sa *nature diagnostique*, l'évaluation formative appelle l'action corrective, sans laquelle il n'existe d'ailleurs pas de véritable enseignement.

En outre, si l'on ambitionne de conduire tous les élèves jusqu'à un niveau de compétences minimum, sinon jusqu'à la maîtrise totale de la notion ou de la technique, la façon d'enseigner est elle-même remise en cause : il n'est plus possible d'appliquer indifféremment à tous une seule et même méthode pendant une même durée.

A ce propos, D. Bain[1] craint que l'on ne recoure à l'évaluation formative indépendamment de la pédagogie pratiquée, ce qui équivaut à dissocier enseignement - diagnostic et remédiation. Dans ce cas, le diagnostic se réduit souvent à une évaluation microsommative identifiant le lieu des erreurs commises, mais non leur cause, ni leur origine. D'où un traitement plutôt symptomatique qu'étiologique. Seule une approche didactique interactive, au cours de laquelle l'enseignement et l'évaluation se remodèlent en permanence en fonction de l'évolution des apprentissages, peut être génératrice d'une éducation de grande qualité.

Se plaçant dans une perspective pratique, L. Allal[2] reformule et développe ces idées de la façon suivante :

> « Pendant la totalité d'une période consacrée à une unité de formation, les procédures d'évaluation formative sont intégrées aux activités d'enseignement et d'apprentissage. Par l'observation des élèves en cours d'apprentissage, on cherche à identifier les difficultés dès qu'elles apparaissent, à diagnostiquer les facteurs qui sont à l'origine des difficultés de chaque élève et à formuler, en conséquence, des adaptations individualisées des activités pédagogiques. Dans cette optique, toutes les interactions de l'élève - avec le maître, avec d'autres élèves, avec un matériel pédagogique - constituent des occasions d'évaluation (ou d'auto-évaluation) qui permettent des adaptations de l'enseignement et de l'apprentissage. La régulation de ces activités est donc de nature interactive. Le but est d'offrir une « guidance » individualisée en cours d'apprentissage plutôt qu'une remédiation a posteriori. »

Et, tenant compte des conditions réelles de la conduite d'une classe, L. Allal précise :

> « Dans la réalité de la pratique pédagogique, le maître sera souvent amené à élaborer des procédures d'évaluation formative qui combinent des modalités de type « évaluation ponctuelle, régulation rétroactive »[3] avec des modalités de type « évaluation continue, régulation interactive ». Nous évoquerons, à titre d'exemple, trois cas de modalités mixtes.
>
> Cas A. Après une série de leçons ou d'autres activités où le maître n'a pas pu observer les élèves en cours d'apprentissage, il y a passation d'un contrôle écrit (test, exercice, etc.). Ayant repéré par ce contrôle les élèves qui ont des difficultés d'apprentissage, le maître poursuit avec eux un mode d'évaluation (par observation, entretien. etc.) qui permet des diagnostics et des régulations individualisés.
>
> Cas B et C. Ayant mis en place un mode d'évaluation continue et interactive, mais ne pouvant pas, pour des raisons pratiques, observer chaque élève lors de chaque activité, le maître a recours, périodiquement, à des moyens de contrôle

1 D. BAIN, Pour une formation à l'évaluation formative intégrée à la didactique. *In* M.G. THURLER et P. PERRENOUD, éds., *Savoir évaluer pour mieux enseigner,* Genève, *Cahiers du Service de la Recherche Sociologique,* 1988, 26, 21-37.

2 L. ALLAL, J. CARDINET, P. PERRENOUD, *L'évaluation formative dans un enseignement différencié*, Berne, Lang, 1979, pp. 142-143.

3 Par exemple, effectuée à l'aide d'un test diagnostique passé en *fin* d'apprentissage.

écrit qui permettent d'identifier des difficultés qui n'ont pas été repérées en cours de route. Ce repérage est suivi, selon les circonstances, soit par des activités de remédiation partiellement standardisées (cas B), soit par des régulations interactives et individualisées (cas C).»

Que deviendrait l'enseignement dans cette perspective?

B.S. Bloom propose une réponse, appuyée sur une multitude de données expérimentales dans son ouvrage *Caractéristiques individuelles et apprentissages scolaires*[1]. Nous y renvoyons le lecteur désireux d'approfondir la question.

Trois problèmes cruciaux se posent:

1. Comment jalonner l'ascension du savoir?
2. Comment conduire l'élève?
3. Où se situe la limite pratique de cette pédagogie de la courbe en *j*?

A ma connaissance, il n'existe pas encore de réponses complètes à ces questions. Pareille imprécision du savoir semble normale en sciences naturelles ou en médecine. Nous devons apprendre à l'accepter aussi dans les sciences de l'éducation. Notre discussion aboutit donc maintes fois sur des recherches à entreprendre ou à continuer.

I. **Jalonner l'ascension du savoir.**

En voyage, pour déterminer à quelle distance on se trouve du but, deux conditions doivent être remplies: d'une part, savoir où l'on est et où l'on va, et, d'autre part, disposer d'une carte indiquant clairement le chemin.

De même, en éducation, nous devons définir les objectifs à atteindre et déterminer avec précision les apprentissages particuliers qui y conduiront. Le problème varie selon que l'on a affaire ou non à des apprentissages de base. Pour ces derniers, il faut découvrir l'enchaînement «critique» des matières, c'est-à-dire celui où l'une n'est accessible que si la précédente est assimilée. Pour cette raison, l'apprentissage des connaissances et des techniques de base doit, en dernière analyse, souvent être linéaire, tandis que les acquisitions et les applications qui vont au-delà voient s'ouvrir devant elles des voies de plus en plus nombreuses.

Par exemple, quelle que soit la méthode d'enseignement, il n'est pas possible d'appliquer complètement une règle de trois, sans avoir - notamment - la notion de la multiplication et de la division. A un niveau plus élevé, comment faire du calcul intégral sans savoir ce

1 Bruxelles, Labor; Paris, Nathan (collection «Education 2000»), 1979.

qu'est une fonction? Mais pareilles propositions sont encore trop vagues. Quels sont exactement les apprentissages nécessaires et suffisants pour pouvoir assimiler la règle de trois? Et, parmi eux, lesquels sont critiques par rapport aux autres?

La *définition des unités d'apprentissage* pose un problème encore loin d'être résolu. Aux extrêmes, deux écoles s'affrontent: l'une est purement empirique, l'autre expérimentale.

Les empiristes se réfèrent à leur expérience professionnelle, à leur logique, à leur intuition aussi, pour diviser une matière en parties relativement homogènes et pour les ordonner. Ainsi procèdent les auteurs de manuels lorsqu'ils répartissent une matière en sections, chapitres, rubriques, paragraphes et alinéas.

Nous avons vu, par ailleurs, que le groupe de Popham a adopté une démarche empirique elle aussi, mais plus serrée: elle arrive au niveau des objectifs spécifiques.

A l'opposé, les tenants de l'analyse hiérarchique des contenus, principalement R. Gagné et J.B. Carroll, recherchent des règles psychologiques ou psychométriques permettant de reconnaître la structure d'une matière et les passages d'apprentissage obligés.

On observe toutefois que rares sont les voies uniques dans les apprentissages. Pour la pédagogie corrective, une connaissance claire des composantes, des unités, importe le plus, chaque individu devant, à la limite, jouir d'une liberté totale pour les structurer, les répartir, les articuler, et arriver à une plus grande maîtrise fonctionnelle du milieu.

Plus simplement, nous dirons que peu importent les voies choisies pour apprendre ou faire apprendre une même matière, du moment que la psychologie de l'enfant soit respectée, et qu'au moment voulu, le maître sache sur quels points il doit faire porter son contrôle pour s'assurer de l'apprentissage effectif.

Il va de soi que certaines branches se prêtent beaucoup mieux que d'autres à une structure d'unités hiérarchisées. Pour une large part, les mathématiques y sont spécialement favorables. A l'opposé, l'histoire universelle, l'enseignement de la composition française, ne sont guère sous-tendus de structures hiérarchiques logiques ou psychologiques aisément discernables. Les faits y abondent, mais les règles sont rares. Quelles sont, par exemple, celles d'une rédaction parfaite?

Par conséquent, les apprentissages dans ces branches, structurellement floues, ne sont pas cumulatifs mais additifs. Aussi, déjà à l'in-

térieur d'une unité, les composantes n'entretiennent pas toujours des relations étroites et ne s'ordonnent pas nécessairement selon un modèle bien articulé. Dans beaucoup de cas, la connaissance se construit comme un puzzle. On sait que pour l'histoire, on essaie de tourner cette difficulté en respectant l'ordre chronologique ou en travaillant par thèmes, mais ce ne sont qu'artifices méthodologiques.

Faut-il en conclure que les unités et donc l'évaluation formative n'ont pas leur place dans ces branches ? Assurément pas, sinon l'enseignement s'abandonnerait à une improvisation permanente dont on imagine les méfaits et les lacunes.

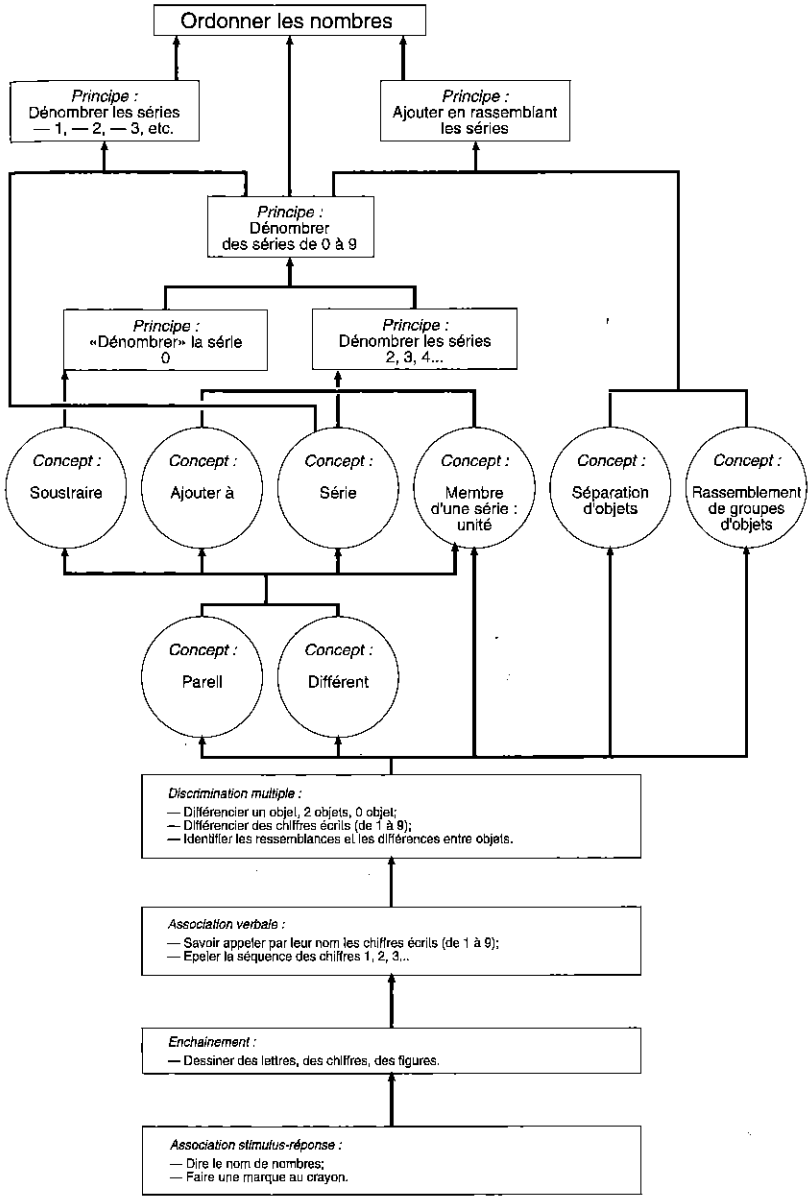
En dehors des impératifs psychologiques et téléologiques, l'éducateur reste libre d'ordonner les apprentissages à réaliser *comme s'ils pouvaient se hiérarchiser,* l'ordre pouvant être remis en cause selon les nécessités pédagogiques.

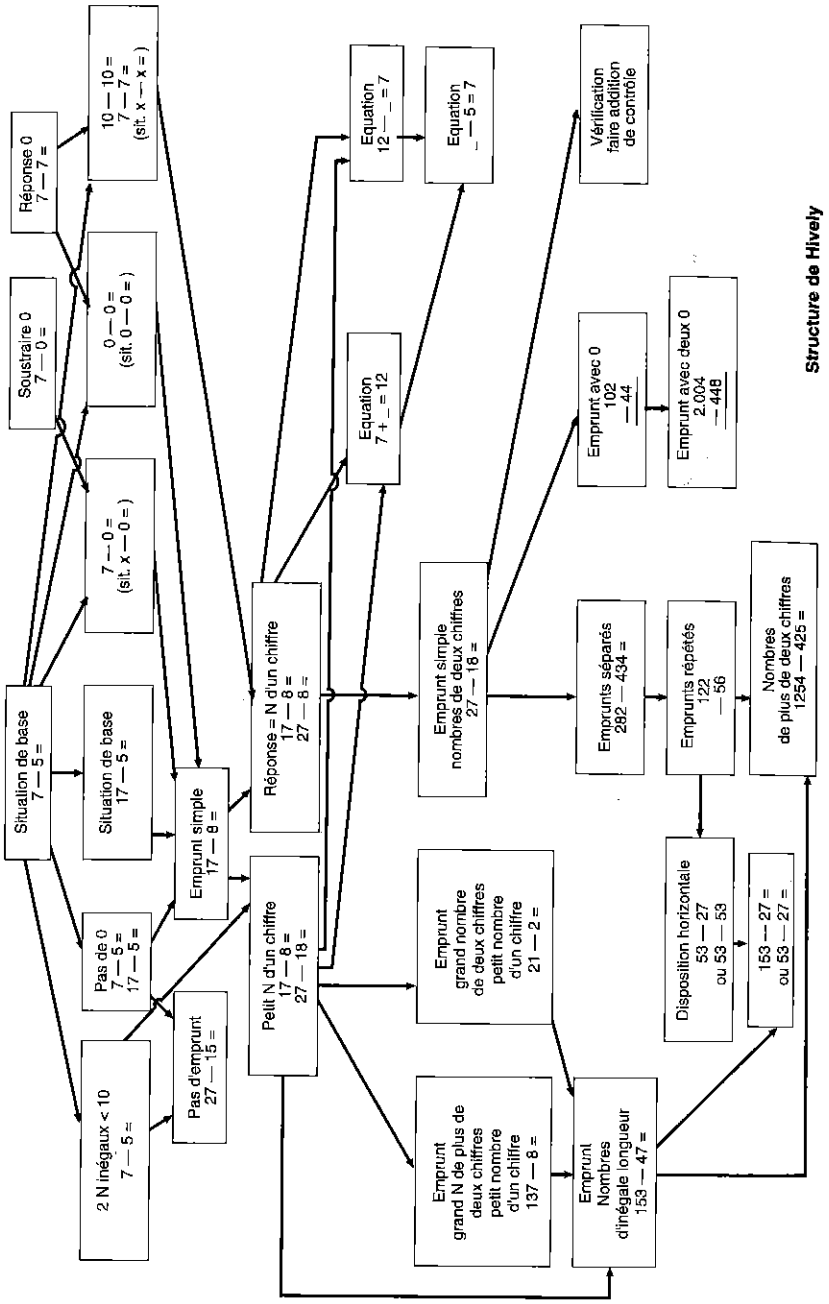
Ainsi, on dispose au moins d'une ligne de conduite et d'une possibilité de contrôle de maîtrise. Les objectifs axés sur les principes, les structures et les transferts se verront naturellement accorder la priorité. Les autres, même s'ils sont de nature plus pointilliste, méritent néanmoins d'être rigoureusement inventoriés, ne fût-ce que pour éviter de stériles accumulations factuelles.

Voici deux exemples de structure hiérarchique d'apprentissage des opérations numériques[1]. L'une, proposée par Gagné, est psychologique; l'autre, proposée par Hively, est logique.

1 Nous empruntons l'adaptation française de cet exemple à D. BONOR, Les buts de l'éducation, in M. REUCHLIN, *Traité de psychologie appliquée,* 6, o.c., p. 15.

Structure de Gagné





II. Guider l'élève.

Une fois clairement défini le chemin qui conduit à un apprentissage, il faut y engager l'élève. Pour les acquisitions fondamentales, nous avons, en outre, décidé que *tous* devraient, en principe, arriver au but.

Dans ce cas, l'évaluation fréquente des progrès est essentielle. D'où la nécessité d'*exercices de maîtrise* et de *tests diagnostiques* portant sur des matières très limitées et utilisés par les maîtres eux-mêmes.

Ces instruments - que seule la collaboration des enseignants et des chercheurs permettra de construire en nombre suffisant - indiquent où l'élève en est (ce qui lui donne sa note) et *où il éprouve des difficultés.*

Les *remèdes* sont multiples et doivent entrer dans notre pratique pédagogique quotidienne:

a) Indiquer de façon précise la partie du cours à réétudier.

b) Le travail par sous-groupes: si un groupe de plus de trois élèves rencontre la ou les mêmes difficultés, le professeur a avantage à consacrer une partie du temps de la leçon à ce groupe, les autres travaillant indépendamment.

c) B. Bloom recommande des groupes hétérogènes de trois élèves. Comme les élèves ne sont plus en concurrence, mais luttent pour la maîtrise d'une matière, l'entraide joue.

d) Dans les écoles comptant plusieurs professeurs d'une même branche, un élève devrait avoir la faculté de demander une explication, voire une courte leçon particulière, à un autre professeur que le sien.

e) Varier la présentation: mettre à la disposition des élèves des manuels expliquant la notion de différentes façons; permettre le recours à l'enseignement programmé, l'apprentissage assisté par ordinateur, aux moyens audiovisuels; varier le niveau d'abstraction.

f) En général, ne pas fixer à tous le même temps, pour les mêmes choses.

Pareille pratique paie. B. Bloom rapporte une expérience conduite dans cette ligne où plus de 80 % des élèves ont obtenu la meilleure note en fin d'année. C'est le triomphe de la pédagogie de la courbe en *j.*

Peut-il toujours en être ainsi? Rien ne permet de l'affirmer.

III. Le rapport temps-apprentissage.

A mesure que la recherche avance, la relation entre le facteur temps et les apprentissages se précise. Une vue d'ensemble de cette question est donnée par B. Bloom[1]. Il envisage successivement l'évolution des apprentissages à mesure que les années civiles passent, en fonction du temps écoulé pour maîtriser un apprentissage, et en fonction du temps pendant lequel l'élève s'adonne à l'apprentissage.

A mesure que les années s'écoulent, des progrès importants interviennent. A l'école primaire, un niveau de performance atteint par les 20 % supérieurs des élèves, une année, est souvent atteint par 50 % des élèves, l'année d'études suivante, et par 80 % des élèves, deux ans après.

Si l'on considère plus spécialement le temps qui s'écoule avant que différents élèves ne maîtrisent un apprentissage particulier, on constate que, en une grossière approximation, les élèves lents y consacrent quatre à cinq fois plus de temps que les élèves rapides. Ces rapports se vérifient assez bien si l'on utilise les scores obtenus aux tests d'aptitudes et aux tests d'intelligence générale pour prédire le temps nécessaire à un apprentissage (corrélation de + .50 à + .70).

Dès 1947, K. Posthumus[2] attire l'attention sur ce phénomène.

«La même performance coûte à tel élève plusieurs fois autant de temps qu'à tel autre.»

Le tableau suivant indique le nombre de minutes nécessaires à huit élèves d'une même classe pour faire leurs devoirs à domicile dans cinq branches A, B, C, D et E.

Elèves	Branches					Total
	A	B	C	D	E	
1	10	25	20	10	15	80
2	20	30	15	10	20	95
3	25	35	45	10	50	165
4	30	60	30	15	30	165
5	25	65	30	20	40	180
6	40	90	30	10	25	195
7	20	90	30	15	45	200
8	30	100	15	5	60	210
Moyennes	25	62	27	12	36	161

L'exactitude de ces données n'est pas sûre, car elles ont été fournies par les parents et les élèves. Elles sont cependant cohérentes avec les résultats de nombreuses recherches. Les élèves les plus lents consacrent à une branche trois à quatre fois plus de temps que leurs compagnons de classe les plus rapides.

1 B.S. BLOOM, *Time and Learning*, Communication au 81e Congrès annuel de l'American Psychological Association, 1973.

2 K. POSTHUMUS, *Levensgeheel en School*, La Haye, s. éd., 1947.

Toutefois, ces données concernent l'enseignement traditionnel. Que se passe-t-il si l'on applique une méthodologie axée sur la maîtrise des apprentissages ?

Lors de l'apprentissage de la première unité, le rapport de $^1/_4$ ou $^1/_5$ se vérifie généralement. Ici, la différence de temps représente, rappelons-le, l'aide supplémentaire apportée directement ou indirectement pour combler les lacunes révélées par l'évaluation formative.

Si cette première unité est indispensable pour l'apprentissage de la deuxième et si l'étude de celle-ci est abordée dès que la première est acquise, le rapport temporel de $^1/_5$ tend à se réduire, pour se situer vers $^1/_3$ après une série d'unités. Arrivé à ce point, B. Bloom fait une observation capitale: «... dans ce processus d'apprentissage axé sur la maîtrise, la valeur prédictive des tests d'intelligence générale ou d'aptitudes, en ce qui concerne le temps à consacrer aux apprentissages, diminue progressivement pour atteindre finalement une valeur très basse. Par contre, le score à un test formatif est bon prédicteur du temps nécessaire à l'apprentissage de l'unité suivante dans la séquence.»

Autrement dit, le taux de performance des élèves s'améliore à mesure qu'ils progressent dans le programme; l'aide particulière à leur apporter pour qu'ils maîtrisent les unités d'apprentissage diminue parallèlement.

Une troisième façon de voir les choses est de ne considérer que le temps pendant lequel l'élève s'adonne pleinement à un apprentissage donné, y consacre son attention et son énergie (*temps effectivement investi dans la tâche*). Il est clair qu'une heure de cours individualisé est différemment utilisée par les élèves: les uns se mettent immédiatement au travail et ne le quittent pas; d'autres ne commencent pas tout de suite et se lassent vite, etc.

D'où l'utilité de faire la distinction entre le temps écoulé globalement et le temps concentré sur la tâche.

Il importe naturellement de tenir compte des comportements observables et des comportements non observables (réflexion sur le problème, activité métacognitive), évalués par introspection. Comme on pouvait s'y attendre, le temps concentré sur l'apprentissage est bon prédicteur de la performance (après correction pour la fidélité, les corrélations rendent compte d'environ trois cinquièmes de la variation des performances entre étudiants)[1].

1 B. Bloom remarque avec raison que le temps pendant lequel les élèves s'adonnent aux apprentissages peut aussi être pris comme indice de la qualité de l'enseignement.

Toutefois, une pédagogie de la courbe en J exerce ici aussi une influence spectaculaire. Considérons deux groupes équivalents d'élèves; l'un reçoit l'enseignement habituel, l'autre passe au système axé sur la maîtrise des objectifs.

Lors de l'apprentissage de la première unité de maîtrise, le *temps de concentration* des deux groupes ne diffère guère et s'élève, en moyenne, à 65% du temps écoulé. A mesure que les unités se succèdent, on observe que les élèves engagés dans le processus de maîtrise se concentrent de plus en plus (jusqu'à 85% du temps écoulé), tandis que l'effort des autres se relâche (jusqu'à 50% du temps écoulé). Autrement dit, le groupe axé sur la maîtrise apprend plus efficacement à apprendre.

Observation capitale: à mesure que l'on avance dans l'apprentissage axé sur la maîtrise, les différences entre étudiants se réduisent. Au début, les variations dans le temps concentré sur la tâche à maîtriser sont de l'ordre de 1 à 3; à mesure que l'on avance, il tend à se réduire jusqu'à 1 à 1,5, voire moins. Il s'agit donc d'un nivellement par le haut dans l'efficacité des apprentissages.

Ces observations ouvrent des perspectives pédagogiques considérables et apportent un grand espoir à beaucoup d'étudiants qu'un milieu d'origine ou une scolarité antérieure, peu favorables aux nouveaux apprentissages à réaliser, vouaient trop souvent à l'échec.

Le nivellement par le haut a néanmoins des limites. On aura déjà constaté que, dans aucun des chiffres avancés plus haut, on ne prétend égaliser entièrement. Il va aussi de soi que, toujours, on s'est adressé à des populations scolaires normales, offrant certes une large marge de variation d'aptitudes au départ, mais excluant les handicaps mentaux graves. On ne peut évidemment permettre de passer vingt ans au lieu de six dans l'enseignement secondaire!

Cette constatation ne doit néanmoins pas remettre le système en cause. Le tout est d'orienter progressivement les élèves en fonction de leurs aptitudes réelles. Le plafonnement dans un enseignement axé sur la maîtrise des apprentissages serait sans doute un bon indicateur de la nécessité de réorienter ou, au moins, de procéder à un examen approfondi de la situation de l'élève. Les cycles d'observation et d'orientation, introduits progressivement dans l'enseignement secondaire de la plupart des pays, offrent cette possibilité.

IV. **Le système d'enseignement mis en cause.**

La pédagogie de la courbe en J appelle le nivellement par le haut; il serait néanmoins chimérique d'imaginer qu'elle permet le nivellement par le... génie.

Nous l'avons vu, moins de cinq pour cent de la population possèdent des aptitudes *exceptionnelles*. Elles doivent être développées au maximum, tant par respect pour la personne que dans l'intérêt de la communauté.

Un système d'enseignement souple permet à la fois d'instruire chacun à l'allure convenable et de laisser s'épanouir les surdoués. A la classe rigide se substituent des activités multiformes. Tantôt le travail est individualisé, en particulier grâce à la technologie éducationnelle; tantôt on travaille par groupes homogénéisés selon les aptitudes spécifiques pour une seule branche[1]; à d'autres moments, des groupes de grandeur variable se constituent selon des critères affectifs.

Dans les conditions actuelles, la classe est un carcan. Pourquoi un élève de six ans qui sait lire en entrant en première primaire ne pourrait-il pas participer aux exercices de lecture de la deuxième année? Pourquoi, s'il le peut, tel élève de quatrième de l'enseignement secondaire ne pourrait-il pas s'il le peut, suivre un cours de mathématique dans une année supérieure?

Dans nos vieilles écoles primaires de campagne, où un seul instituteur se voyait parfois confier les six années, il était commun de voir de tels déplacements. Dans des écoles secondaires qui, actuellement, comptent plusieurs classes de même âge, il est possible de travailler à quatre ou cinq niveaux d'aptitudes différents, au moins dans quelques branches principales. Pareil système fonctionne parfaitement, depuis de nombreuses années, dans des écoles comme la New Trier High School[2].

Dès 1923, O. Decroly et R. Buyse préconisent ce système de «classes mobiles, organisées en tenant compte des différences, d'acquis dans certaines branches»[3].

1 Système à ne pas confondre avec le système qui consiste à constituer des classes homogènes dites fortes, moyennes ou faibles à l'aide de tests d'aptitudes générales ou, pire encore, en fonction des résultats scolaires globaux. On commet ainsi une double erreur scientifique: dans la grande majorité des cas, un même sujet est inégalement doué pour les diverses branches du programme; de plus, homogénéiser globalement fait baisser le rendement de l'ensemble. Dans ce système, on aboutit d'ailleurs souvent à une semi-ségrégation sociale.

2 Voir annexe 3.

3 O. DECROLY et R. BUYSE, *Les applications américaines de la psychologie à l'organisation humaine et à l'éducation*, Bruxelles, Lamertin, 1923, p. 45.

Toutefois, le recours aux groupes homogénéisés n'est réellement profitable que si les enseignants varient leur stratégie pédagogique en fonction de la qualité des groupes, et ne se contentent pas simplement de modifier leur niveau d'exigences sans rien changer d'autre.

Prévenons immédiatement l'objection financière. Une école souple ne coûte pas nécessairement plus cher qu'une autre, au contraire. Si l'on ajoute le gain de temps, de talent et la raréfaction des échecs que le système permet, on peut affirmer sans crainte de se tromper que le bénéfice est considérable...

Que des apprentissages scolaires puissent être maîtrisés par le plus grand nombre des élèves, sinon par tous, ne doit cependant pas faire croire à la disparition des différences entre les individus.

CHAPITRE 4

L'EVALUATION SOMMATIVE

Alors qu'une évaluation formative est normalement effectuée au terme de chaque tâche d'apprentissage, notamment pour intervenir immédiatement là où une difficulté se manifeste, l'évaluation sommative revêt le caractère d'un bilan. Elle intervient donc après un ensemble de tâches d'apprentissage constituant un tout, correspondant, par exemple, à un chapitre de cours, à l'ensemble du cours d'un trimestre, etc. Les examens périodiques, les interrogations d'ensemble sont donc des évaluations sommatives.

L'évaluation formative revêt, en principe, un caractère privé (sorte de dialogue particulier entre l'éducateur et son élève), tandis que 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...[1]

L'évaluation sommative ne doit cependant pas rouvrir la porte aux examens traditionnels ou, plus exactement, au mauvais usage que l'on a fait de leurs résultats. Elle aussi doit s'insérer dans l'action éducative globale et, avant tout, aider au développement de l'élève.

En matière d'évaluation de curriculums, l'évaluation sommative a pour objet de déterminer dans quelle mesure un programme fonctionne bien dans son ensemble.

Bormuth distingue :

- l'évaluation sommative du programme comme tel (*time trial study*) où l'on mesure, soit le temps mis par les étudiants pour arriver à un niveau de maîtrise des contenus pris pour critère, soit le niveau de maîtrise atteint en un temps préalablement fixé; on calcule le rapport entre le nombre d'unités de contenu apprises et le temps mis pour réaliser cet apprentissage;

1 Cf. B. BLOOM, J.P. HASTINGS et G.F. MADAUS, *Formative and Summative Evaluation of Student Learning*, New York, McGraw-Hill, 1973.

- l'évaluation par comparaison à d'autres programmes.

Il semble évident que, pour tester les effets d'un programme, les questions ou les tâches à effectuer doivent s'inscrire dans ce programme. Interroger sur ce qui n'a pas été enseigné semble ici injustifiable.

Or, l'évaluation des effets de deux curriculums différents, à l'aide d'un même instrument, soulève presque toujours cette difficulté. L'exemple le plus frappant de ces derniers temps se trouve dans la comparaison des rendements des écoles secondaires sélectives et des écoles uniques («compréhensives»). T. Husen écrit excellemment: «Les critères d'évaluation des systèmes scolaires élitistes ou unifiés dépendent de jugements de valeur: les systèmes à différenciation hâtive prennent souvent pour critère de succès un produit final particulier (par exemple les connaissances en mathématiques), tandis que les systèmes unifiés insistent sur l'aspect affectif de l'éducation, l'ouverture d'esprit, la disponibilité intellectuelle et sociale. Comment démontrer qu'un choix vaut mieux que l'autre?»

En pareils cas, la comparaison devrait s'opérer en deux temps. Une première série de mesures tenterait d'épuiser comparativement la partie commune des deux programmes; une seconde série porterait distinctement et explicitement sur les objectifs spécifiques différant d'un programme à l'autre.

Pour la première série, portant sur la partie commune, ou bien un même test s'impose, ou bien on appliquera deux instruments dont on puisse démontrer qu'ils sont chacun la *somme d'items* générés à partir des mêmes règles, à de mêmes niveaux taxonomiques, et qu'ils résultent tous de tirages au hasard dans l'univers des questions. Nous retrouvons ainsi la théorie de Bormuth, déjà rencontrée.

Enfin, signalons que le principe de l'évaluation «indépendante des buts assignés» ne manque pas de défenseurs.

En ne faisant porter l'évaluation que sur les buts assignés à un programme éducatif ou à une expérience, on risque d'ignorer des effets obliques (*side effects*) ou des effets seconds positifs ou négatifs. Il se peut que des effets positifs imprévus soient assez importants pour justifier la poursuite d'un programme qui n'atteint cependant pas les résultats escomptés. A l'opposé, les effets négatifs peuvent revêtir une telle gravité que l'arrêt du programme, autrement réussi, s'impose.

C'est pourquoi M. Scriven propose qu'en certains cas, les évaluateurs analysent les situations en toute indépendance et choisissent de mesurer les effets qu'ils croient observer plutôt que de se laisser guider par les objectifs explicites du programme (*goal-based evaluation*).

CONCLUSIONS

Les conclusions partielles se sont imposées à mesure que nous avancions dans cette étude déjà longue. Le sujet est pourtant loin d'être épuisé et les solutions proposées ne sont certainement pas définitives.

Sortir des routines et ambitionner de traduire les grandes théories pédagogiques dans la pratique quotidienne de l'éducation, découvre tant d'écueils, tant de conditions à remplir que l'on risque d'être envahi par le doute et le découragement.

Au risque de laisser nos lecteurs insatisfaits, nous n'avons pourtant pas voulu voiler les difficultés, et moins encore laisser croire à l'existence d'une docimologie achevée, capable de conduire à une évaluation parfaite, si on lui en donne les moyens. Vraisemblablement, et heureusement sans doute, la mesure rigoureuse des traits humains les plus fins restera toujours impossible : le sort nous garde de la machine à peser les âmes !

Que l'évaluation continue, formative s'insère fonctionnellement dans le processus d'enseignement et d'apprentissage dont elle devrait être indissociable, n'en supprime pas, pour autant, la nécessité d'une technique d'examen et de notation.

Par ailleurs, l'exigence d'évaluations normatives, que l'éducateur peut regretter pour des raisons idéales, dépasse le domaine scolaire : elle correspond à un caractère fondamental de notre civilisation.

Peut-être en sera-t-il un jour autrement. Si notre culture continue à s'intellectualiser, et donc à croître en complexité, ne voit toutefois pas comment elle pourrait renoncer complètement aux examens et aux concours.

La position docimologique actuelle est résolument éclectique dans son désir de concilier les avantages certains de théories et de techniques nouvelles et l'apport indéniable et fondamental de l'analyse qualitative.

Pour nous limiter à un seul exemple de cet éclectisme, nous ne pouvons concevoir que les décisions de passage intervenant en fin de chaque année d'études universitaires soient prises seulement en fonction des résultats obtenus à un test à choix multiple se prêtant à la notation automatique. Assurément, l'introduction de ce type d'épreuves est hautement souhaitable pour le contrôle objectif et approfondi des connaissances, mais il ne suffit pas. L'évaluation du travail de l'année doit intervenir et la rencontre ultime du maître et de son élève apporter toutes les nuances que l'approche quantitative a laissées dans l'ombre.

De même que l'introduction des machines dans l'industrie a permis à l'intelligence de prendre le pas sur la force musculaire et la routine avilissante, de même le contrôle automatique de la connaissance des faits, des méthodes et des techniques rend possible un examen final réellement centré sur les processus mentaux supérieurs et sur la personnalité.

Dans le présent ouvrage, seule l'évaluation des élèves ou, plus exactement, de ce qu'ils réalisent dans le domaine cognitif au cours de leurs études a été envisagée. Toutefois, il ne faut jamais perdre de vue qu'ils sont des personnes entières traversant un moment de leur vie, crucial pour leur développement physique, cognitif, affectif et social. Même si elle y prend beaucoup de place, l'école n'est pas toute leur vie. Loin s'en faut !

Les professeurs, les établissements, le système scolaire doivent aussi être évalués. Chacun de ces autres aspects nécessiterait un précis au moins aussi complexe que celui-ci. On en trouvera une ébauche dans V. DE LANDSHEERE, *L'éducation et la formation. Science et pratique*, Paris, P.U.F., Collection 1er Cycle, 1992.

ANNEXES

ANNEXE I

ETUDE COMPAREE D'UNE QUESTION D'EXAMEN PRESENTEE SELON LA METHODE TRADITIONNELLE ET SELON LA METHODE PAR QUESTIONS A CHOIX MULTIPLE

Exposé de la question[1] (Méthode traditionnelle).

1° a) Donner la liste des causes d'hypoglycémie.
b) Décrire les lésions anatomo-pathologiques résultant de l'hypoglycémie.

Pour permettre aux correcteurs un certain degré d'uniformité dans l'appréciation, la grille suivante a été établie par l'ensemble des correcteurs:

A. Valeur égale pour les deux parties de la question;

B. Pour obtenir une note de 75/100, le candidat doit avoir indiqué dans la liste des causes:
1) excès d'insuline;
2) tumeur fonctionnelle des îlots de Langerhans;
3) influence endocrinienne par hypofonctionnement de l'hypophyse et des surrénales;
4) maladie du foie empêchant la mise en réserve du glycogène (nécrose aiguë) ou la libération du glycogène (maladie de Von Gierke).

2° A quoi reconnaît-on la différence des lésions causées par une seule crise aiguë d'hypoglycémie de celles causées par une hypoglycémie chronique?

3° Décrire les modifications visibles au niveau du cerveau.

1 MOORE, Robert A., *Methods of Examining Students in Medecine*, in « Journal of Medical Education », janvier 1954, vol. 29, n° 1.
Traduction empruntée au rapport français sur la réforme des examens de médecine, o.c.

C. Pour obtenir une note allant de 75 à 90/100, le candidat doit avoir fourni:
- soit une liste de causes montrant qu'il a compris le mécanisme d'action de chacune des causes,
- soit une liste comprenant d'autres causes, mais en indiquant qu'il comprend que les mécanismes d'homéostase ont habituellement tendance à corriger l'hypoglycémie, dans les cas tels que:
1) ingesta diminués;
2) absorption perturbée;
3) métabolisme augmenté comme dans l'hyperthyroïdisme;
4) infection grave;
5) surmenage physique;
6) glycosurie rénale.

D. Pour obtenir une note supérieure à 90/100, le candidat doit avoir indiqué:
- une liste logique des quatre causes majeures et des six causes mineures, en plus d'un type idiopathique,
- une différenciation de l'hypoglycémie aiguë et chronique par une description détaillée des lésions cérébrales.

Exposé de la question (Méthode par questions à choix multiple).

Voici comment une même question pourrait être présentée selon la nouvelle méthode.

1° On peut déterminer si le candidat est capable de reconnaître les quatre causes les plus importantes et les différencier des moins importantes dans la question suivante:

INSTRUCTIONS. Quatre des cinq phrases numérotées sont communes à l'un des trois troubles fonctionnels de la liste alphabétique (A.B.C.).

Indiquez celle qui est l'exception et le trouble fonctionnel commun aux quatre phrases restantes.

1. Excès d'insuline
2. Tumeur fonctionnelle des îlots de Langerhans
3. Glycosurie rénale
4. Hypofonctionnement hypophysaire
5. Maladie de Von Gierke

A. Hypoglycémie cliniquement décelable
B. Hyperglycémie cliniquement décelable
C. Glycosurie cliniquement décelable

Si le candidat répond correctement, il montre qu'il sait que 1, 2, 4 et 5 peuvent produire une hypoglycémie cliniquement décelable, que ce n'est pas le cas de 3, et qu'aucune combinaison des quatre cas parmi les cinq ne peut être associée ni à l'hyperglycémie, ni à la glycosurie. En d'autres termes, la possession de connaissances positives et négatives est prouvée.

2° Si l'on veut savoir ce que sait le candidat sur les phénomènes qui commandent l'équilibre du niveau du sucre sanguin et de l'insuline, on peut poser la question suivante:

INSTRUCTIONS. Chacune des phrases suivantes est composée de deux parties: l'exposé d'un fait et la raison de ce fait.

Indiquez pour chacune des phrases numérotées la lettre A, B, C, D ou E, selon que:

A. L'exposé du fait et sa raison sont vrais et ont une relation de cause à effet.

B. L'exposé du fait et sa raison sont vrais, mais n'ont pas de relation de cause à effet.

C. L'exposé du fait est vrai, mais la raison est fausse.

D. L'exposé du fait est faux, mais la raison est un fait ou un principe accepté.

E. L'exposé du fait et la raison sont faux.

(A) 1. Le taux du sucre sanguin tombe brutalement après hépatectomie parce que le glycogène contenu dans le foie est la source principale du sucre sanguin.

(E) 2. Le diagnostic anatomo-pathologique d'un adénome des îlots de Langerhans du pancréas implique que la maladie était hyperinsulinique parce que tous les adénomes des îlots sont fonctionnels et secrètent de l'insuline.

(C) 3. Les malades atteints d'hyperthyroïdisme ont toutes chances d'avoir une hypoglycémie parce qu'il existe un hyperinsulinisme associé.

(C) 4. Les malades atteints de la maladie de Von Gierke présentent une diminution du taux du sucre sanguin parce que, dans cette maladie, le glycogène n'est pas mis en réserve par le foie.

(E) 5. Une hypoglycémie durant depuis plusieurs mois n'est pas suivie de séquelles parce que les modifications cellulaires produites par l'hypoglycémie sont réversibles.

Si le candidat répond correctement à cette série, cela montre qu'il sait:

a) que le glycogène hépatique est la source principale permettant le maintien du taux du sucre sanguin;

b) que toutes les tumeurs des îlots de Langerhans ne sont pas fonctionnelles;

c) que les malades atteints d'hyperthyroïdisme ont une hypoglycémie, mais que la cause n'est pas un hyperinsulinisme associé;

d) que les malades atteints de la maladie de Von Gierke ont une hypoglycémie, mais qu'elle n'est pas due au fait qu'il n'y a pas de glycogène dans le foie;

e) qu'il y a des séquelles après hypoglycémie chronique et que les modifications cellulaires résultantes ne sont pas réversibles.

3° On peut déterminer si le candidat a quelques notions simples concernant les modifications au niveau du cerveau dans la question suivante:

INSTRUCTIONS. Chacun des exposés incomplets (numérotés) est suivi de cinq compléments au choix. Indiquez, dans chaque cas, le complément qui convient le mieux.

1. Les modifications anatomiques secondaires à une hypoglycémie chronique sont mises en évidence le plus souvent dans:
   a) la rate
   b) le rein
   c) les surrénales
   d) le cerveau
   e) la thyroïde

2. Parmi les modifications provoquées par l'hypoglycémie chronique au niveau du cerveau, les plus importantes sont sur:
   a) les neurones
   b) les astrocytes
   c) les cellules de l'épendyme
   d) les cellules oligogendrogliques
   e) les cellules microgliques

3. Parmi les altérations des cellules nerveuses provoquées par l'hypoglycémie aiguë, la plus évidente est:
   a) le déplacement du noyau
   b) la disparition de la paroi cellulaire
   c) l'altération de la substance de Nissl
   d) l'éclatement du noyau de la cellule
   e) la fusion de mitochondries.

4. Parmi les modifications provoquées au niveau du cerveau par l'hypoglycémie chronique, la plus significative est:
   a) l'hydrocéphalie interne
   b) l'épaississement fibreux de l'arachnoïde
   c) la destruction des neurones
   d) la prolifération des astrocytes
   e) la prolifération des cellules de l'épendyme

Si le candidat répond correctement à cette série, il montre qu'il sait que les altérations principales de l'hypoglycémie chronique atteignent les neurones du cerveau, que l'hypoglycémie aiguë provoque une altération de la substance de Nissl des neurones et que l'hypoglycémie chronique provoque la destruction de cellules nerveuses.

4° On peut déterminer si le candidat comprend le mécanisme des troubles, dans la question suivante:

INSTRUCTIONS. Dans la liste alphabétique sont indiqués cinq mécanismes différents pouvant conduire à l'hypoglycémie. Inscrire la lettre appropriée après chacun des exposés numérotés en associant au trouble le mécanisme responsable:

A. Augmentation de l'activité métabolique.
B. Hyperinsulinisme.
C. Mise en réserve d'un glycogène anormal dans le foie.
D. Absence de mise en réserve de glycogène dans le foie.
E. Hypofonctionnement de l'hypophyse ou des surrénales.

(B) 1. Adénomes des îlots de Langerhans.
(A) 2. Exercice physique violent.
(A) 3. Hyperthyroïdisme.
(E) 4. Maladie de Simmonds.
(C) 5. Maladie de Von Gierke.
(D) 6. Hépatite épidémique.
(C) 7. Maladie d'Addison.

Si le candidat répond correctement à cette série, cela montre qu'il comprend les bases du métabolisme des glucides et connaît les facteurs influençant ce métabolisme.

Récapitulation.

Ainsi, au moyen de ces dix-sept «questions objectives», nous avons mis en évidence toutes les connaissances requises pour pouvoir donner une note supérieure à 90/100, à savoir:

1. Liste des quatre causes principales.
2. Reconnaître la différence entre les effets de l'hypoglycémie chronique et aiguë.
3. Décrire les altérations au niveau du cerveau.
4. Compréhension des mécanismes.
5. Liste des causes mineures.
6. Différencier les lésions de l'hypoglycémie aiguë et chronique.

ANNEXE II

EXEMPLE DE QUESTIONS POUR UNE COMPOSITION EN LANGUE MATERNELLE[1]

Questions.

EPREUVE I. Deux questions - 1 ½ h.

1. Choisissez un des sujets suivants. Consacrez-y environ 1 h.
   a. Une nuit brumeuse.
   b. Un marchand achète et revend, le même jour, un objet d'occasion. Décrivez les deux scènes.
   c. Quelle serait votre politique si vous dirigiez les programmes de radio ou de télévision ?
   d. Une grande foule se disperse. Décrivez la scène.
   e. Plaisir de la photographie ou du dessin ou de la danse ou du cyclisme.
   f. Pensez-vous que garçons et filles ont les mêmes chances de carrière ?
   g. Quelles sont vos réactions devant les progrès et les réalisations de l'exploration spatiale ?

2. Choisissez un des sujets suivants. Consacrez-y environ ½ h.
   a. Décrivez un entraînement destiné à améliorer vos performances dans un sport de votre choix.
   b. Après avoir visité une entreprise, faites un rapport à vos compagnons sur les conditions de travail et les perspectives d'avenir qu'elle offre.
   c. Décrivez clairement un des appareils suivants et expliquez comment il fonctionne : un « walkie talkie », un sèche-cheveux, un aspirateur, un mélangeur-batteur (mixer).
   d. Dans votre ville, on veut créer un centre commercial où la circulation sera interdite aux véhicules. Ecrivez une lettre à un journal local pour exposer vos vues sur le projet.

1 Angleterre, *General Certificate of Education*, 1967. Fin du secondaire - Niveau ordinaire.

EPREUVE II. Quatre questions - 1 ¾ h.

1. Résumez le passage suivant en bonne prose continue et en 110 mots maximum. A la fin, indiquez combien de mots vous avez utilisés. Le passage compte 314 mots.

   Suit un texte sur la confiance exagérée dans la science et la technologie.

2. Lisez le passage suivant; ensuite, répondez aux questions.

   (Le texte décrit deux grands types de promeneurs - ceux qui suivent un guide et ceux qui partent à l'aventure - et souligne l'intérêt des excursions géologiques.)

   a. Expliquez avec vos propres mots la différence entre les deux types de promeneurs.
   b. Quel conseil l'auteur donne-t-il à propos des grottes ? Formulez-le avec vos propres mots.
   c. Expliquez brièvement pourquoi l'auteur croit que la géologie est : un *hobby* amusant - un *hobby* instructif.
   d. L'auteur écrit qu'il utilise l'expression « creuser un fossé » au sens littéral. Expliquez pourquoi le sens est ici littéral.
   e. Expliquez les expressions suivantes : précautions prescrites; être parfaitement conscient de la nature de ses actes.
   f. Choisissez quatre des mots suivants. Remplacez-les par des synonymes ou des périphrases qui pourraient être utilisés dans le texte sans en changer le sens (...).

3. Répondez, au choix, à une des deux questions suivantes :
   a. Choisissez trois mots parmi les suivants. Construisez des phrases (6 en tout) montrant que ces mots peuvent être employés dans deux sens différents (...).
   b. Définissez en une phrase trois des mots suivants : monopole - interlude - préface - microscope - antidote.

4. Répondez au choix à une des deux questions suivantes :
   a. Expliquez clairement, mais brièvement, la différence de sens entre chaque paire de phrases (porte sur *could-should; can-may; will-shall; might-must*).
   b. Récrivez correctement le passage suivant, en respectant toutes les idées. Vous pouvez changer l'expression, l'ordre des mots et des idées, l'orthographe et la ponctuation.
   (Suit un texte défectueux d'une centaine de mots.)

«Par exemple si on projetait de faire passer une route à travers une ville mais qu'une maison historique était dans son chemin alors les plans devraient être changés, entraînant des dépenses considérables, pour contourner le bâtiment créant un virage dans la route et la rendant aussi dangereuse pour les autos. Par la déviation de cette route non seulement des frais sont causés rnais le prix de différentes choses augmente spécialement si la déviation est grande parce que si la distance ajoutée est disons dix kilomètres et qu'un camion portant certains articles parcourt la route, le camion consommerait alors plus d'essence prendrait plus de temps pour arriver et dix kilomètres seraient enlevés de sa vie.»

Consignes pour la correction.

EPREUVE I. - Maximum 50 points.

Le schéma de notation qui figure ci-dessous ne constitue qu'un guide préliminaire. Des additions et des amendements pourront être apportés lors de la réunion des examinateurs qui seront convoqués après une première lecture des travaux.

Question 1 (maximum 35 points).

On attend un minimum de 400 mots, mais les compositions ne doivent pas être principalement notées en fonction de la longueur. Tenir compte du sujet choisi et de la faon dont il est traité: Même si elle est courte, une composition où l'argumentation est serrée et où l'expression est bonne doit obtenir plus de points qu'une longue narration informe.

L'examinateur devrait avoir une idée claire de ce qu'est une composition recevant tout juste la note de réussite (16 points). Pareille composition doit contenir des idées raisonnables, mais non très originales. L'expression doit être claire, mais sans distinction particulière. On ne devrait rencontrer, dans le travail, que quelques erreurs mécaniques. Les candidats qui ont dépassé ce niveau général doivent être récompensés et ceux qui ont travaillé en dessous de ce niveau doivent être pénalisés.

On trouvera dans le document annexé des notes détaillées sur les qualités à observer dans les compositions. On compte toutefois que les examinateurs noteront en fonction de leur impression générale, n'alloueront donc pas une proportion fixe des points pour les différents aspects. Si le sujet se prête à la controverse, les idées et leur enchaînement peuvent être plus importants que dans une composition descriptive où le vocabulaire pourrait prendre une plus grande place. Nous nous fions au jugement des examinateurs.

En réservant 35 points à cette première question, on a voulu marquer son importance majeure dans l'ensemble des examens. Beaucoup de candidats sont médiocres. Toutefois, si l'examinateur ne disperse pas largement ses notes, cette première question ne pèsera pas d'un poids suffisant dans l'ensemble des résultats.

Groupe A (29-35 points).
La composition est de qualité exceptionnelle.

Groupe B (22-28 points).
La qualité du travail est au-dessus de la moyenne.

Groupe C (14-21 points).
Le travail est de qualité moyenne.

Groupe D (7-13 points).
Le travail est d'un niveau inférieur à ce que l'on considère comme satisfaisant.

Groupe E (0-6 points).
Le candidat est incapable de présenter ses idées avec cohérence.

Cette épreuve est un examen de langue maternelle. Dans l'épreuve 1, nous examinons la capacité du candidat à exprimer ses opinions, ses expériences, ses impressions, ses sentiments et ses intérêts. Il ne s'agit ni d'un test de connaissances générales, ni d'une évaluation des aptitudes du candidat. Si une jeune fille décrit clairement une scène en anglais, elle doit obtenir une note favorable si les matériaux utilisés sont pertinents, même si l'examinateur pense que la jeune fille en question aborde le sujet de façon trop sentimentale. Par ailleurs, si derrière la masse de prose incohérente, l'examinateur a l'impression que se trouvent des émotions profondes et des attitudes morales élevées, il n'a pas à s'occuper de ce dernier aspect: seul ce qui est écrit compte. Pour réussir l'épreuve, la clarté de l'expression et la précision du style sont essentiels.

Question 1 (maximum 35 points)

(a) Une narration ou une description sont acceptables. Dans une narration, la nuit brumeuse doit jouer un rôle essentiel.

(b) En gros, les deux scènes doivent être équilibrées. Récompenser la vivacité de narration, de dialogue, de description et le contraste.

(c) On attend une définition claire de la politique dont plusieurs points doivent être développés. Le candidat doit avoir choisi la radio ou la télévision; il ne peut avoir réuni les deux.

(d) Le sujet du travail est la dispersion; autoriser néanmoins une courte introduction.

(e) Une réponse cohérente, claire, de longueur modérée fait plus que de longs errements et des répétitions.

(f) Ce sujet n'est pas facile, récompenser généreusement la bonne ordonnance des arguments et les exemples bien choisis.

(g) Un traitement purement narratif ne peut pas être accepté, mais quelques exemples de progrès réalisés peuvent être nécessaires pour expliquer les réactions.

Question 2 (maximum 15 points).

(a) On attend une description claire et logique.

(b) Le rapport doit traiter les trois aspects de la question. On insiste surtout sur l'information rapportée; on n'exigera donc pas une forme de rapport particulier.

(c) On exige à la fois une description de l'objet et une explication de son fonctionnement. Ne punissez pas sévèrement les erreurs matérielles; notez simplement la clarté de l'expression.

(d) On peut envisager de nombreux aspects. Aussi, pénalisez toute idée inadéquate.

Décomptez:
1 point pour une mauvaise rédaction de l'adresse.
1 point pour un manque de cohérence entre la vedette et les salutations.
1 point si l'élève a signé «M. John Smith» ou «Mademoiselle Jeannette Smith».
$\frac{1}{2}$ point pour d'autres erreurs de disposition ou de ponctuation, ou d'orthographe dans des mots essentiels.

Tous ces sujets fournissent un matériau suffisant pour une demi-heure de travail. On attend un minimum de 200 mots. Le choix est riche.

EPREUVE II - Maximum 50 points.

Question 1 (maximum 16 points).

Les points attribués à cette question sont habituellement beaucoup plus bas que pour les autres. Les examinateurs sont priés de ne pas considérer 11 comme le maximum attribuable au résumé.

A. - Attribuer un maximum de 2 points pour chacun des aspects suivants. Pour obtenir 2 points, le candidat doit avoir clairement compris l'idée et l'avoir exprimée correctement. Nuancez vos notes par 1 $\frac{1}{2}$, 1 ou $\frac{1}{2}$ point. A la fin du travail, comptez une nouvelle fois l'ensemble en fonction de la fluidité et de la cohérence du résumé complet. Si un passage est incohérent, supprimez au maximum $\frac{1}{4}$ des points attribués; s'il est plutôt rocailleux, soustrayez $\frac{1}{8}$. Cette correction devrait apparaître sur la composition en écrivant par exemple 10 - 1 = 9. Toute soustraction de points destinée à pénaliser un texte plus long que la limite fixée ou une proposition non seulement inexacte mais absurde doit apparaître séparément. Le total final doit être entouré d'un cercle dans la marge.

1° L'homme de la rue accepte aujourd'hui les découvertes scientifiques,

2° sans douter de leur origine, de leur validité, ou de leurs effets.

3° et 4°. La demande de nouveaux progrès destinés à élever le niveau de vie ne se ralentit jamais.

5° Les hommes ont confiance dans l'homme de science et dans son travail,

6° et croient qu'on ne peut arrêter le progrès.

7° Bien qu'ils reconnaissent que les hommes de science ne sont pas toujours d'accord sur les sécurités apportées par les nouvelles découvertes,

8° le public est convaincu que les hommes de science finiront par trouver l'unanimité ou au moins un large accord.

B. - La limite de 120 mots donne une marge généreuse. Supprimez un point pour chaque tranche de 5 mots dépassant le maximum. Ne comprenez pas dans le comptage des mots introductifs tels que « Dans ce passage, l'auteur explique que... ». Recomptez les mots. N'acceptez pas simplement le nombre inscrit par le candidat.

Question 2 (maximum 20 points).

En notant cette question, accordez le maximum dans chaque section au candidat qui expose nettement le point.

(a) (1) Ceux qui aiment que l'on prévoie tout à leur place (1)
(2) et ceux qui préfèrent disposer simplement d'un plan général permettant de suivre l'inspiration du moment (esprit d'exploration). (2)

(b) Visiter des grottes. (1)
A moins que vous ne soyez guidé par quelqu'un qui connaît très bien le terrain. (1)
Respectez toutes les règles de sécurité. (2)

(c) (1) Il ne faut guère d'équipement spécial. (2)
(2) Le géologue fait continuellement de nouvelles découvertes (1 ½)
à petite échelle. (½)

(d) Quand le géologue casse une pierre, il pose, en fait, un acte unique, car personne après lui ne pourra casser à nouveau cette même pierre. (2)

(e) (1) Mesures de sécurité (1)
imposées ou recommandées. (1)
(2) Pour comprendre exactement (1)
les conséquences de ce qu'on fait. (1)

(f) Souterrain :

qui est sous le sol, sous la surface de la terre; (1)
sous terre. (1)

Spéléologues :

personnes qui étudient les grottes scientifiquement;
explorateurs de grottes. (1)

Consciencieusement :

de bonne foi, sans se laisser distraire,
avec beaucoup d'application (1)
honnêtement (½)
fidèlement (½)

Etc.

Pour les sous-questions (a), (b) et (e), n'attribuez aucun point aux élèves qui se bornent à recopier une partie du texte.

Question 3 (maximum 6 points).

(a) Attribuez un point pour toute phrase construite correctement. N'attribuez aucun point en cas de construction incorrecte.

(b) Accordez un point pour chaque définition exacte exprimée dans une phrase correcte (maximum 3 points).
Accordez un point pour chaque phrase où le sens du mot commençant par le même préfixe ressort clairement (maximum 3 points).
½ point pour une définition exprimée en une phrase incorrecte.
½ point pour une définition fournie en phrase incomplète.
Aucun point si le mot commençant par le même préfixe n'est pas présenté dans une phrase.
Un monopole est une propriété exclusive détenue par une firme (ou)
Un monopole est le nom donné à une firme qui détient des droits commerciaux exclusifs.
Un *interlude* est un intervalle ménagé au cours de la représentation d'une pièce (ou un intervalle dans le déroulement d'un événement).
Etc.

Question 2 (maximum 8 points).

(a) Accordez un point par phrase.
Vérifiez la présence des idées suivantes...
N'accordez aucun point si l'on ne voit pas clairement à quelle phrase le candidat se réfère.

(b) Les candidats répondront de façons différentes.

Déduisez 1 point par faute d'orthographe, 2 points pour toute construction boiteuse, 1 point pour une expression incorrecte, ½ point pour l'omission ou l'emploi erroné d'une virgule essentielle, 1 point pour chaque idée omise.

Revoyez la note totale en fonction de l'impression générale du passage.

Pour les questions 1 et 2, si le candidat obtient la moitié des points, arrondissez à l'unité supérieure (ex.: 6 ½ = 7).

Si les deux questions donnent une note comprenant ½ point, arrondissez l'une par excès et l'autre par défaut.

Mêmes remarques pour les questions 3 et 4.

ANNEXE III

EXEMPLE D'ENSEIGNEMENT SEMI-INDIVIDUALISE

La New Trier Township High School, Winnetka.

La *New Trier Township High School* est une grande école du degré secondaire supérieur où l'enseignement est semi-individualisé et où l'étudiant peut corriger son orientation jusqu'au terme de l'adolescence.

Elle accueille les élèves à partir de 14 ans[1] et jouit d'une grande réputation tant aux Etats-Unis qu'à l'étranger.

Au moment de notre visite, 3 740 étudiants suivaient régulièrement les cours[2]. Le corps professoral comptait 255 membres dont:

40	professeurs	d'anglais
31	"	de mathématiques
30	"	d'éducation physique
29	"	de langues étrangères
27	"	de «Social Studies» (Histoire, géographie, civisme, sociologie et économie
17	"	de sciences
9	"	de musique
7	"	de cours techniques
7	"	de commerce
6	"	d'art dramatique
6	"	de peinture-dessin
6	"	d'automobile
4	"	d'économie domestique
2	"	d'hygiène
1	"	d'enseignement spécial (retardés mentaux éducables).

On comptait un enseignant pour 15-16 étudiants et les classes réunissaient généralement 25 élèves environ.

1 L'organisation de l'enseignement des Etats-Unis n'est pas uniforme. La N.T. High School relève du système «NK 8-4,: un an de Nursery School, un an de jardin d'enfants, 8 ans de primaire et 4 ans de secondaire. Les deux autres systèmes les plus fréquents sont: NK 6-3-3 et NK 6-6. On rencontre aussi NK 7-5, NK 6-2-4 et NK 6-4-4.

2 Ces chiffres se réfèrent à l'année scolaire 1959-1960. Ils sont extraits de: *Information for College Admission Officers*, N.T. Township High School, nov. 1959, ou ont été recueillis sur place.

Les prestations d'un professeur qui n'assurait pas de responsabilités spéciales (telles que président d'un département, par ex.) comportaient 24 périodes de 40 minutes par semaine : 4 cours de 5 périodes et 4 périodes de *counseling*.

Dans ce type d'école, le niveau intellectuel des élèves, jugé sur la base des tests classiques, est élevé. En 1960, environ 80 % de ceux qui terminaient leurs études à New Trier obtenaient des résultats supérieurs à la moyenne nationale, dans les tests d'aptitudes et de connaissances (SCAT, STEP, *National Merit*). 92 % des diplômés continuaient des études supérieures.

L'individualisation des programmes.

Le principe fondamental de l'action pédagogique de la New Trier High School est défini dans la première phrase de son programme de cours : « Personne ne peut croire à la dignité des hommes sans éprouver un plaisir profond au spectacle de leur variété infinie. »[1]

Restant imprégnées de l'esprit du plan de Winnetka de Washburne, les études sont organisées de façon à fournir à chacun la possibilité de se développer à son rythme propre, selon ses capacités et ses penchants. Ceci ne signifie nullement que la fantaisie stérile ou que les solutions de facilité soient permises.

Selon le système répandu dans tout le pays, l'étudiant ne peut obtenir un diplôme de fin d'études que si, au cours de celles-ci, il a gagné un nombre total de points ou de *crédits* fixé par le conseil d'administration de l'école qui tient lui-même compte de normes générales. Ces *crédits* expriment en une unité conventionnelle l'importance qualitative et quantitative des différents cours, pour un semestre (5 *crédits* correspondent, par exemple, à un semestre de cours mineur, à raison de cinq périodes par semaine). Les *crédits* ne sont acquis que si l'étudiant atteint une note supérieure à une limite minima fixée.

Pour obtenir le certificat d'études de la New Trier High School, il faut y avoir gagné 350 *crédits* en 4 ans.

Le jeu des *crédits* est à la fois source de souplesse et de sécurité, car il permet de délimiter exactement le champ de liberté de l'élève.

En principe, celui-ci établit son programme d'études comme on compose le menu d'un repas en choisissant, sur la carte, ce qui l'attire et lui convient le mieux.

1 NEW TRIER TOWNSHIP HIGH SCHOOL, *Curriculum Guide*, déc. 1959, p. 1.

Voici, par exemple, la liste des cours offerts pour la première année, avec mention des crédits qu'ils rapportent s'ils sont suivis avec succès pendant un semestre[1].

Cours « majeurs »

(Rapportant 10 crédits par semestre, sauf indication contraire, ils doivent être suivis pendant 2 semestres consécutifs)

Anglais	Langues étrangères
Algèbre	Mathématiques générales
Alimentation (1 semestre)	Peinture-dessin (2 périodes par jour)
Civisme	Photographie
Commerce	Radio amateur
Dessin industriel	Travail du bois
Electricité (1 semestre)	Travaux manuels
Histoire universelle	Vêtement (1 semestre)

Cours « mineurs »

(Nombre de *crédits* entre parenthèses)

Art dramatique	(6)	Dessin industriel (1 période par jour)	(5)
Chant choral	(3)	Diction	(6)
Dactylographie	(5)	Harmonie	(5)
Cours d'harmonie	(6)	Orchestre symphonique	(5)
		Travaux manuels (1 période par jour)	(5)

Néanmoins, plusieurs restrictions influencent le choix de l'élève. En premier lieu, un certain nombre de cours, jugés indispensables à la culture de base de tous les membres de la nation, sont obligatoires :

Cours	*Durée obligatoire*	*Crédits*
Langue maternelle	4 ans	80
Mathématiques	2 ans	40
« Social Studies » (Histoire - Géographie)	2 ans[2]	40
Sciences	1 an	20
Education physique	4 ans	16
Automobile : théorie et pilotage	1 semestre	3
	Total	199

1 Cf. NEW TRIER TOWNSHIP HIGH SCHOOL, *Registration Bulletin for Freshmen 1959-1960. Planning a Course of Study*, p. 2.

2 Dont un an obligatoirement consacré à l'histoire des Etats-Unis.

Environ $^2/_3$ du total des crédits exigés sont donc fournis par des cours imposés. Toutefois, comme nous le verrons, il peut être satisfait à ces exigences de façon fort libre; ainsi, nous avons relevé une dizaine de possibilités différentes permettant d'accomplir valablement l'année de sciences réclamée.

Un second facteur important guide l'étudiant dans l'élaboration de son programme: la profession ou les études supérieures auxquelles il aspire. En particulier, chaque université détermine ses conditions d'admission et spécifie notamment combien de crédits l'étudiant doit avoir acquis dans l'enseignement secondaire, pour des branches déterminées.

Enfin, les parents, les maîtres, les conseillers pédagogiques et les orienteurs s'efforcent de guider l'étudiant au mieux de ses intérêts, veillant à ce qu'il tire le meilleur profit possible de ses potentialités.

La semaine scolaire comptant 5 jours de 8 périodes de 40 minutes - réunion quotidienne avec le conseiller pédagogique 20 min) et durée du lunch toujours pris à l'école (25 min) non comprises -, on suggère généralement à l'élève de se constituer un programme moyennement chargé, ménageant du temps pour l'étude personnelle. Voici quatre exemples types pour la première année[1]:

I		II		III		IV	
Cours	*pér. sem.*	*Cours*	*pér. sem.*	*Cours*	*pér. sem.*	*Cours*	*pér. sem.*
Anglais	5	Anglais	5	Anglais	5	Anglais	5
Algèbre	5	Algèbre	5	Algèbre	5	Algèbre	5
Latin	5	Sciences	5	Histoire	5	Histoire	5
Peint./dess.	10	Commerce	5	Aliment.	7	Sciences	5
Gymn. (filles)	4	Gymn. (garç.)	5	Chant	3	Lang. étrang.	5
	-	Musique instr.	5	Gymn. (filles)	4	Gymn. (garç)	5
	29		-		-		-
Etude			30		29		30
personnelle	11		10		11		10
	-		-		-		-
	40		40		40		40

On remarquera combien ces plans de travail contrastent avec l'émiettement de l'effort si fréquent chez nous. Pratiquement, tous les cours académiques peuvent être concentrés dans la matinée.

1 NEW TRIER TOWNSHIP HIGH SCHOOL, *Courses for Freshmen*, avril 1959, p. 4.

Environ un quart du temps passé à l'école est réservé aux études personnelles. Les heures ainsi laissées libres seront souvent consacrées à des travaux de recherche dans la bibliothèque scolaire qui, aux Etats-Unis, jouent un rôle incomparablement plus important qu'en Belgique ou en France.

L'individualisation de l'enseignement.

Non seulement l'élève choisit les branches qui lui conviennent le mieux, mais encore l'enseignement de chacune de celles-ci sera adapté à ses possibilités.

En effet, toutes les branches importantes peuvent être étudiées à cinq niveaux d'aptitudes différents: inférieur, moyen faible, normal, accéléré, avancé. De cette facon, l'effort réclamé à l'étudiant, qu'il possède une intelligence supérieure ou soit peu doué, qu'il soit fort dans un domaine et en retard dans un autre, est toujours en proportion avec ses possibilités.

Comme l'élève suit rarement tous ses cours à un même niveau, le danger d'une ségrégation générale, selon les aptitudes, semble minime: l'école reste d'ailleurs très attentive à ce problème, regroupant systématiquement tous les étudiants à l'occasion de certaines activités. On met aussi tout en œuvre pour inculquer une véritable tolérance vis-à-vis du plus ou moins grand talent des compagnons d'études et pour encourager chacun à se dépasser: En sport, tout le monde n'est pas capable de jouer en excellence; c'est aussi le cas dans les études. Mais on attend de chacun le meilleur de lui-même[1].

Le système d'enseignement à différents niveaux est pratiqué à New Trier, depuis la fin de la Première Guerre mondiale, à la plus grande satisfaction de tous. En moyenne, les étudiants se répartissent selon les pourcentages suivants:

- supérieurs 15-20%
- normaux (moyens forts) 40-55%
- moyens faibles 36-40%
- limités 5- 8%
- avancés (seniors) ± 10%

Quand les élèves entrent à la New Trier High School, les groupes sont provisoirement déterminés sur la base des résultats scolaires antérieurs et d'autres renseignements réunis selon un système que nous étudierons plus loin. Par après, les résultats obtenus dans l'école même corrigeront et guideront les affectations, celles-ci n'étant jamais définitives, quelle que soit la branche.

1 NEW TRIER HIGH SCHOOL, *Guide Book to New Trier*, Winnetka, 1959, p. 39.

Pareille souplesse conduit certes à une organisation complexe, mais il ne faut cependant pas en exagérer la difficulté.

Dans le tableau suivant, nous faisons apparaître toutes les possibilités offertes pendant le second semestre de l'année scolaire 1959-1960. Quelques notes marginales précisent l'esprit des études[1].

Chaque cours est désigné par un nombre de trois chiffres:

a) Le chiffre des centaines indique en quelle année il se donne; on aura donc 1, 2, 3 et 4.

b) Le chiffre des dizaines indique le semestre:
   1 = 1er semestre
   2 = 2e semestre
   0 = peut être suivi au 1er ou au 2e semestre
   3 = cours de vacances d'été.

c) Le chiffre des unités indique le niveau d'aptitude:
   1 = niveau inférieur
   2 = moyen faible
   3 = normal
   4 = enseignement accéléré
   5 = avancé
   6 = séminaires
   9 = tous niveaux réunis.

Branches et niveaux de cours	*Observations*
Langue maternelle	
125	- Pour étudiants supérieurement intelligents, excellents en langue maternelle (intègre la langue maternelle, la biologie et l'histoire)
124	- Cour enrichi, fondé sur les œuvres littéraires de niveau universel
123	- Cours moyen
122	- Cours moyen inférieur: s'adresse plus particulièrement aux élèves qui n'ont pas encore acquis une méthodologie de travail rationnelle
122 R	- R = «remedial». Destiné aux élèves identifiés par le bureau de testing comme présentant une déficience marquée en lecture (compréhension et rapidité) et en orthographe
121	- Accueille les élèves présentant une faiblesse générale dans toutes les branches de la langue maternelle.
224 - 223 - 222 - 221 324 - 323 - 322 - 321 424 - 424 (+) - 423 - 422 - 421	Les différences de niveaux en 2e, 3e et 4e années sont *mutatis mutandis* parallèles à celles que nous indiquons ci-dessus pour la 1re. 323 = journalisme. 424 (+) = Littérature classique

1 Les deux documents de base suivants ont été utilisés: NEW TRIER TOWNSHIP HIGH SCHOOL, *Curriculum Guide*, 1959 et *Program of Classes*, 1959-1960.

Branches et niveaux de cours		*Observations*
Mathématiques		La différence entre les niveaux 2, 3 et 4 réside moins dans l'accélération que dans la méthode d'enseignement et l'approfondissement. *Au niveau 2,* les explications sont détaillées: beaucoup d'applications; pas d'incursions dans les domaines voisins. *Au niveau 3,* l'étudiant doit plus travailler par lui-même: théorie plus rigoureuse, enrichissement. *Au niveau 4:* les concepts sont traités rapidement; étude plus approfondie; nombreuses incursions dans les domaines voisins. Le même manuel est employé pour ces 3 niveaux et tous les élèves étudient le même chapitre en même temps, ce qui permet, à tout moment, le passage d'un niveau à l'autre.
Algèbre A (accéléré)	124	
" –	124	
" E (expérimental)	124	
" –	123	
" E	123	
" –	122	
" E	122	
Mathématiques de base	121	
Mathématiques approfondies *(20 crédits)*	225	
Géométrie	224	
"	223	
"	222	
Mathématiques	222	
Mathématiques de base	221	
Mathématiques approfondies *(20 crédits)*	315	
Mathématiques	324	
Algèbre	323	
"	322	
Géométrie	303	
Mathématiques approfondies	425	4 semestres de mathématiques sont obligatoires: toutefois, 2 de ceux-ci peuvent être consacrés à la comptabilité ou à l'arithmétique commerciale.
Mathématiques	424	
Algèbre (niveau universitaire)	403	
Trigonométrie	403	
Usage de la règle à calculs *(2 crédits;* va avec algèbre 322)	323	
«Social studies»		4 semestres de «Social Studies» sont obligatoires; 2 de ceux-ci doivent être consacrés à l'histoire des Etats-Unis.
Civisme	123, 122, 121	
Histoire universelle	125, 124, 123	
Histoire: temps modernes	324	
Histoire: antiquité	204	
Histoire: moyen âge	202, 203, 204	
Histoire: universelle	221	
Géographie	222 - 223	
Histoire: temps modernes	323	
Histoire: Grande-Bretagne	303	
Histoire: U.S.A.	324 - 323 - 322 - 321	
Histoire: U.S.A. (approfondie)	425	
Histoire: contemporaine (Europe)	425	
Histoire: Grande-Bretagne (approf.)	404	
Histoire: Amérique latine	403	
Histoire: Extrême-Orient	403	
Histoire: Universelle, XXe S.	402	
Histoire: Etats-Unis, XXe S.	402	
Civisme	403	
Sociologie	403	
Sciences économiques	403	

Branches et niveaux de cours		Observations
Sciences		Les élèves qui désirent continuer des études dans les Facultés de Sciences des Universités sont invités à suivre 6 semestres de sciences et 8 de mathématiques.
Biologie	123 - 124 - 125	
Radio amateur	123	
Biologie	224 - 223 - 222	
Chimie	225	
Biologie	323 - 322 - 321	
Electronique	323	
Physique	325	
Chimie	324 - 323	
Chimie (séminaire)	426	
Physique	425 - 424 - 423	
Sciences	409	
Langues		Aucun cours de langue n'est obligatoire. Les étudiants sont cependant fort encouragés à les suivre s'ils se destinent aux études supérieures. On pense que, très prochainement, les universités exigeront à l'entrée que l'étudiant ait étudié au moins une langue étrangère pendant 3 ans.
Latin	124 - 123 - 122	
	224 - 223 - 222	
	324 - 323	
	424 - 423	
Allemand	124 - 123	
	324 - 323	
	426	
Russe	124	
	224	
	324	
Français	125 - 124 - 123	C indique qu'il s'agit de la continuation d'un cours précédent. = cours de conversation.
	224 C - 224 - 223 C - 223	
	324 C - 324 - 323 C - 323 - 324 =	
	424 C - 423 C	
Espagnol	125 - 124 - 123 - 122	
224 - 224 C - 223 - 223 C - 222		
	324 - 324 = - 323 - 322	
	424 - 423	
Arts		Tous ces cours peuvent être suivis à raison de 2 périodes par jour. (= cours «majeur» 10 *crédits*) ou de 1 période par jour (= cours «mineur»: 5 *crédits*). Toutefois, Histoire de l'art 329 est toujours un «majeur».
Dessin-peinture	129	
Travail manuel	121	
Dessin-peinture Travail manuel	229 C - 229	
Dessin-peinture	329	
Céramique	329	
Joaillerie	329	
Histoire de l'art	329	
Céramique	429	
Joaillerie	429	
Peinture-dessin	429	

Branches et niveaux de cours		Observations
Commerce		*Objectifs poursuivis par ce cours:* 1. Préparation aux études supérieures de sciences commerciales; 2. Formation générale; 3. Préparation pour les étudiants qui veulent travailler à temps partiel dans le commerce pendant leurs études universitaires; 4. Préparation professionnelle pour les élèves qui ne feront pas d'études supérieures; 5. Possibilité pour les étudiants qui cherchent leur voie de voir si la carrière commerciale pourrait les intéresser. Le 422 = une réunion de 30 min tous les matins à l'école; le reste consiste en stages pratiques (20 h/semaine).
Commerce général	122 - 123	
Commerce général	222	
Comptabilité	223	
Le point de vue du consommateur	323	
Vente	302	
Publicité	303	
Organisation commerciale	303	
Sténographie	323	
Dactylographie	303	
Pratique du bureau	322	
Secrétariat	423	
Sténographie	423	
Droit commercial	423 - 403	
Pratique commerciale	422	
Automobile		Une loi, passée en 1955, rend ce cours obligatoire dans toutes les High Schools de l'Illinois. (3 «crédits»).
Théorie et pratique (pilotage)	209	
Economie domestique		Destiné principalement aux jeunes filles; toutefois: - les garçons s'intéressant à la restauration ou à l'hôtellerie peuvent suivre les cours d'alimentation; - les garçons s'intéressant à l'architecture peuvent suivre les cours de décoration intérieure.
Alimentation	109	
Vêtement	109	
Alimentation	209	
Vêtement	209	
Décoration intérieure	329	
Cours techniques		- Le but poursuivi n'est pas de donner une formation professionnelle, mais de faire de l'étudiant «un consommateur intelligent des produits de l'industrie» (p. 64). - Chacun de ces cours rapporte 5 ou 10 «crédits»; de plus en plus, elles souhaitent que les étudiants acquièrent une formation technique.
Dessin industriel	123 - 122	
Bois	123 - 121	
Electricité	103	
Dessin (architecture)	223	
Bois	222	
Métallurgie générale	223	
Dessin (mécanique)	323	
Métallurgie générale	323	
Notions d'architecture	323	
Lecture de plans	302	

Branches et niveaux de cours		*Observations*
Musique		Il s'agit soit d'une initiation pour l'amateur, soit d'études approfondies pouvant préparer aux carrières musicales.
Chant	129	
Orchestre	129	
Cours d'harmonie	129	
Appréciation musicale	129	
Chant	229	
Petit groupe vocal	229	
Appréciation musicale	229	
Composition	329	
Madrigaux (groupe vocal)	429	
Composition	429	
Chorale (sélection de 32 étud.)	029	
Opéra	029	
Grande chorale	029	
Harmonie (cadets - groupe de sorties - groupe d'honneur)	029	
Orchestre (seniors)	029	
Ensemble instrumental	029	
Piano	029	
Education physique - Sports - Danse		
Art oratoire - Art dramatique		
Photographie		
Croix-Rouge		

L'étude de chaque élève avant son entrée à New Trier.

Nous avons dit que l'école recueille des informations détaillées sur ses futurs élèves afin que ceux-ci puissent trouver immédiatement une place, au moins provisoire, dans le système complexe que nous venons de décrire. Voici comment on procède.

La majorité des étudiants qui fréquentent New Trier proviennent de six écoles publiques et de sept écoles confessionnelles des alentours.

Dès le mois de janvier, les services de testing de la *High School* fournissent aux six écoles publiques des tests d'intelligence, d'aptitudes et de connaissances à administrer aux enfants qui ont manifesté l'intention de s'inscrire aux cours de New Trier. Les élèves des écoles confessionnelles viendront à Winnetka, en avril, pour y subir des épreuves similaires.

Chaque école doit en outre remplir une fiche individuelle qui porte, outre les résultats des tests mentionnés, un tableau résumant l'appréciation générale des professeurs[1]

APPRECIATIONS GLOBALES					
Intelligence	1	2	3	4	5
Application	1	2	3	4	5
Sens des responsabilités	1	2	3	4	5
Conduite à l'école	1	2	3	4	5
Qualités de chef	1	2	3	4	5
APPRECIATIONS SPECIALES					
Anglais	1	2	3	4	5
Mathématiques:					
Raisonnement	1	2	3	4	5
Connaissances de base	1	2	3	4	5
«Social Studies»	1	2	3	4	5
Sciences	1	2	3	4	5

Code: 1 = supérieur; 2 = au-dessus de la moyenne; 3 = moyen; 4 = en dessous de la moyenne; 5 = pauvre. Le chiffre retenu est encadré.

Après l'administration des tests, le directeur du service de *counseling* de New Trier rencontre les conseillers des différentes écoles afin de recueillir des renseignements complémentaires concernant non seulement l'instruction, mais aussi l'histoire, la famille, la santé des futurs étudiants.

Ensuite, chacun de ceux-ci est interviewé dans son école. L'entretien porte d'abord sur les talents (arts, sports, bricolage) et sur les goûts (branches préférées), puis sur la famille, l'enquêteur tâchant d'identifier les problèmes éventuels. Une question choc termine l'entrevue.

Au mois de mai, parents et élèves - qui ont déjà pu étudier le programme des cours - sont invités à assister à une réunion d'information à la High School. Les inscriptions aux différents cours sont recueillies.

La direction de New Trier possède alors assez d'éléments pour constituer des groupes d'environ 30 étudiants qui, pendant toutes leurs études, seront confiés au même conseiller pédagogique et le rencontreront chaque matin. Dans ces groupes, l'hétérogénéité est systématiquement recherchée, tant en ce qui concerne les aptitudes que les ori-

1 Cf. NEW TRIER HIGH SCHOOL, *Test and Personal Data Card.*

gines socio-économiques. Toutefois, les sexes sont séparés, le conseiller étant toujours du même sexe que ses étudiants. De plus, on s'efforce de donner une personnalité équilibrée à chaque groupe en évitant notamment que trop d'adolescents «à problèmes» ne s'y trouvent réunis.

Le jour de la rentrée, non seulement les cours et les horaires sont organisés, mais chaque étudiant sait exactement ce qu'il doit faire: il sait, par exemple, que pour telle branche, il doit se joindre au groupe des moyens et pour telle autre, au groupe supérieur; il sait qu'un conseiller l'attend; il sait même déjà quels livres, quels cahiers, quel équipement sportif il doit acheter et on lui a fait connaître d'avance le prix exact de chacun de ces articles.

Terminons ces indications concernant les débuts scolaires en signalant qu'environ 15 jours après la rentrée des classes, les parents sont invités à une première réunion où ils rencontreront le conseiller de leurs enfants et où ils seront éclairés sur les raisons du classement aux différents niveaux, sur la signification des manifestations extrascolaires, etc.

La semaine suivante, les pères et les mères se rendent à l'école avec leur enfant et suivent une journée de classe complète à ses côtés (la journée est, à cette occasion, répartie sur l'après-midi et la soirée). On devine aisément combien une telle expérience facilite la compréhension entre l'école et la famille.

Ayant commencé ses études secondaires supérieures dans ces conditions excellentes, l'étudiant sera suivi, jour après jour, par son conseiller.

Les conseillers pédagogiques.

Grâce à ses conseillers pédagogiques, la New Trier High School réussit ce véritable tour de force qui consiste, non seulement à guider efficacement chacun de ses 3.740 étudiants dans le dédale des programmes, mais aussi à les aider tant sur le plan psychologique que sur le plan social et médical.

Le *counseling* est assuré par:

126 conseillers pédagogiques à temps partiel, professeurs consacrant $^1/_5$ de leur temps à cette mission.
10 conseillers à temps plein dirigeant les précédents.
6 psychologues à temps plein.

Au cours de la réunion quotidienne du matin, chaque groupe de 30 étudiants dont nous avons signalé la constitution rencontre son conseiller pour discuter de problèmes de discipline, de carrière et, en général, de tout ce qui peut intéresser la majorité. L'atmosphère détendue de ces rencontres constitue une excellente transition vers le travail ardu qui va bientôt commencer.

L'étudiant peut aussi consulter individuellement son conseiller, chaque fois qu'il le désire.

De plus, pendant l'année, le conseiller rendra au moins une visite personnelle à la famille de chaque élève appartenant à son groupe.

Il va sans dire qu'une mission aussi délicate ne s'improvise pas: elle réclame des éducateurs ouverts aux problèmes de la jeunesse et spécialement formés pour l'aider.

Lors du recrutement des professeurs, le conseil d'administration de l'école attache une importance toute particulière aux qualités intellectuelles et morales indispensables au parrainage des étudiants et donne la priorité à ceux qui les possèdent.

Pendant sa première année de fonction, le conseiller doit assister à 50 réunions de formation d'environ 1/2 heure chacune et tous les quatre ans, il est soumis à un nouvel entraînement.

La préparation de base se déroule de la façon suivante:

I. *Avant la rentrée des classes:*

Un certain nombre de séances sont d'abord consacrées à l'étude de l'administration de l'école. Ensuite, le conseiller reçoit en communication toutes les informations recueillies sur ses futurs élèves.

Enfin, quelques jours avant la fin des vacances, le directeur des études de première année réunit les «freshman helpers», c'est-à-dire des étudiants de dernière année qui vont aider le conseiller dans sa tâche. Pendant les 9 premières semaines, l'assistant participera à toutes les réunions matinales. Pendant les 9 semaines suivantes, il ne fournira plus que deux prestations par semaine; par après, il n'interviendra plus que sur demande du conseiller.

II. *Pendant l'année scolaire.*

Sous la présidence d'un directeur spécialisé, la formation se poursuit lors de causeries familières faites par les directeurs des départements (anglais, mathématiques, sciences, etc.), par les chefs

des différents services (directeurs du testing, psychologue, bibliothécaire, médecin-assistant social, délégué à l'association des parents) et par les adultes responsables des clubs estudiantins.

Le directeur des conseillers convoque aussi certaines réunions spéciales pour envisager des problèmes nouveaux concernant l'admission d'élèves, la révision des programmes, etc.

Enfin, en avril, une conférence de clôture permet de dresser le bilan d'activité de l'année qui touche à sa fin et d'établir les premières prévisions pour l'année suivante.

BIBLIOGRAPHIE

AGAZZI A., *Les aspects pédagogiques des examens*, Strasbourg, Conseil de l'Europe, C.C.C., 1967.

ALLAL L., CARDINET J., PERRENOUD P., *L'évaluation formative dans un enseignement différencié*, Berne, Lang, 1979.

ANGOFF W.H., Scales, Norms and Equivalent Scores. *In* R.L. THORNDIKE, Ed., *Educational Measurement*, Washington, American Council on Education, 1971, 2e éd.

BACHER F., La normalisation de la notation, in *Docimologie et Education*, numéro spécial de la revue «les Sciences de l'Education», nos 2-3, 1969, 131-156.

BACHER F., L'évaluation des résultats scolaires au niveau de l'école moyenne, in *Le Travail humain*, 1965, 28, 219-230.

BACHER F., La docimologie, in *Traité de Psychologie appliquée*, VI, Paris, P.U.F., 1973, pp. 27-86.

BAIN D., Pour une formation à l'évaluation formative intégrée à la didactique. *In* M.G. THURLER et P. PERRENOUD, Ed., *Savoir évaluer pour mieux enseigner*, Genève, Cahiers du Service de la Recherche Sociologique, 1988, 26, 21-37.

BAKER E.L., Beyond objectives: Domain referenced achievement. *In* W. HIVELY Ed., *Domain referenced testing*, Englewood Cliffs, N.J., Educational Technology Publication, 1974.

BAZIN R., Les Français d'aujourd'hui et leurs examens, in *Education et Gestion*, 4, 1970, 3-10.

BERK R.A., A Consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, 1986, 56, 1, 137-172.

BETZ D., Rhythmische Schwankungen als Fehler in der Notengebung, in *Psychologie in Erziehung und Unterricht*, 21, 1974, p. 1-4.

BLOCK J., *Mastery Learning*, New York, Holt, Rinehart and Winston, 1970.

BLOOM B., HASTINGS, J.T. and MADAUS, G.F., *Formative and Summative Evaluation of Student Learning*, New York, Mc Graw-Hill, 1970.

BLOOM B.S., *Time and Learning*, Communication au 81e Congrès de l'American Psychological Association, Montréal, 1973.

BLOOM B.S., *Caractéristiques individuelles et apprentissages scolaires*, Bruxelles, Labor, Paris, Nathan, 1979.

BONBOIR A., *La docimologie*, Paris, P.U.F., 1972.

BONNARDEL R., Application de la méthode d'analyse factorielle de Thurstone à l'étude de la notation des copies d'examens, in *Le Travail humain*, VIII, 1946, 130-139.

BOURDIEU P. et PASSERON, J.-C., *La reproduction. Eléments pour une théorie du système d'enseignement*, Paris, Ed. de Minuit, 1970.

BRITTON J., Experimental Marking of English Composition Written by Fifteen-Year-Olds, in *Educational Review*, Birmingham, Vol. 16.1, 1963, 17-23.

BRITTON J., MARTIN N. et ROSEN H., *Multiple Marking of English Composition, An Account of an Experiment*, London, H.M.S.O., 1966.

BRUCE G., *Secondary School Examinations. Facts and Commentary*, Oxford, Pergamon Press, 1969.

BRUNELLE L., *Pourquoi des examens ?* Paris, Société des Editions rationalistes, 1968.

CARDINET J., *L'adaptation des tests aux finalités de l'évaluation*, Neuchâtel, Institut Romand de Recherche Pédagogique et de Documentation Pédagogique, 1973, 31 pp.

CARDINET J., L'évaluation en classe : mesure ou dialogue. *European Journal of Psychology of Education*, 1987, II, 2, 133-144.

CAVERNI J.-P., FABRE J.-M., NOIZET G., Dépendances des évaluations scolaires par rapport à des évaluations antérieures, in *Le Travail humain*, 1975, 38, 213-222.

CHOPPIN B. et PURVES A., A comparison of open-ended and multiple choice items dealing with literary understanding, in *Research in the Teaching of English*, 3, 1, 1969, 15-24.

COMBER, L.C. and KEEVES J., *Science Education in Nineteen Countries*, I.E.A., Stockholm, Malmqvist, 1973.

DE BAL R., DE LANDSHEERE G., PAQUAY-BECKERS J., *Construire des échelles d'évaluation descriptives*, Bruxelles, Ministère de l'Education nationale, Organisation des Etudes, 1976.

DE LANDSHEERE G., *Dictionnaire de l'évaluation et de la recherche en éducation*, Paris, P.U.F., 1992, 2e éd.

DE LANDSHEERE V., *Faire réussir - Faire échouer. La compétence minimale et son évaluation*, Paris, Presses Universitaires de France, 1988.

DE LANDSHEERE V., *L'éducation et la formation. Science et pratique*, Paris, P.U.F., Collection 1er Cycle, 1992.

DE LANDSHEERE V. et G., *Définir les objectifs de l'éducation*, Liège, Dessain, Paris, P.U.F., 1992, 7e éd.

DEMANGEON M. et LARCEBEAU S., Une expérience de correction multiple, in *BINOP*, 1958, 14, 131-156.

Docimologie et Education, numéro spécial de la revue *Les sciences de l'Education*, 2-3, 1969, 166 pp.

EBEL, R.L., *Essentials of Educational Measurement*, Englewood Cliffs, Prentice Hall, 1979, 3e éd.

EDGEWORTH F.V., The Statistics of Examinations, in *Journal of the Royal Stat. Society*, 1988, 51, 599-635.

EISNER E.W., Instructional and expressive educational objectives. *In* J. POPHAM, Ed., *Instructional Objectives, op. cit.*.

ELLEY B.E. and LIVINGSTONE, I.D., *External Examinations and Internal Assessments. Alternative Plans for Reform, Wellington*, New Zeland Council for Educational Research, 1972.

Examen des examens, N° spécial des *Cahiers de pédagogie*, 92, septembre 1970.

Examinations Bulletins, Londres, H.M.S.O.

N° 1. *The Certificate of secondary education : some suggestions for teachers and examiners*, 1963.
N° 2. *The C.S.E. : Experimental examinations - Mathematics*, 1964.
N° 3. *The C.S.E. : An introduction to some techniques of examining.*
N° 4. *The C.S.E. : An introduction to objective-types examinations*, 1964.
N° 5. *The C.S.E. : School-based examinations*, 1965.
N° 6. *The C.S.E. : Experimental examinations : Technical drawing*, 1965.
N° 7. *The C.S.E. : Experimental examinations - Mathematics 2*, 1965.
N° 8. *The C.S.E. : Experimental examinations : Science*, 1965.
N° 9. *The C.S.E. : Trial examinations : Home economics*, 1966.
N° 10. *The C.S.E. : Experimental examinations : Music*, 1966.
N° 11. *The C.S.E. : Trial examinations - Oral English*, 1966.
N° 12. *Multiple marking of English compositions*, 1966.
N° 13. *The C.S.E. : Trial examinations : Handicraft*, 1966.
N° 14. *The C.S.E. : Trial examinations - Geography*, 1966.

N° 15. *Teachers' experience of school based examining (English and Physics)*, 1967.
N° 16. *The C.S.E. : Trial examinations - Written English*, 1967.
N° 17. *The C.S.E. : Trial examinations - Religious knowledge*, 1967.
N° 18. *The C.S.E. : The place of the personal topic - History*, 1968.
N° 19. *The C.S.E. : Practical work in science*, 1969.
N° 20. *The C.S.E. : A Group Study Approach to Research and Development*, Londres, Evans-Methuen, 1970.

FABRE J.M., *Jugement et certitude, Bern, Lang*, 1980.

FISCHER H., Wechselwirkungen zwischen Unterrichtszielen, Didaktik und Prüfungen, in *Eidgenössischen Technischen Hochschulen Bulletin* (Zürich), août 1970, 9-14.

FRENCH J.W., *Schools of Thought in Judging Excellence in English Themes*, Princeton, E.T.S., 1961.

GLASER R., Instructional technology and the measurement of learning outcomes. Some questions. *American Psychologist*, 1963, 18, 519-521.

GRISAY A., *Rendement du français : notes et échecs à l'école primaire. Les mirages de l'évaluation scolaire*, Liège, Laboratoire de pédagogie expérimentale de l'Université de Liège, 1982.

GUILFORD J.P., *The Nature of Human Intelligence*, New York, McGraw-Hill, 1967.

HAMBLETON R.K., Criterion-referenced measurement. *In* T.HUSEN et T.N. POSTLE THWAITE, *International Encyclopedia of Education*, Oxford, Pergamon, 1985, 1108-1113.

HARTOG P. and RHODES, E.C., *An Examination of Examinations*, London, McMillan, 1936.

HARTOG P., *The Marking of English Essays*, London, McMillan, 1941.

HINTON, E.M., *An Analytical Study of the Qualities of Style and Rhetoric Found in English Compositions*, New York, 1940.

HOTYAT F., *Les examens*, Paris, Bourrelier, 1962.

INGENKAMP K., *Die Fragwürdigkeit der Zensurengebung*, Weinheim, Beltz. 1971.

INIZAN A., *Le temps d'apprendre à lire*, Paris, Bourrelier, 1964.

KAUFMANN J., Note sur les problèmes de métrique en matière de notation scolaire, *Le travail humain*, 38, 1975, 133-148.

LAUGIER H. et SCHREIDER E., Recherche docimologique sur un examen de l'enseignement supérieur, in *Biotypologie*, 1958, 19, n° 2, 61-72.

LAUWERYS J.A. et SCANLON, D.G., Ed.,*Examinations*, The World Year Book of Education, 1969, London, Evans, 1969.

LECLERCQ D., *La conception des questions à choix multiple*, Bruxelles, Labor, 1986.

LECLERCQ D.,Computerized tailored testing, *European Journal of Education*, 1980, 15, 3.

LINN R.L., E.L., BAKER E.L., et DUBAR S.B., Complex, performance-based assessment: expectations and validation criteria, *Evaluation Comment*, Hiver 1991-92, 3-9.

LLOYD W.A., Les examens en Angleterre, in *Revue Française de Pédagogie*, janvier 1968.

MAGER R.F., *Comment définir les objectifs pédagogiques ?* Paris, Gauthier- Villars, 1972.

MATALON B., *L'analyse hiérarchique*, Paris Gauthier-Villars, 1975.

MATHER D., FRANCE N. et SARE G., *The C.S.E., A Handbook for Moderators*, London, Collins, 1965.

McINTOSH D., WALKER D. and McKAY D., *The Scaling of Teachers' Marks and Estimates*, Edinburgh, Oliver and Boyd, 1962, 2e éd.

MISLEVY R.J., Foundations for a new test theory. *In* N. FREDERIKSEN and I. BEJAR, *Test theory for a new generation of tests*, Lawrence Erlbaum et ass., 1990.

MONTGOMERY R.J., *Examinations, An Account of their Evolution as Administrative Devices in England*, Londres, Longmans, 1965.

NOIZET G. et CAVERNI J.-P., *Psychologie de l'évaluation scolaire*, Paris, P.U.F., 1978.

JOINT MATRICULATION BOARD, *The Marking of Scripts in Advanced Level History*, Universities of Manchester, Liverpool, Leeds, Sheffield and Birmingham, 1964.

OTTER H.S., *A Functional Language Examination*, Oxford Univ. Press, 1968.

PASSERON J.-C., Sociologie des examens, in *Education et Gestion*, 1970, 2, 6-16.

PEDLEY, F.H., *A Parents' Guide to Examinations*, Oxford, Pergamon Press, 1964.

PERRENOUD P., *La fabrication de l'excellence scolaire*, Paris, Droz, 1984.

PIAGET J., *Psychologie et pédagogie*, Paris, Denoël, 1969.

PIDGEON D. et YATES A., *An Introduction to Educational Measurement*, Londres, Routledge et Kegan Paul, 1968.

PIERON H., *Examens et docimologie*, Paris, P.U.F., 1963.

PIERON H., REUCHLIN M., et BACHER F., *Une recherche expérimentale de docimologie sur les examens oraux de physique au niveau du baccalauréat de mathématiques*, in *Biotypologie*, 1962, 23, 48-73.

PIOBETTA J.-B., *Examens et Concours*, Paris, P.U.F., 1943.

POPHAM W.J., *Criterion-referenced measurement*, Englewood Cliffs, Prentice-Hall, 1978.

POSTHUMUS K., *Levensgeheel en School*, La Haye, s. éd., 1947.

PURVES, A.C., *Literature Education in Ten Countries*, I.E.A., Stockholm, Malmqvist, 1973.

REMONDINO C., Recherche sur les systèmes numériques d'évaluation scolaire, in *Le travail humain*, 1965, 18, 3-4, 263-265.

Reports of the Secondary School Examinations Council, Londres, H.M.S.O., 1947, (1st) - 1964 (8st).

REUCHLIN M., *L'orientation pendant la période scolaire*, Strasbourg, Conseil de l'Europe, C.C.C., 1964.

REUCHLIN M. et BACHER F., L'appréciation des élèves par leurs professeurs, in *Revue française de Pédagogie*, 1968, 2, 19-25.

ROLLER S., L'évaluation du travail pédagogique, in *Educateur et bulletin corporatif* (Montreux), 1970, 36, 694-696.

ROT N. et BUJAS Z., Les distributions de notes scolaires comparées aux distributions des résultats obtenus dans les tests de connaissances, in *Le travail humain*, 1959, 22, 19-26.

SAUNDERS J.C., et MAPUIS L.L., Accuracy and consistency of expert judges in setting passing scores on criterion-referenced tests, *Communication à la Conférence annuelle de l'AERA*, La Nouvelle-Orléans, 1984.

STEWART J. et al., *Assessment for better learning*, Wellington, Ministère de l'Education, 1989, p. 19.

STIGGINS R.J., Design and development of performance assessment, *Educational measurement: Issues and practice*, 1987, 6, 3, 33-42.

THORNDIKE R.L., Marks and Marking Systems, in R.L. EBEL, *Encyclopedia of Educational Research*, Londres, McMillan, 1969, pp. 759-766.

THORNDIKE R.L., *Reading Comprehension in Fifteen Countries*, I.E.A., Stockholm, Malmqvist, 1973.

Symposium sur la docimologie, XIIIe congrès de l'Association Internationale de Psychologie (Rome, 1958), in *Le travail humain*, XXII, 1-2, janvier-juin 1959.

TYLER R., GAGNE R. et SCRIVEN N., *Perspectives of Curriculum Evaluation*, AERA Monograph series on curriculum eval., n° 1, Chicago, Rand McNally, 1967.

VALENTINE C.W., *The Reliability of Examinations*, University of London Press, 1932.

VERNON P., *Secondary School Selection*, Londres, Methuen, 1957.

VYGOTSKY L.S., *Mind in Society: the development of higher psychological processes*, Cambridge, Mass, Harvard University Press, 1978.

H. WALBERG H., *The implications of cognitive psychology for measuring school achievement*, Chicago, University of Illinois, 1991, ronéotypé, pp. 19-22.

WALKER A.S., *Pupils' School Records*, Newnes, Educ. Publ., 1955.

WENDLER J., *Standardarbeiten; Verfahren zur Objektivierung der Notengebung*, Weinheim, J. Beltz, 1969.

WISEMAN S., The Marking of English Composition in Grammar School Selection, in *British Journal of Educational Psychology*, XIX, 1949, 200-209.

WISEMAN S., *Examinations and English Education*, Manchester, University Press, 1961.

WOOD R., *Multiple choice: A State of the Art report*, Oxford, Pergamon Press, 1977 (collection *Evaluation in Education).*

WRIGLEY J., The Relative Efficiency of Intelligence and Attainment Tests as Predictors of Success in Grammar Schools, in British Journal of Educational Psychology, 25, 1955, 107-116.

YATES A. and PIDGEON D., *Admission to Grammar Schools*, London N.F.E.R., 1957.

TABLE ANALYTIQUE

TROISIEME PARTIE

CONSTRUCTION DE L'EXAMEN

QUATRIEME PARTIE

LES PROCEDURES DE MODERATION

CINQUIEME PARTIE

UNE PEDAGOGIE DE LA MAITRISE

CONCLUSIONS

ANNEXES

TABLE DES MATIERES

ACHEVÉ D'IMPRIMER
EN 1992.

Imprimé en Belgique.

La contestation des examens traditionnels a donné naissance à des ouvrages dénonçant les méfaits multiples des systèmes employés. Quelques résultats expérimentaux, presque toujours les mêmes d'ailleurs, appuient parfois la démonstration. Ainsi s'est répandue une docimologie essentiellement négative.

Réagissant contre ce courant, le professeur G. de Landsheere se contente de rappeler en quelques pages ce qu'il ne faut pas faire, puis élabore une docimologie positive, constructive.

Les maîtres, auxquels cet ouvrage est destiné, disposent d'un outil pratique pour construire les instruments d'évaluation, en synthétiser et en interpréter les résultats.

Les principaux systèmes permettant de rendre comparable les notes de différents maîtres ou de différentes écoles sont aussi étudiés.

Enfin, la dernière partie de ce précis ouvre largement la perspective vers les nouvelles théories de l'évaluation formative et de l'enseignement générateur de la maîtrise des connaissances.

Dans son introduction, l'auteur, dont la clarté d'exposition est bien connue, promet de mettre les notions de statistique nécessaires à la portée de quiconque connaît les quatre opérations arithmétiques fondamentales. Cette promesse est fidèlement tenue d'un bout à l'autre de l'ouvrage.

Mis au service de la réforme de l'enseignement et des examens, ce livre serein et réaliste, n'hésitant pas, au besoin, à présenter le pour et le contre, à reconnaître des ignorances ou à réagir contre des novateurs trop enthousiastes, doit être lu et utilisé par les maîtres de tous les niveaux.

ISBN 2-8040-0813-4
D/1992/258/112