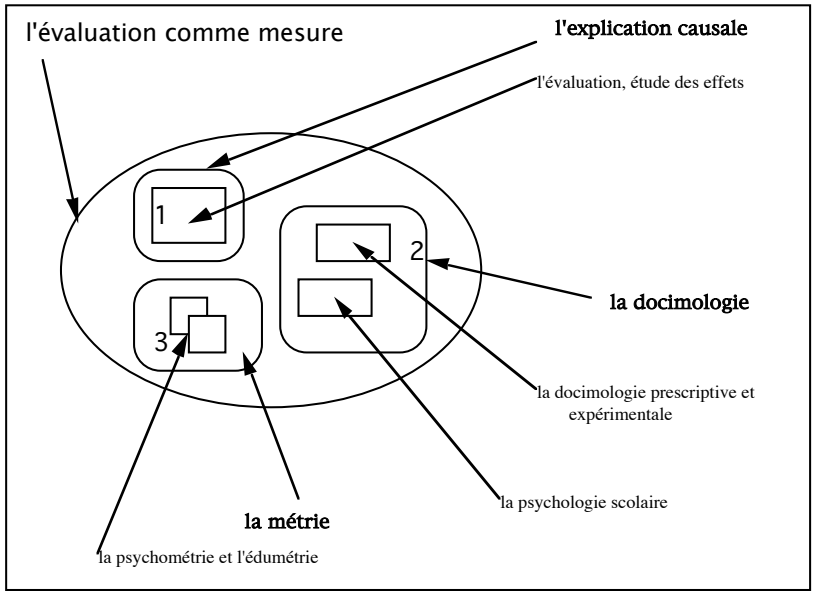
L'évaluation comme mesure

Chapitre I



Dans la conception, le système d'idées ou l'*épistémê de La mesure*

coexistent des modèles, des modes de pensée qui servent de références théoriques.
Ces modèles sont
le modèle nomade de l'explication causale
et les modèles transrégionaux de la docimologie et de la métrie

Des écoles, des mouvements, développent des *syntagmes ,* des *motifs* identifiables, des modèles locaux*:*
dans le modèle 2 :
-la*docimologie prescriptive et expérimentale*
-la *psychologie scolaire -*
dans le modèle 3:
-l'*édumétrie*
-la *psychométrie*

I.
L'évaluation comme mesure, perspective générale

Evaluation est ici synonyme d'opérations de mesure. On situe traditionnellement cette matrice théorique (1) de l'évaluation dans les milieux scolaires mais elle a largement débordé le monde des enseignants et règne sur la formation des adultes, la formation à la recherche et l'univers des entreprises.

On peut dire que Mesurer est le mot qui vient "tout naturellement" à l'esprit quand on parle d'évaluation. C'est sans doute le sens le plus ancien, le plus solidement ancré dans les mentalités, dans l'idéologie.

L'évaluation confondue avec la mesure veut répondre à la question : "Qu'est-ce que ça vaut ? (ce que nous faisons)". Résoudre le problème de la valeur à accorder à tel ou tel acte passe ici par la situation sur une échelle de mesure.(2)

(1) ou "système d'idées" (Morin, Tome 4, chapitre 2 *La méthode*), ou épistémê c'est-à-dire l'ensemble des connaissances réglées (conceptions du monde, sciences, philosophies...) propres à un groupe social, à une époque.

(2) Morissette, D. *La mesure et l'évaluation en enseignement*, Ste-Foy, Presses Université de Laval, 1984

Monteil, J-M. "L'évaluation scolaire : fragments de recherches en psychologie", *Connexions* n°56, 1990-2, p. 51/65 :

(...) *L'évaluation : un objet de recherche en psychologie :*

Objet d'étude en psychologie, l'évaluation présente, à ce titre, la caractéristique d'être appréhendable à plusieurs niveaux : celui des processus, celui des pratiques ou encore celui des techniques et méthodes. Selon le niveau choisi, les référents théoriques, voire les méthodologies sont susceptibles de marquer certaines différences.

Abordée en tant que processus, l'évaluation appartient à la catégorie des objets de recherche relevant des processus de jugement, elle est alors inscrite dans un champ théorique où le modèle de la décision occupe une place privilégiée. La psychologie d'inspiration cognitive constitue ici la référence paradigmatique de base du travail scientifique, quels que soient les points d'application du jugement. En effet, que l'on étudie l'évaluation avec comme support la notation des productions scolaires (Noizet & Caverni, 1978), que l'on aborde les modes d'explication des conduites d'autrui ou des siennes propres comme dans les recherches sur l'attribution causale (Deschamps & Clemence, 1987), que, préoccupé de perception sociale, l'on s'attache à la formation d'impressions, et au jugement des personnes (Srull & Wyer, 1989 ; La Haye, 1989, 1990), ou, encore, que l'on s'efforce de déterminer comment nous utilisons notre connaissance d'autrui pour faire des inférences, aux résultats d'ailleurs souvent faux, à propos de la personnalité de cet autrui, comme dans l'étude des théories implicites de la personnalité (Leyens, 1983 ; Beauvois, 1984) les processus impliqués dans ces différents comportements ressortissent tous à l'étude plus générale des activités mentales que les psychologues se donnent pour objectifs de comprendre et de modéliser. (...)

Appréhendée comme une méthode, l'évaluation participe de l'élaboration d'outils. Aussi, à ce titre, elle emprunte à la psychophysique puisqu'il s'agit, à l'aide

d'instruments de mesure, de mettre en relation des variations issues de l'univers physique et les réponses psychologiques qui en découlent. (...)

Noizet, G. , Caverni, J-P. *Psychologie de l'évaluation scolaire,* Paris, PUF, 1978
Deschamps, J-C. & Clemence, A. *L'explication quotidienne*, Fribourg, Delval, 1987
Srull, T-K. ; Wyer, J.R. "Person memory and judgement", *Psychological Review*, 96, 1, 1989, p 58/83
La Haye, A.M. "La mémoire des personnes I Les fondements mémoriels du jugement" *L'année psychologique,* 1989, p 585/613 et " II La construction cognitive des individus et des groupes", *L'année psychologique*, 1990, p. 93/108
Leyens , J-P. *Sommes-nous tous des psychologues* ?, Bruxelles, Mardaga, 1983
Beauvois, J-L. *Psychologie quotidienne*, Paris, PUF, 1984

Monteil, J-M.

I.1.
Le modèle de l'explication causale ; l'évaluation, étude des effets

Champ d'étude :

Le travail de l'évaluateur consiste à rechercher la cause des effets constatés : l'évaluation est l'étude des effets, par exemple les effets d'une formation sur les formés. On est donc dans le modèle (1) de l'explication causale : l'évaluateur souscrit à l'idée d'une liaison " naturelle ", automatique entre les faits-causes et les faits-conséquences.

Les outils d'évaluation produits par ce modèle sont tous des appareils de mesure, de quantification, de distribution sur une échelle graduée. Evaluer, c'est situer sur une échelle dite " de valeur ", dont le prototype reste la notation de zéro à vingt. Les statistiques s'imposent et tiennent lieu de méthodologie. Autrement dit, la méthode de l'évaluation est confondue avec les procédures d'administration de la preuve, dans une scientificité stricte, empruntée aux Sciences de la nature : déjouer les événements trompeurs, prendre de la distance par les mathématiques, gérer le hasard et généraliser les résultats.

(1) en lien direct avec le "schème causal", cf. Berthelot, J-M. *L'intelligence du social*, PUF, 1990, p.62 : ..." B dépend de A selon une relation telle que, dans l'absolu, c'est-à-dire dans une situation où A serait la cause unique de B, l'on ne puisse avoir B sans A et qu'à toute variation de A corresponde une variation de B (implication réciproque). Il s'en suit que A et B sont distincts soit réellement (objets ou réalités différentes), soit analytiquement (niveaux différents d'une réalité globale) et que l'élément A est conçu comme étant nécessairement antérieur chronologiquement ou logiquement à l'élément B." En somme le schème se donne à un moment de la vie des idées comme évident, il se présente comme la référence obligée, il structure si profondément les modes de pensée qu'il peut, comme ici, ne même pas se nommer comme tel : l'évaluation EST la recherche du rapport causal, on parle l'évaluation sans parler le schème. Le schème joue alors le rôle d'un postulat.

MORIN, M. "Evaluation et éducation des adultes", *Education permanente* n°9, 1971, p. 21/38 :

(...) Dans les années cinquante, c'est pour des actions de formation du personnel d'encadrement dans les entreprises que commencent à apparaître dans les pays anglo-saxons "les premiers compte-rendus d'évaluation systématiques" (2). Encore, ces premiers travaux possèdent-ils un caractère assez impressionniste et se contentent-ils très souvent "d'étudier les effets immédiats de la formation en considérant, par exemple, les modifications des attitudes et du comportement des individus formés à l'issue du programme de formation". En Europe, c'est également à propos du personnel d'encadrement en milieu industriel (de maîtrise en particulier) que s'organise dans les années 60 un projet de recherche sur l'évaluation des résultats de la formation, à l'échelon européen. (...) La préoccupation première est alors une interrogation sur l'efficacité des actions de formation d'agents de maîtrise ou, éventuellement, de cadres moyens d'entreprises industrielles.(...)

C'est à partir des Universités américaines et de leurs laboratoires de sciences sociales appliquées que des réponses méthodologiques ont été construites. Les chercheurs européens se sont contentés "d'adapter" les "instruments" élaborés aux Etats Unis aux divers "publics" nationaux. (...)

L'orientation des travaux a progressivement quitté la référence d'un modèle strict de psychologie de l'apprentissage pour tenter de situer l'évaluation comme moment ou comme pratique dans le cadre d'une théorie des systèmes et de l'équilibre organisationnel. (...)

(2) organisation de Coopération et de développement Economiques. "Les techniques d'évaluation de la formation du personnel d'encadrement". Publications de l'OCDE, juin 1963.

MORIN, M.

CARDINET, J. "L'élargissement de l'évaluation", dans *Hommage à Cardinet*, Fribourg, Delval, 1990, p.109/137, article publié dans *Education et recherche*, vol 1, n°1, 1979, p. 15/34 :

(...) *L'évaluation quantificatrice*

.(...) L'évaluation quantificatrice est appelée ainsi parce qu'elle recueille en priorité des données chiffrées, alors que les données subjectives, les attitudes et les jugements par exemple, ne sont recueillis qu'à titre de reflets indirects de ces efforts objectifs. L'accent est mis sur les résultats, qui doivent pouvoir se mesurer.(...)

Pour l'évaluation quantificatrice la causalité est une relation observée de l'extérieur entre trois ordres de faits. Il existe des variables indépendantes dont on peut étudier l'effet sur des variables dépendantes, à condition de contrôler l'effet de l'ensemble des autres variables. Seront variables indépendantes, celles que l'expérimentateur modifiera à sa guise (en général le curriculum étudié). Seront variables dépendantes, celles qui révéleront l'effet du curriculum (les dimensions d'objectifs pédagogiques). Les autres influences seront contrôlées en partie par des plans expérimentaux ingénieux, mais surtout par l'attribution au hasard des élèves et des maîtres aux diverses conditions et par les tests statistiques ultérieurs.(...)

En principe, l'évaluation quantificatrice est indépendante du temps. Une première raison est le caractère immuable d'un dispositif expérimental. La nature des hypothèses détermine les procédures à suivre pour le recueil des données. On a pu dire que, dans l'idéal (de la méthode hypothético-déductive), il devrait être possible d'écrire un rapport scientifique avant même d'avoir fait la moindre observation. Les hypothèses de travail devraient être suffisamment bien précisées et la déduction à partir de ces hypothèses devrait être assez rigoureuse pour que l'expérience n'ait plus qu'à apporter la confirmation attendue. Il s'ensuit qu'un plan expérimental est rigide.

Il faut donc au chercheur une situation stable, où les choses se passent, jusqu'à la fin des études, de la façon dont le chercheur les avaient prévues.

Une seconde raison tient aux techniques d'analyse. Comme il lui faut l'ensemble des données pour effectuer ses tests statistiques, le chercheur est amené à prendre en charge de préférence des évaluations sommatives. De tels bilans sont essentiellement statiques, parce que définitifs. Ils sont donc extra temporels.(...)

L'évaluation quantificatrice fonde ses généralisations sur la répétabilité des résultats observés. Les tests statistiques ont pour but d'assurer le chercheur de la stabilité de ses conclusions d'un échantillon à un autre. La valeur de ses mesures est assurée par des mises en relation préalables (études de fidélité) qui contrôlent également leur caractère répétable.

La généralisabilité d'une relation observée se fonde dans ce cas sur la même démarche que l'inférence statistique, par laquelle on estime un paramètre d'une distribution à partir d'une série d'observations tirées au hasard dans cette population. Dans la réalité sociale globale où se situent les innovations pédagogiques, les situations ne sont pourtant ni contrôlables, ni répétables, comme le notent Weiss et Rein (1970) en relevant les difficultés théoriques et pratiques que soulève l'approche quantificatrice.(...)

Ce qui caractérise enfin les recherches d'évaluation actuelles, c'est justement l'intérêt pour la méthodologie pour elle-même. Alors que le modèle scientifique de l'évaluation comparative avait été considéré plus ou moins comme acquis et intangible pendant de nombreuses années, il est maintenant remis en cause. Chaque étude de cas sert à expérimenter, en même temps qu'une réforme particulière, une nouvelle façon d'aborder le problème de l'évaluation en général.(...)

Weiss R.S. et Rein , M. The evaluation of board - aim programs : experimental design, its difficulties and an alternative, in *Admin, Science quart,* 197O , 15 (1), p. 97/112

CARDINET, J.

HUTEAU, M., & LOARER, J., "Comment évaluer les méthodes d'éducabilité cognitive ?", *L'orientation scolaire et professionnelle*, n° 1. vol 21, 1992, p. 47/74 :

(...) Si l'évaluation de l'efficacité des méthodes d'éducabilité cognitive est une préoccupation largement partagée par l'ensemble des partenaires de la formation (enseignants, formateurs, psychologues, pédagogues...), ce que l'on entend par évaluation l'est moins. Les auteurs proposent une réflexion sur les objectifs à retenir et présentent les principaux aspects méthodologiques à prendre en compte pour mener à bien une évaluation permettant de valider scientifiquement les méthodes d'éducabilité cognitive. (...)

Dans de nombreux cas, les formateurs procèdent eux-mêmes à cette évaluation qui consiste alors à noter en fin de formation (ou à recueillir auprès des stagiaires eux-mêmes), un certain nombre d'appréciations subjectives sur les changements de comportements observés durant le stage. Ces appréciations portent généralement sur les méthodes de travail mises en oeuvre par les stagiaires, et sont souvent complétées par des témoignages sur les modifications qui interviendraient dans les représentations qu'ont les stagiaires de la formation, de la pratique professionnelle, ou encore d'eux-mêmes et de la possibilité qu'ils ont de développer leurs capacités et leurs connaissances. Les conclusions de ces études sont alors généralement très favorables. Mais quel crédit leur apporter ?

L'évaluation nécessite, d'une part, une analyse préalable des éléments sur lesquels on va faire porter l'évaluation et, d'autre part, la construction d'un dispositif d'observation qui permette d'aboutir à des conclusions valides sur ces observations. Bref, il faut à l'évaluation, des objectifs pertinents et une méthodologie adaptée. (...)

1. Les objectifs de l'évaluation.

1 Sur quoi doit porter l'évaluation ?

C'est une évidence, on ne peut envisager un dispositif d'évaluation qui ne tiendrait pas compte des objectifs poursuivis par la méthode que l'on cherche à évaluer. Evaluer, c'est avant tout vérifier si les objectifs de la formation ont été atteints. Le choix des critères d'évaluation et des indicateurs que l'on va observer dépendent donc nécessairement des objectifs de la formation elle-même. (...)

2. La méthodologie de l'évaluation.

Les objectifs de l'évaluation d'un programme d'éducabilité cognitive étant définis, il reste alors à mettre en place ce programme dans des conditions telles que l'on puisse savoir si les objectifs visés sont effectivement atteints. Ceci pose plusieurs grandes questions de méthode

- Comment observer, caractériser les conduites ?
- Comment montrer que ceux qui ont suivi le programme ont changé davantage que ceux qui ne l'ont pas suivi ? Ou, en d'autres termes, comment organiser le recueil d'observations ?
- Comment s'assurer que ce changement est bien attribuable au programme et non aux paramètres psychosociaux de la situation expérimentale ? (...)

Ces questions ne sont pas spécifiques à l'évaluation des méthodes d'éducabilité cognitive, elles se posent dès que l'on cherche à évaluer les effets d'une intervention visant à modifier les individus quelles que soient la nature ou la finalité de ces interventions (interventions éducatives diverses, interventions thérapeutiques, recherche de l'impact de certaines conditions de travail, effet de politiques...). (...)

Conclusion

(...) l'évaluation des méthodes d'éducabilité cognitive impliquait une démarche contraignante, parfois lourde et relativement complexe. Au niveau de l'observation des conduites, toute une série de biais doivent être contrôlés. Ceci nécessite une définition stricte des conditions de l'observation. Au niveau du recueil des données de nombreuses précautions doivent être prises afin de pouvoir conclure avec un minimum d'ambiguïté et d'éviter des erreurs d'interprétation. L'impossibilité d'expérimenter, au sens fort du terme, est bien sûr un handicap. Une véritable expérimentation supposerait une simplification drastique des procédures mises en oeuvre dans l'éducabilité cognitive. On améliorerait ainsi la validité interne, mais au prix de la quasi-disparition de la validité externe. En d'autres termes, nos résultats seraient plus sûrs mais ils perdraient leur pertinence. C'est donc bien sur le terrain, avec toutes les difficultés que cela comporte, que doivent être évaluées les méthodes d'éducabilité cognitive.

Le type d'évaluation que nous proposons doit être distingué nettement des évaluations pratiquées fréquemment par les enseignants et les formateurs à l'issue ou au cours même de la formation. Celles-ci sont évidemment utiles car elles permettent une régulation, par corrections et ajustements, du processus de formation. Par ailleurs, elles sont une source irremplaçable d'idées et d'hypothèses. Mais elles n'apportent généralement pas d'informations fiables sur la valeur des méthodes. Pour que celle-ci puisse être éprouvée un ensemble de conditions favorables doivent être réunies : la volonté des formateurs d'harmoniser leurs pratiques, leur acceptation de certaines modalités d'observation, la possibilité d'une collaboration formateurs-chercheurs. (...)

HUTEAU, M., & LOARER, J.,

LE POULTIER, F. "Principes d'une recherche évaluative en travail social", *Connexions* n°56, 1990, p.37/49 :

(...) *Etablir une relation de causalité entre des actions et des effets observés :*

(..) Evaluer dans un esprit expérimental conduit à formaliser cette relation de causalité entre une action d'éducation spécialisée ou d'assistance sociale et des modifications observées chez des personnes prises en charge ou suivies. Il n'est pas dans les us et coutumes des travailleurs sociaux d'exprimer les choses de cette manière. Le problème de la mise en relation des résultats observés avec les moyens déployés n'est jamais posé aussi explicitement. Cela ne signifie pas que les travailleurs sociaux vivent dans une complète indifférence quant aux effets de leurs pratiques, mais cela veut dire que le rapport moyens/résultats n'est jamais posé comme tel. En règle générale, les moyens mis en oeuvre ne sont pas inventoriés de matière systématique mais évoqués a posteriori lors d'un réunion ou de la rédaction d'un bilan. Les résultats sont également rapportés de manière occasionnelle pour une personne donnée. La mise en relation de causes et d'effets s'apparente davantage à une juxtaposition de deux observations concernant une personne singulière plutôt qu'à une vérification objective d'un lien de causalité entre des actions menées auprès d'un ensemble de personnes et l'évolution des tendances générales de ces dernières. Mais, parce que des données de contexte politique et économique ont changé, le travail social est confronté à une obligation de résultats. (...)

Ceci est la marque d'un changement considérable dans les mentalités, une véritable petite révolution. Pourtant, le raisonnement expérimental appliqué à l'évaluation dans le travail social repose sur des considérations somme toute élémentaires : il importe de montrer que des personnes suivies ou placées arrivent dans un état A et repartent quelques mois ou années après dans un état B sensiblement meilleur, et il faut pouvoir attribuer objectivement le passage de A à B à l'action des travailleurs sociaux. Caricatural, simpliste, rétorquent certains praticiens : chaque personne est un cas particulier, les causes et les effets sont trop étroitement liés pour être dissociés, les symptômes d'une amélioration sont des apparences parfois trompeuses. Cette position est surtout tenue par des psychologues cliniciens et des psychiatres. Il n'est pas étonnant qu'une analyse en terme de causalité et qu'un traitement collectif des usagers du travail social suscitent des résistances chez ceux qui tirent leur légitimité de la complexité et de l'individualité. (...)

LE POULTIER, F

I,1. Le modèle de l'explication causale, l'évaluation, étude des effets,

point de vue des détracteurs

Si ce type d'évaluation n'a plus satisfait les évaluateurs, c'est qu'on lui a reproché de cautionner une conception mécaniste du monde et de s'inscrire dans l'idéologie positiviste, mais surtout d'avoir tendance à ériger en dogme l'idée de la mono-causalité linéaire : la causalité n'est plus l'explication suffisante d'un phénomène. Comprendre n'est plus chercher la cause. La constatation que, dans les situations de vie et les pratiques sociales, l'explication peut être pluri-causale et non-linéaire a atteint la pérennité de ce modèle et a ouvert à la nécessité d'autres évaluations.

CARDINET, J., "Evaluation interne, externe ou négociée?", conférence de 1987, *Hommage à Jean Cardinet*, Fribourg, Delval, 1990, p.139/156 :

(...) *Le modèle des sciences de la nature.*

L'évaluation va donc avoir de préférence pour objet des produits finis, dont le caractère stable favorise la mesure. Comme ce sont les résultats qui comptent, et non les intentions, ou les raisons, qui servent trop souvent de justifications fallacieuses, les observateurs se contenteront de décrire les niveaux atteints du point de vue de quelques variables critères, en n'introduisant qu'un minimum d'interprétation personnelle.

L'évaluation reste donc essentiellement comparative, se réduisant à un contrôle du résultat observé par rapport à la mesure attendue (une réalité multidimensionnelle étant ramenée à une simple moyenne par le jeu d'un vecteur de coefficients de pondération). Selon l'écart de la moyenne observée au seuil prévu, les décisions correctives peuvent être prises quasi-automatiquement, dès qu'une stratégie a pu être définie à ce sujet. (...)

Pour assurer l'objectivité de cette évaluation, une série de précautions sont proposées par les chercheurs. La principale source d'arbitraire provenant du choix des questions d'examen, un tirage aléatoire est organisé dans une banque d'items représentant l'univers de toutes les questions admissibles (Schoemaker, 1973). L'analyse statistique de la variabilité entre items permet ensuite de déterminer la taille de l'échantillon d'observations à prendre, en fonction de la marge d'erreur que l'on peut accepter autour du seuil de réussite (Cardinet et Tourneur, 1985). Des procédures formalisées sont aussi prévues pour la fixation de ce seuil. En cas d'échec, les activités à proposer à chaque élève en difficulté sont choisies d'avance, selon la nature de ses difficultés (Bloom, Hastings et Madaus, 1971). Rien d'étonnant dans ces conditions à ce que l'ensemble de la démarche individuelle d'évaluation par objectif puisse être confiée à un ordinateur (Nitko et Tse-chi-Hsu, 1984). (...)

Discussion de l'évaluation externe :

Personne ne songe à mettre en doute la fécondité de la démarche expérimentale, ni les succès de son application au domaine de l'enseignement. Ce que l'on peut récuser, par contre, c'est le positivisme, c'est-à-dire l'affirmation que cette démarche est la seule voie légitime pour parvenir à la connaissance. Il est facile de montrer que cet éclairage particulier détermine, lui aussi, ses propres zones d'ombre.

- Illusion de simplicité :

A chacun des niveaux examinés, ce serait une erreur de croire que les mesures que l'on a prises rendent comptent de l'essentiel de ce que l'on voulait contrôler. Le fait qu'un élève soit capable de répondre à des questions d'examens, par exemple, n'a qu'une relation assez faible avec d'autres critères, tout aussi essentiels, comme le fait qu'il sache faire appel spontanément aux mêmes connaissances dans une situation extra-scolaire.

A un autre niveau, il ne suffit pas de reconnaître que certains objectifs d'un curriculum ne sont pas atteints : les tenants de l'Ecole Active, par exemple, ont toujours soutenu qu'il fallait aussi mesurer l'effet indirect de la démarche d'étude qu'ils avaient choisie sur les motivations et les stratégies d'apprentissage des élèves, c'est-à-dire qu'il fallait évaluer les bénéfices à longue échéance, avant de faire un bilan. Comment tenir compte à la fois de critères à court et à long terme, s'ils sont contradictoires ? L'évidente complexité des problèmes se double pourtant d'obstacles encore plus fondamentaux.

- Illusion de généralité :

Il n'est plus guère de chercheurs dans les sciences sociales qui prétendent mettre en évidence des lois valables dans toutes les cultures. Le temps oblige aussi à relativiser les théories, dont la durée de vie ne dépasse guère celle d'une génération. Les techniques évoluent encore plus rapidement. La constitution de banques d'items, par exemple, se heurte au fait que les contenus et les formes de questionnement se renouvellent au même rythme que les programmes et les méthodes d'enseignement. Enfin, les recherches expérimentales mettent en cause elles-mêmes la validité de leurs conclusions, en révélant qu'une approche didactique favorable pour tel type d'élève ne l'est pas pour tel autre et que ces interactions se multiplient à l'infini, pour chaque nouvelle variable étudiée (Cronbach, 1975).

- Illusion d'objectivité :

L'idée même qu'il existe dans la réalité un niveau vrai de connaissances de l'élève, ou une opinion véritable de la personne interrogée est maintenant mise en doute, lorsqu'on voit qu'il suffit de changer la forme du questionnement pour transformer la réussite en échec, ou l'acceptation en refus (Cardinet, 1986) . On peut, bien sûr, esquiver le problème en fixant la forme du questionnement (par des tests standardisés, ou des questionnaires écrits), mais on sacrifie alors presque totalement la portée (généralisabilité) du résultat. Enfin, la réactivité de la plupart des mesures utilisées dans les expériences pédagogiques oblige à douter de l'objectivité de leurs résultats : un groupe de contrôle n'est jamais neutre et le simple fait d'avoir été choisies pour faire partie d'une expérience modifie positivement l'attitude des personnes concernées. Inversement, devoir simplement appliquer ce que d'autres ont précédemment expérimenté et mis au point est généralement mal vécu. Le modèle d'un produit industriel développé d'abord en laboratoire, puis diffusé à l'extérieur, ne s'applique pas en éducation.

- Nécessité de "comprendre" :

La majorité des chercheurs, dans toutes les sciences sociales, sont maintenant d'avis qu'il est dangereux de faire confiance à une régularité statistique purement empirique, si l'on ne peut pas rendre compte de ce phénomène émergeant, en termes de réactions de sujets à leur environnement, si l'on ne peut pas en faire, comme le disent Boudon et Bourricaud (1986) "un composé d'actions compréhensibles". Il est donc impossible de gérer de l'extérieur n'importe quel niveau du système scolaire, en se référant seulement à des indicateurs objectifs. Il faut pouvoir connaître cette réalité de l'intérieur. (...)

Schoemaker, D. *Principles and procedures of multiple matrix sampling*, Cambridge, Mass : Ballinger Publisching Co, 1973

Cardinet, J & Tourneur, Y. *Assurer la mesure,* Berne : Peter Lang, 1985

Bloom, B. , Hasting, T. & Madaus, G. *Handbook on formative and summative evaluation of student learning*, New York : Mc Graw Hill, 1971

Nitko, A. & Tse-chi Hsu, A comprehensive microcomputeur system for classroom testing, *Journal of Educational Measurement*, vol 21, N°4, 1984, p. 377/390

Cronbach, L. Beyond the two disciplines of scientific psychology, *American Psychologist*, vol 30, 1975, p. 116/127

Cardinet, J. Les modèles de l'évaluation scolaire, in *Evaluation scolaire et pratique*, Bruxelles, De Boeck, 1986, p. 239/262

Boudon, R. & Bourricaud, F. *Dictionnaire critique de la sociologie*, Paris, PUF, 1986, 2° ed.

I 1. Le modèle de l'explication causale, l'évaluation, étude des effets, figure de l'évaluateur

La formation évaluée, par exemple -ou tout autre pratique sur laquelle porte l'évaluation-, n'a d'intérêt, dans ce modèle, que parce qu'elle produit des effets. L'évaluateur ne se distingue pas du commanditaire de l'évaluation pour qui la formation, par exemple, est pensée comme une transformation, une production inscrite dans un processus de rentabilité, d'efficacité. L'évalué devient une chose, le résultat d'un ensemble de pressions : la mesure, comme Midas, transforme tout ce qu'elle touche en or.

Remettre en question la double assimilation de la chaîne industrielle de production avec la formation, ainsi que du produit manufacturé avec le formé, a pu faire apparaître que ce type d'évaluateur ne tenait qu'un des rôles possibles : l'expertise de la productivité. Il participe d'une "technologie sociale" : l'évaluateur, délibérément externe, est celui qui rationalise l'expérience de l'Autre.

L'évaluateur dans le modèle de l'explication causale est une doublure du formateur qui se fantasme comme transformateur de l'autre ou du commanditaire qui fantasme la formation comme une procédure de transformation, de fabrication. L'évaluateur partage, sans l'interroger, une conception "économiste" de la pratique évaluée.

DUPUY, J-M. "Guide du bon usage des indicateurs d'évaluation, l'exemple des politiques de développement social des quartiers", *Pour* n°107, *L'évaluation au pouvoir*, p.29/35 :

(...) Parler des effets d'une politique revient à poser deux hypothèses : premièrement, on suppose qu'il existe bien une relation de cause à effet, entre les moyens mis en oeuvre et le phénomène (par exemple, l'échec scolaire ou le nombre des impayés), sur lequel on cherche à mesurer l'impact d'une politique ; or, la causalité ne se démontre pas mais se construit à travers un modèle théorique qui décrit les canaux de cheminement de l'influence d'un facteur sur une variable (par exemple l'influence de l'effectif réduit d'une classe de Zone d'Education Prioritaire sur l'échec scolaire) ; deuxièmement, on suppose que l'effet en question résulte bien de la politique examinée, et d'elle seule.

Idéalement, il faudrait pouvoir déterminer ce qui se serait passé en l'absence de cette politique. Comme c'est impossible, il convient de limiter au maximum le risque de biais dans l'appréciation des effets. Il faut bien sûr commencer par expliciter les mécanismes par lesquels se produisent les influences supposées : par exemple quels sont les facteurs déterminants de la délinquance et pourquoi ? Idem pour le chômage des jeunes, ou pour la multiplication des familles monoparentales.

Puis on cherche des groupes de contrôle, en l'occurrence des quartiers, les plus semblables possibles aux quartiers analysés, sauf du point de vue de la politique étudiée (ce qui ne sera pas facile si la sélection des quartiers pour l'opération Développement Social des Quartiers a été faite de façon assez objective, en retenant effectivement les quartiers les plus défavorisés).

Faute de respecter cette précaution méthodologique, on risque fort d'attribuer à l'opération Développement Social des Quartiers des conséquences qui ne lui sont pas dues. En se bornant à enregistrer les modifications survenues dans le quartier depuis 1981, on mêlerait aux effets propres de cette politique, l'impact d'autres phénomènes concomitants (aggravation du chômage dans l'agglomération, lancement d'autres opérations, etc.).

La difficulté à isoler les effets attribuables au Développement Social des Quartiers provient justement de ce que, dans les mêmes quartiers, se superposent et interfèrent de multiples politiques, émanant de l'Etat et des collectivités locales (mission locale, Zone d'Education Prioritaire, opération "été chaud", conseils de prévention de la délinquance, etc.).

Aussi, le groupe de travail a-t-il recommandé la prudence, d'abord dans la méthodologie en procédant à des comparaisons avec d'autres quartiers ou groupes de contrôle, ensuite dans l'interprétation des liens de cause à effet.

Une des difficultés de l'identification des effets tient à leur variété, certains se produisant par exemple là où on ne les attendait pas. Ces effets inattendus sont parfois qualifiés de "pervers" alors qu'ils peuvent être fort souhaitables bien qu'imprévus. **DUPUY, J-M.**

I. 2
La docimologie comme modèle de l'évaluation

Champ d'étude :

La docimologie (1) s'intéresse aux examens, elle repose sur un principe non apparent de causalité : le correcteur de copies d'examens est surdéterminé par un ensemble de facteurs que l'on veut mettre à jour, il est agi. On peut y voir une actualisation d'un thème déterministe présent dans le schème causal (que Berthelot appelle "la causalité structurelle") : "... un système B est sous la dépendance d'un système A, antérieur à lui, ou le plus souvent, plus fondamental que lui. (...) La correspondance structurale entre systèmes sera ainsi la preuve d'une relation de détermination, que l'antériorité logique de l'un sur l'autre permettra de définir comme causale. (...) ce schème est donc de type vertical. Il tend à chercher, derrière une construction sociale B (...) le système A qui le fonde et dont B est l'effet ou le "reflet". (p.64).

(1) la métrie reprendra le même principe (cf. I.3)

PARISOT, J-C., "Le paradigme docimologique : un frein aux recherches sur l'évaluation pédagogique?", *L'évaluation en question*, CEPEC, 2°ed, Paris, ESF, 1988 :

(...) Dans le vocabulaire de la psychologie de Piéron, la docimologie est définie comme "l'étude systématique des examens (modes de notation, variabilité interindividuelle et intra-individuelle des examinateurs, facteurs subjectifs, etc.)". Ce terme proposé par Pieron lui-même est dérivé du grec dokimé qui veut dire épreuve, la docimastique (dokimastikos) désignant l'étude des techniques d'examens. Dans son

ouvrage : Examens et docimologie, il rapporte la naissance de ce champ scientifique. (...)

Gilbert De Landsheere reprend dans ses définitions des éléments indiqués par Pieron tout en commentant : "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.". Noizet et Bonniol, pour nommer l'aspect prescriptif éventuel de la docimologie, proposent le terme de docinomie et J. Guillaumin, celui de doxologie pour qualifier l'étude systématique du rôle que l'évaluation joue dans l'éducation scolaire (référence psychologique ou psychosociologique). Ces termes n'ont pas été retenus par l'usage qui seul a consacré le terme de docimologie. (...)

Piéron, H. *Vocabulaire de la psychologie*, 6° éd revue et augmentée, Paris, PUF, 1979
Piéron, H. *Examens et docimologie*, Paris, PUF, 1963

PARISOT, J-C.

CARDINET, J. "Remettre le quantitatif à sa place en évaluation scolaire", *Les nouvelles formes de la recherche en éducation,* colloque AFIRSE d'Alençon, Matrice Andsha, 1990, p. 58/66 :

(...) *Le modèle suivi : la quantification en psychologie*

Pendant longtemps la psychologie est apparue comme la discipline de base dont la pédagogie devait s'inspirer : elle avait réussi en effet à établir des lois générales, dans le domaine de l'apprentissage notamment, dont on pensait pouvoir tirer directement des applications dans le milieu scolaire.

Les conditions de ce progrès semblaient claires à la fin du siècle dernier, parce que les sciences de la nature traçaient la voie du point de vue méthodologique. Pour tirer profit de la méthode expérimentale, érigée par Claude Bernard (1856) en véritable épistémologie générale, il fallait pouvoir mesurer avec précision les phénomènes que l'on étudiait. D'où l'obsession de la quantification chez Fechner, Wundt et les premiers psychologues, qui affirmaient en postulat : "Tout ce qui est existe dans une certaine quantité que l'on peut donc mesurer".(...)

CARDINET, J.

I 2. La docimologie comme modèle de l'évaluation

I. 2.1.
La docimologie prescriptive et expérimentale : la docinomie

Champ d'étude de ce premier modèle local (1) :

L'objet travaillé est ici la fabrication des notes.
On étudie les examens pour identifier les biais, les effets perturbateurs qui expliqueraient les variations entre les notations. Il s'agit de trouver des lois qui rendraient compte des problèmes de la "fidélité" des notes (aboutir au même

résultat quel que soit le nombre de passations ou de correcteurs), de la "validité" des examens (n'évaluer que ce qui est affiché) et de la "sensibilité" des outils d'évaluation (utiliser harmonieusement les échelles de mesure).

(1) Les modèles locaux ou syntagmes sont souvent concurrentiels, Ainsi docinomie et doxologie se construisent l'un contre l'autre, la doxologie se voulant un "dépassement" de la docinomie.

CARDINET, J., *Les modèles de l'évaluation scolaire*, Neuchâtel, IRDP, 1986 :

(...) Le défaut d'objectivité des examens et des notes scolaires a été mis en évidence de façon répétée dans tous les pays.

Piéron (1963) a par exemple comparé les moyennes des notes données par des jurys parallèles du baccalauréat français. Les différences atteignaient 4 à 5 points sur 20, en mathématique et en physique, faisant passer les taux de réussite de 53% à 31% d'un jury à l'autre, pour une même population de candidats.

Laugier et Weinberg (1938), donnant de façon expérimentale le même lot de copies à corriger à des jurys parallèles, trouvent des différences moyennes de 2 à 3 points en mathématique et en physique, et de 4 points en français et en philosophie. En prenant les deux extrêmes de la distribution des notes données à une même dissertation par 76 correcteurs différents, ces chercheurs ont même mis en évidence un écart de 13 points sur 20 entre deux examinateurs .

Les travaux plus récents de Noizet et Caverni (1978), puis ceux de Bonniol (1981), sont parvenus à analyser expérimentalement les déterminants psychologiques de ces différences d'appréciation, qui tiennent au choix des compétences à évaluer, à l'estimation du seuil de suffisance, aux attentes induites par des informations extérieures sur les élèves, aux effets de contraste dus à l'ordre de succession des copies, etc. (...)

Piéron, H. *Examens et docimologie,* Paris PUF, 1963
Laugier, H. & Weinberg, D. *Recherche sur la solidarité et l'interdépendance des aptitudes intellectuelles d'après les notes des examens écrits du baccalauréat*, Paris, Chantenay; 1938
Noizet, G. & Caverni, J-P., *Psychologie de l'évaluation scolaire*, Paris, PUF, 1978
Bonniol, J-J. "Déterminants et mécanismes des comportements d'évaluation d'épreuves scolaires", Thèse, Université de Bordeaux II, 1981

CARDINET, J.

Est mise à jour une série de variables influençant les notations, appelées “effets” : effets d'ordre et de contraste dans la succession des copies, effet de contamination de l'avis des confrères, effet de stéréotypie systématisant les appréciations antérieures, effet de halo des représentations sociales du candidat chez le correcteur.

ABERNOT, Y. *Les méthodes d'évaluation scolaire*, BORDAS, PARIS, 1988 :

(...) *Effets d'ordre et de contraste :*

(...) JJ. Bonniol présenta en 1972 une thèse sur l'estimation par contraste. Il proposa à des groupes d'enseignants de corriger des séries de copies identiques, mais dans des ordres différents. En répétant cette expérience un nombre de fois suffisant, il constata deux phénomènes :

-les correcteurs notent par contraste, c'est-à-dire que la note d'une copie dépend en partie de la ou des copies précédentes et leur est en quelque sorte opposée

-d'une manière générale, les correcteurs sont plus sévères à la fin de la série qu'au début.(...)

A partir de cet effet de contraste, JJ. Bonniol utilise la notion d'ancre (référence à l'instrument de marine). Dans une série de copies sélectionnées après plusieurs corrections, pour leurs notes moyennes, le chercheur introduit une ou plusieurs très bonnes (ancres hautes) ou très mauvaises copies (ancres basses) pour constater leurs effets sur les copies suivantes.

Il montre, par exemple, qu'en anglais, l'effet de contraste est important surtout après les ancres basses ; alors qu'en mathématiques, ce sont davantage les ancres hautes et lourdes (plusieurs copies excellentes) qui entraînent la sous-estimation. (...)

Effets de contamination

L'avis des confrères a également un rôle influent sur le jugement du correcteur.

JP. Caverni (1975) propose à deux groupes d'enseignants de sciences naturelles de noter quatre copies différentes. Il donne à chaque correcteur un faux dossier scolaire par copie : le premier groupe de correcteurs dispose d'appréciations antécédentes élevées et stables, le deuxième groupe se fonde sur des dossiers aux notes faibles et dispersées.(...)

Les résultats font apparaître le peu d'écart absolu (0,25/20) dû à l'influence d'un dossier pour une copie notoirement faible (de 2,75/20 à 3,00). Mais l'écart grandit au fur et à mesure que la qualité des copies augmente. Il culmine à 3,25 sous l'influence de dossiers opposés pour des copies de haut niveau. En conclusion, un dossier ne rachète pas une mauvaise copie, alors qu'il est très influent pour un bon travail.(...)

Effets de stéréotypie :

(...) Après quelques mois, l'enseignant connaît suffisamment les élèves pour avoir une idée de leurs compétences actuelles. (...) L'effet de stéréotypie est une systématisation de l'appréciation établie. (...)

Effet de halo :

(...) M. Gilly (1980) consacre plusieurs chapitres de son ouvrage aux représentations respectives des élèves et des maîtres, émanant de stéréotypes sociaux de natures très diverses : habillement, verbalisation, attitudes face à l'école, etc. Ces variables sont très importantes dans la relation pédagogique, et se retrouvent dans l'évaluation où l'effet de halo est à l'oeuvre.(...)

Il convient de réfléchir à la pertinence d'une note censée être l'aboutissement d'une mesure, tout particulièrement dans les domaines scolaires qui ne se laissent pas quantifier aisément ; (...) Les notes sont relatives, non seulement au groupe de référence, et à l'établissement scolaire, mais aussi à l'enseignant qui les distribue. Ceci n'a rien de scandaleux, mais il est indispensable d'en avoir conscience pour tirer le meilleur parti de l'évaluation.(...)

Bonniol, J-J. "Les comportements d'estimation d'une tâche d'évaluation d'épreuve scolaire, étude de quelques uns de leurs déterminants", Thèse de 3° cycle, Aix-en-Provence, Université de Provence, 1972.

Caverni, J-P., Fabre, J-M., Noizet, G. "Dépendance des évaluation scolaires par rapport à des évaluations antérieures, études en situation simulée", *Le travail humain*, 38, 1975

Gilly, M. *Maître-élève, rôles institutionnels et représentations*, Paris, PUF, 1980

ABERNOT, Y.

Afin de "pondérer" ces biais, (ces “effets parasites”), pour réduire les variations de notes d’un correcteur à l’autre, des procédures ont été proposées, comme la standardisation des

questions posées de façon à mieux cibler l'objet à évaluer dans la réponse fournie par le candidat ou "l'appel à des données d'information complémentaires permettant de "rectifier" les jugements produits : c'est le cas, par exemple lors d'examens relativement importants du recours aux dossiers scolaires ou aux différentes techniques de "repêchage".(1) ; ou "la mise en place de situations permettant une harmonisation empirique des critères utilisés : c'est le sens notamment de la constitution de jurys ou de réunions d'entente entre correcteurs " (1) et "l'intervention a posteriori sur les jugements produits ou les notes attribuées : c'est le sens général des "procédures de modération". " (1)

(1) cf Barbier, J-M. *L'évaluation en formation,* PUF, 1985, p. 43

I. 2. La docimologie
I 2 1. La docimologie prescriptive et expérimentale,

point de vue des détracteurs :

La notion de biais, d'effet perturbateur ou parasite suppose qu'on cherche à expliquer l'écart entre la note donnée et la note méritée. Autrement dit, travaillant sur les déformations, on cautionne l'idée d'une note objective, même si c'est pour montrer que la note ne l'est pas. Le dysfonctionnement, la prise en défaut n'ont de sens que par rapport à la norme.

C'est cette obsession de la "vérité docimologique" qu'accusent les détracteurs : la vraie note, la note juste, n'a plus grand intérêt quand on a compris qu'on n'arrivera jamais à neutraliser tous les biais possibles, que l'objectivité n'existe pas ; que ni les situations d'évaluation ne seront purifiées, ni les outils d'évaluation ne seront parfaits, ni l'évaluateur ne sera libéré de se condition humaine.

CARDINET, J. «L'objectivité de l'évaluation", *Formation & technologie*s, n°0, 1992, p. 17 /25 :

L'objectivité en question.

La critique des examens n'est pas nouvelle. Déjà la recherche de Laugier et Weinberg, publiée en 1938, mettait en cause la fidélité des correcteurs du baccalauréat et les recherches en docimologie de Piéron (1969) sont venues confirmer les doutes concernant l'objectivité de telles épreuves. Mais une série d'approches nouvelles, surtout sociologiques et psychosociologiques, ont pu récemment éclairer le problème sous des angles différents, dont il est important de tenir compte.

De Landsheere (1979) définit l'objectivité comme "le caractère de ce qui donne une image non déformée... des choses". L'objectivité d'un test est pour cet auteur "le fait qu'il est relativement exempt d'erreurs de jugement ou de correction...que ses résultats dépendent seulement de la performance du sujet. "

Si, comme le dit encore De Landsheere, "le garant de cette objectivité réside dans l'emploi de règles de correction et de notation précises», on peut se demander pourquoi, après un demi-siècle de recherches, le système scolaire n'est pas parvenu à assurer cette objectivité dans le jugement des performances des élèves. Formuler ainsi la question est tendancieux, répondra-t-on, parce que c'est jouer sur deux sens

différents du mot "objectivité", pris chez De Landsheere comme une des qualités métrologiques d'un test et non pas comme valeur de vérité d'un jugement. Mais ces deux sens sont adjacents et la confusion n'est pas fortuite. Malgré ce qu'en dit un certain discours officiel, l'évaluation scolaire a beaucoup de raisons de rester subjective (...)

Laugier, H. & Weinberg, D. *Recherche sur la solidarité et l'interdépendance des aptitudes intellectuelles d'après les notes des examens écrits du baccalauréat*, Paris, Chantenay; 1938
Piéron, H. *Examen et docimologie*, Paris PUF, 1969
De Landsheere, G. *Dictionnaire de l'évaluation et de la recherche en éducation*, Paris PUF, 1979
CARDINET, J.

Bonniol (1981, p. 487), quant à lui, oriente la question de l'évaluation vers une autre voie en posant la distinction entre la recherche-évaluation qui "se demande comment évaluer" et la recherche sur l'évaluation qui, elle "se demande comment fonctionne le système qui évalue...". Car la docimologie s'est davantage préoccupée de "méthodologies d'exécution" que de "méthodologies de recherche" (p. 30) :

BONNIOL, J-J., *Déterminants et mécanismes des comportements d'évaluation d'épreuves scolaires*», Thèse de Doctorat ès Lettres et Sciences Humaines, Bordeaux, 1981 :

(...) Le problème (jusqu'à présent traité) est bien "comment évaluer ?" et non pas "que fait celui qui évalue ?". Les populations étudiées sont "les formés", non "les évaluateurs". Or, il nous semble qu'il y a là pour le moins une ambiguïté, une confusion des rôles et des questions, dont le rapport ambigu entre chercheurs et praticiens est un symptôme, mais un symptôme qui ne rend pas suffisamment compte de la gravité de la situation : il semble que l'on cherche à construire une science appliquée sans s'être posé la question des fondements théoriques de cette science ; ainsi on étudie l'évaluation :

-qui est d'abord un comportement, sans s'intéresser à ceux qui manifestent ce comportement ;

-qui est ensuite une tâche, sans s'intéresser à ceux qui effectuent cette tâche

- qui correspond à un rôle et implique un statut, sans se poser le problème de ce rôle et de ce statut ;

-qui pose des problèmes à l'évaluateur, sans s'intéresser à l'évaluateur, mais en prenant les problèmes à son propre compte pour tâcher de les résoudre à la place de l'évaluateur.

Bref, on n'étudie pas l'évaluation, on innove en matière d'évaluation, ce qui empêche que le comportement d'évaluation soit jamais étudié en tant que tel, tandis que par ailleurs on continue d'étudier pour lui-même le comportement de "prise de décision" ou celui de "résolution de problèmes", sans confondre l'expérimentateur et le sujet, mais en distinguant au contraire le problème de l'un et celui de l'autre. (...)

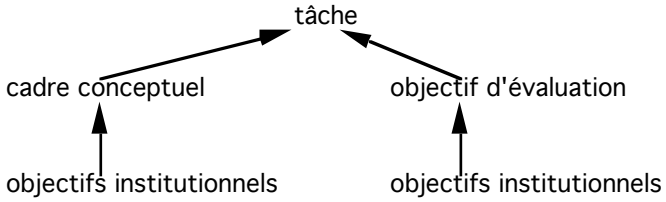
1.2.3.2. Les fondements du modèle

- Il semble que toute évaluation porte sur une production, sur un comportement qui est une réponse déterminée par un problème, par une tâche : il y a

une consigne, un énoncé, qui marque de manière plus ou moins explicite ce qui est demandé. Il ne semble pas qu'il puisse y avoir d'exception à cette règle, quel que soit l'objet évalué.

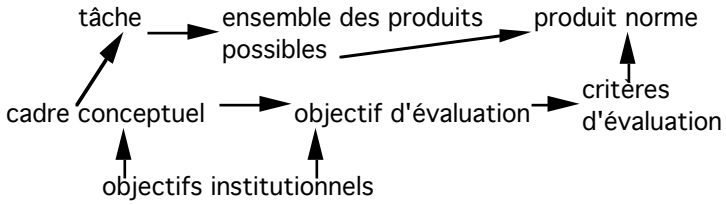
- Cette tâche est généralement définie à partir d'objectifs d'évaluation qui spécifient les objectifs institutionnels visés, qu'il s'agisse d'objectifs d'évaluation formative, subordonnés à des objectifs de formation (quand l'institution demande une transformation des comportements) ou qu'il s'agisse d'objectifs d'évaluation sommative, de sélection ou de certification (quand l'institution demande une garantie de compétence). En apparence la tâche est parfois définie indépendamment de tout objectif d'évaluation ; on serait tenté de le croire en particulier quand l'initiative en est prise par le sujet qui doit l'accomplir : formateur, enseignant en formation ou élève - Mais une telle interprétation impliquerait qu'une tâche peut être décidée indépendamment du réseau des déterminants sociaux, indépendamment des objectifs institutionnels explicitement ou implicitement pris en compte - Or ce n'est pas soutenable.

Il semble donc justifié de poser la relation suivante qui unit les premiers rouages de l'évaluation :



- Une tâche, ou du moins la représentation que l'évaluateur peut s'en faire, conduit à se représenter abstraitement une infinité de réalisations possibles de cette tâche, selon tout un ensemble de paramètres (le temps imparti, les moyens disponibles, la mobilisation des producteurs, etc...). Il y a théoriquement un ensemble de produits possibles dont le produit réel sera un élément. L'évaluateur, avant même d'être confronté au produit réel va restreindre cet ensemble de produits possibles pour le constituer en modèle de référence opérationnel auquel sera comparé le produit réel.

- Cette restriction de l'ensemble des produits possibles s'effectue d'abord en fonction des critères d'évaluation, eux-mêmes dérivés des objectifs d'évaluation : il s'agit des dimensions, des caractéristiques, qui seront privilégiées lors de l'évaluation, et qui définissent le produit-norme dans l'ensemble des produits possibles. Cela permet d'ajouter au schéma de nouvelles relations :



- La filiation entre le produit-norme et les objectifs institutionnels est donc double : par l'intermédiaire des objectifs d'évaluation, plus ou moins ambigus généralement, la tâche (plus ou moins adaptée aux objectifs) et les critères (plus ou moins explicités, plus ou moins pertinents aux objectifs et plus ou moins différents entre deux moments de la formation) constituent la matrice du modèle de référence de l'évaluation, le produit-norme : et il semble bien qu'il y ait toujours un produit-norme

quand il y a évaluation, quel que soit le caractère vague et flou qu'il possède dans les représentations de l'évaluateur, quel que soit le degré de conscience que l'évaluateur peut en avoir.(...)

Néanmoins le produit-norme ainsi défini reste théorique et abstrait. C'est pourquoi l'évaluateur opère une nouvelle restriction dans le sens du réalisme, en limitant ses attentes, grâce à des informations d'une autre nature dont il dispose déjà ou qu'il essaie d'obtenir : informations sur les agents, les producteurs qui effectuent la tâche, informations sur les conditions de réalisation de cette tâche.

L'application sur le produit-norme de ces informations a priori concernant les producteurs, détermine le produit attendu. Les recherches rapportées montrent l'influence relative de ces informations : leur présence ou leur absence conditionne des produits attendus différents en fonction desquels sont effectuées des évaluations différentes. C'est principalement l'influence d'informations concernant les personnes des producteurs, leur statut social, ethnique, scolaire, qui produit un effet d'assimilation des évaluations, tandis qu'un effet dans ce sens d'informations sur les évaluations antérieures reste tout à fait hypothétique. Il faut remarquer qu'en situation simulée les informations du premier type déterminaient des effets nets, mais que les hypothèses n'ont pas été éprouvées en situation réelle, tandis que les informations du second type déterminaient, en situation simulée, des effets moins nets qui n'ont pas été retrouvés en situation réelle d'évaluation sommative.

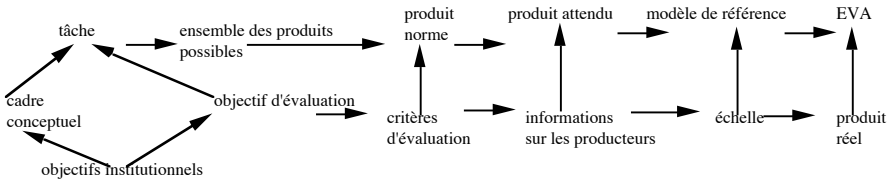
On peut penser que des critères explicites rigoureusement utilisés limiteraient l'influence d'informations relatives aux personnes des producteurs, qui répondent à des normes sociales globales parmi les plus discutables : seule la nécessité dans laquelle se trouve l'évaluateur de relativiser le produit-norme permet d'expliquer le phénomène de prise en compte de ces informations, d'autant qu'elles constituent un ensemble hétérogène et que certaines d'entre elles sont légitimement souhaitables pour que l'évaluateur puisse fixer son niveau d'exigence : l'expérience antérieure des candidats, leur niveau dans la discipline, permettent a priori de définir le seuil d'acceptabilité impliqué dans toute épreuve critériée selon la définition de Harris et Steward (1971) : "échantillon de tâches défini en fonction d'un ensemble bien défini de performances attendues, échantillon utilisable pour estimer la proportion de performance que l'étudiant peut réaliser". Hambleton et Nowick (1973) précisent la définition en termes de seuil : "épreuve qui permet de déterminer si le niveau réel de performances est au-dessus ou au-dessous d'un seuil".

On peut alors poser l'hypothèse que la recherche et la prise en considération des informations a priori concernant les personnes est substituée à celle d'informations techniques que l'évaluateur ne sait peut-être pas prendre en compte, et qui spécifieraient l'ensemble des produits attendus avec autant d'efficacité : il suffirait que le produit-norme idéal soit relativisé par une norme empirique qui peut être soit statistique, soit réglementaire, soit pédagogique, selon le type d'objectif d'évaluation déterminé.

Le procédé aurait le mérite d'être clair en évitant que les stéréotypes des normes sociales globales soient subrepticement réinjectés dans le dispositif lors de la restriction du produit-norme en ensemble des produits attendus.

- Le modèle de référence est constitué par l'application d'une échelle sur le produit attendu : il ne s'agit pas d'une nouvelle restriction mais plutôt d'un étalonnage qui complète l'opérationalisation du produit-norme en instrument d'évaluation. On a pu mettre en évidence que le type d'échelle (en particulier le sens dans lequel elle est utilisée, en positif ou en négatif) conduit à des évaluations différentes : mais surtout il semble que l'échelle elle-même soit plus ou moins pertinente aux objectifs initiaux.

Les derniers paragraphes permettent de compléter le modèle :



-L'évaluation est alors définie comme une comparaison entre un produit réel et un modèle de référence, comme un jugement comparatif dont il serait absurde de se demander s'il est "objectif" ou s'il n'est pas "objectif", mais dont on peut se demander s'il est fondé sur un modèle de référence pertinent ou peu pertinent aux objectifs qu'il représente, dont on peut se demander selon quelles règles de dérivation il est issu de ces objectifs.

- Le modèle est ainsi complété, hormis les boucles de feed-back qui peuvent rétroagir, à partir des évaluations, sur une instance du modèle ou sur une autre : on a pu montrer en effet que les écarts entre modèle de référence et produit réel peuvent être d'une autre nature ou d'une amplitude qui conduise à reconsidérer telle ou telle source de variation, à modifier ou à préciser la modalité de telle variable, ou son importance : échelle, informations concernant les producteurs, critères d'évaluation, et même tâche et objectifs d'évaluation, ce qui pourrait même aboutir à la remise en question des objectifs institutionnels ou de certains de leurs aspects. Cette voie reste à explorer ; elle peut être féconde si la recherche est associée à une formation des évaluateurs. Une autre voie de recherche serait celle de la relation unissant l'objet virtuel d'évaluation et le produit réel. Cette question est liée au problème général de l'ambiguïté de l'évaluation que nous avons systématiquement schématisé en le formulant en termes d'alternatives : objectifs explicites ou implicites, évaluation formative ou sommative, normes scolaires ou normes sociales globales, critères appropriés ou normes intériorisées.

La production, la performance ou la conduite n'est sans doute jamais qu'un représentant de ce qui intéresse l'évaluateur, c'est l'indice d'un niveau, d'une organisation ou d'un processus que vise l'évaluation ; or la même performance d'élève peut être considérée par différents évaluateurs, et selon leurs objectifs, comme l'indice de "ce que vaut l'élève", de "ce que vaut la méthode pédagogique" adoptée par l'enseignant, ou de "ce que vaut" le programme enseigné. Dans la mesure où la performance est effectivement une fonction composée de variables en interactions, dont les modalités ne sont pas toujours indépendantes, le choix qui est opéré d'un objet d'évaluation que représenterait la performance considérée est un choix arbitraire ; cela ne signifie pas qu'il est fortuit ou gratuit, cela signifie que la relation qui unit cette performance et cet objet doit être définie parmi d'autres relations, et non comme si elle était la seule possible, assimilée alors à une relation d'identité ou d'implication.

- Il est à remarquer que ce modèle simple, sans doute un peu trop simple, permet a priori d'analyser n'importe quel dispositif d'évaluation, quels que soient les produits réels et les objets virtuels sur lesquels portent les évaluations, et quel que soit le sujet de l'évaluation ; peu importe en réalité le statut du sujet de l'évaluation : si les évaluations de l'enseignant différent de celles de l'inspecteur, de l'élève ou de l'examinateur, alors qu'elles sont effectuées à partir du produit réel, c'est peut-être parce que le statut des évaluateurs n'est pas le même, mais c'est certainement parce que les modèles de référence des uns et des autres ne sont pas les mêmes, ce qui les conduit à repérer et à catégoriser des indices différents dans le même produit réel, ou catégoriser différemment les mêmes indices. (...)

Les déterminants de la note sont donc d'abord des déterminants abstraits, globaux ou spécifiques : disciplines, critères, types d'échelle, qui définissent eux-mêmes les indices à privilégier, et de ce fait construisent l'objet à évaluer. Nous sommes loin de la conception de l'évaluation déterminée par la "valeur intrinsèque" du devoir et de la notion corollaire de "vraie note".

Les raisons sont donc multiples pour que le modèle ne soit pas utilisé dans sa totalité pour décrire tous les comportements d'évaluation d'épreuves scolaires, tels qu'ils se produisent effectivement. En revanche si les différences constatées ou envisageables justifient la construction de modèles plus fins ou plus spécifiques, aucune ne contredit la plausibilité d'un modèle général, représentant, au sens ou l'entend Richard (1974) "à un niveau global une approximation de modèles plus fins, différents les uns des autres... et nécessaires pour décrire des mécanismes différents".

Une telle généralité peut présenter, sous certaines conditions, plus d'avantages que d'inconvénients si l'on s'interroge, comme le préconise Ribeill (1973) sur la nature, les fonctions, et l'usage qui peut être fait de ce modèle. Il est certain que d'autres modèles aussi généraux sont possibles, mais cela n'exclut pas que celui-ci se réfère à un champs de questions qui pour être limité n'en est pas moins important : réduisant sans doute la réalité qu'il représente de manière globale et simplifiée, il distingue et met en relief des rouages trop souvent confondus ou au contraire envisagés dans une indépendance fallacieuse : objectifs d'évaluation, critères, normes, barèmes... la définition des termes et des relations a permis d'introduire des notions nouvelles, comme celle de produit-norme et de remodeler des notions familières ambiguës, comme celles de critères ou d'utilisation d'une échelle de notation. Par le choix initial de sa forme et de son contenu, un grand nombre de présupposés conditionnent ses possibilités de représentation, aussi ne prétend-il ni à l'objectivité, ni à l'impartialité : si un évaluateur prétendait évaluer pour connaître la valeur d'une performance indépendamment de tout objectif, ou de tout rôle, joué par cette évaluation, il ne serait sans doute ni objectif ni impartial de récuser cette affirmation ; c'est pourtant ce que ce modèle inciterait à faire, décrivant une structure de l'évaluation, plutôt que l'évaluation telle qu'elle apparaît dans l'expérience quotidienne. Et cette structure, qui nous semblait, lors de sa construction, neutre, logique et pour tout dire innocente, laisse apparaître à l'usage la fonction politique qui lui est associée et l'option méthodologique qui la relativise. Après avoir constaté, au cours de ce travail, l'utopie de l'option initiale qui était celle de décrire ou de représenter les faits d'évaluation au moyen d'un modèle descriptif, la justification à ne pas abandonner maintenant cet instrument est l'utilité qu'il représente pour la construction ou la reconnaissance du sens de l'évaluation par les évaluateurs ; c'est une autre option, compatible avec les objectifs de recherche qui demeurent évidemment au premier plan, mais qui modifie le rôle que l'on peut attendre du modèle: il devient un stimulus, ou un ensemble de stimulus qui permet de relativiser les variables qui sont manipulées dans la recherche, qui peut aussi transformer les comportements d'évaluation des "sujets" qui l'utilisent. S'il perd ainsi d'un côté de la dignité épistémologique, essayons de montrer qu'il en gagne d'un autre côté par les conditions méthodologiques de son usage. En effet la recherche peut être, en ce domaine comme dans d'autres, un facteur de transformation sociale autant qu'un facteur de conservation sociale, l'un et l'autre sans doute dans un même mouvement qui les intègre dialectiquement. Il est possible que ce modèle occulte à certains moments, comme le pense Guigou (1972) les tensions et les antagonismes sociaux et renforce le pouvoir de l'idéologie dominante ; il est probable, bien qu'il ne le pense pas (mais faut-il lui renvoyer le reproche de se mouvoir dans un univers unidimensionnel ?) qu'en d'autres phases de la recherche il les renforce et les souligne : utiliser ce modèle, non seulement pour le chercheur mais pour l'évaluateur,

professeur ou élève, n'est-ce pas s'approprier des degrés de liberté de l'évaluation et se donner plus de puissance pour maîtriser les phénomènes selon ses objectifs ? N'est-ce pas enrichir son rapport à l'environnement en étant plus conscient des phases de changement dans son propre comportement, et de leurs déterminants ? (...)

Harris, M-L. et Steward, D-M., *Application of classical strategies to criterion-referenced tests construction*, American educational research association, New York, 1971
Hambleton, R-K., Nowick, M-R., Toward in integration of theory and method for criterion-referenced tests, *Journal of education measurement*, 1973, 10, 3, p. 159/169, 1973
Richard, J-F., *Attention et apprentissage*, Paris, PUF, 1974
Ribeill, G. Modèles et sciences humaines, Métra, 1973, 12, 2, p. 271/280
Guigou, J. *Critique des systèmes de formation*, Paris, Anthropos, 1972

BONNIOL, J-J.,

I, 2. La docimologie
I 2 1. La docimologie prescriptive et expérimentale,

figure de l'évaluateur produite

L'évaluateur docimologue projette une image de l'évaluateur qui se confond avec une sorte de fantasme du psychologue qui se croirait au-dessus de la situation, pur regard scientifique, juge, maître de la situation parce qu'externe, non impliqué, étudiant en toute innocence les dérives praticiennes des autres. Cette figure de l'évaluateur comme marionnettiste, tireur de ficelles - (penser au conte de Pinocchio) - n'a pas résisté à la perte de l'objectivité comme indiscutable critère de la scientificité.

PARISOT, J-C., "Le paradigme docimologique : un frein aux recherches sur l'évaluation pédagogique?", *L'évaluation en question*, CEPEC, 2°ed, Paris, ESF, 1988, p.37 56 :

(...) *Que retenir des travaux de docimologie ?*

Dans la très importante moisson de données rapportées par différents auteurs, ce qui frappe ce sont " les violents désaccords qui se produisent entre des juges compétents et appliqués".(10) La preuve en a été apportée, assénée pourrait-on dire, et cet apport reste déterminant pour une nécessaire relativisation. Elle fonde à l'évidence la nécessité de recherches sur des modalités d'examen qui seraient plus fiables. Au-delà des examens, la docimologie a montré que la notation, outil habituel et prédominant d'évaluation dans notre système scolaire est peu fiable. On peut retenir qu'en cas d'évaluation sommative pratiquée avec des outils normatifs et particulièrement dans le cas de décisions de certification ou de sélection, il importe de s'inspirer de certaines recommandations docimologiques.

Nous pensons que les techniques de standardisation sur spécification d'objectifs peuvent être utiles et amener des changements significatifs. Par contre, nous ne croyons guère, dans le contexte actuel, à l'efficacité des techniques de modération qui ne constitueraient qu'un pis-aller en évacuant le problème. Quant aux méthodes dites de centration des variables, elles nous paraissent relever de l'orthopédie statistique ou encore constituer "des systèmes perfectionnés pour continuer à mal faire les choses", comme l'écrit malicieusement G. de Landsheere.

Au total, si la docimologie a eu le mérite d'alerter sur un problème majeur, elle nous laisse relativement démunis quant aux solutions. (...)

La centration sur l'examen : une fausse piste

Dès l'origine, la docimologie s'est attachée à étudier les examens. Rappelons que la première recherche fondatrice a concerné le certificat d'études primaires (1922) et, comme en témoigne l'ensemble des travaux rappelés par Piéron, cette centration sur l'examen s'est maintenue jusque dans les années 1960. En privilégiant ainsi des problématiques bien précises, ces travaux ont découpé dans l'évaluation un champ d'investigation très local sur lequel pendant longtemps tout a été focalisé. Cette centration excessive s'explique assez facilement par le contexte historique. L'institution des examens et concours fait partie en France des conquêtes post-révolutionnaires. La plupart des charges et fonctions publiques étaient, sous l'Ancien Régime, des charges héréditaires - des offices. L'abrogation de ce système liant naissance et fonction a rendu nécessaire la mise en place d'un système de remplacement. C'est Napoléon qui, en 1808, a instauré les examens pour la délivrance des grades universitaires et les concours pour le recrutement ordinaire des fonctionnaires. d'où la préoccupation majeure, qui ne s'est pas démentie tout au long du XIX° siècle, de voir la justice garantie aux examens et concours. Cette préoccupation peut se répéter aussi dans l'ensemble des règles formelles de sérieux traditionnel telles que l'anonymat ou encore d'autres mesures, toujours en oeuvre aujourd'hui.

En s'intéressant au certificat d'études primaires dans le contexte de l'entre-deux-guerres, Piéron est dans un terrain hautement symbolique. Le certificat d'études - archétype de l'examen dans une France rurale - constitue dans l'inconscient collectif et dans l'idéologie républicaine la sanction sociale qui reconnaît le travail et le mérite de l'enfant sorti du peuple. Il importe donc au plus haut point d'en assurer la justice. C'est une préoccupation éthique.

A l'autre bout du système, les études de docimologie se sont concentrées sur le baccalauréat et les concours d'admission ou de sortie des grandes écoles. Là encore, on peut tout à fait relier la date de ces travaux (1950-1960) à des évolutions socio-économiques générales qui amènent à recruter les meilleurs sujets pour des filières de haut niveau. Si on lit attentivement Henri Piéron, on se rend compte que l'évolution socio-économique, l'histoire des réformes du baccalauréat, les questions posées par la docimologie sont en interactions permanentes.

D'une part, il s'agit de s'assurer que les examens et concours participent de manière pertinente au tri des compétences dont la nation a besoin (c'est une préoccupation de rendement social), d'autre part, il s'agit de recommander des systèmes plus performants et scientifiques pour gérer l'orientation et la sélection des populations scolaires de plus en plus nombreuses.(...)

Si la docimologie ne met pas en cause les examens ni les concours et qu'en cherchant à les améliorer, en quelque sorte, elle les consolide, c'est aussi parce que le régime des examens et des concours (pourtant assez sélectif) mesurant ou organisant l'échec d'une grande partie de la population scolaire a été cohérent jusqu'à une époque récente avec les exigences du fonctionnement économique et les lois de la stratification sociale. Dans une France des années 1950, personne ne parlait de l'exigence de mener 80% d'une classe démographique à la fin d'un second cycle, court ou long ! (12) .

Ces raisons historiques nous paraissent rendre compte pour l'essentiel de l'émergence de la docimologie et de ses orientations premières. Les représentations sociales de ce qui était important à étudier scientifiquement dans le domaine de l'évaluation appartiennent au paradigme que nous élucidons et dont les effets induits

ne sont pas négligeables Si l'on a pu admettre comme une évidence, à une certaine époque, que la question importante était de rendre les examens fiables, nous pouvons aujourd'hui voir les choses autrement : bien d'autres problématiques nous apparaissent plus essentielles dans le rapport à instaurer entre logique d'évaluation et logique de formation.(...)

Une certaine idée de la science

La docimologie permet d'analyser et de commenter certaines représentations de la scientificité dans les sciences humaines qui, malgré leur caducité, continuent sur un mode mineur ou travesti à marquer certaines approches ou conceptions épistémologiques de la recherche en éducation.

Nous pensons que ces représentations imposent des limites abusives et biaisent la recherche et qu'elles résultent d'une histoire des sciences que nous voudrions rappeler. Même s'il convient d'être prudent, il faut essayer de re-situer la docimologie dans le contexte général et la tonalité des travaux de sciences humaines au début du XX° siècle. Ces travaux sont largement tributaires et comme baignés dans une ambiance encore scientiste. Le XIX° siècle n'est pas loin ; on a alors une image un peu naïve de la toute puissance de la SCIENCE et d'une omniscience possible.

Les travaux de Darwin teintent encore un vaste sous-ensemble de travaux scientifiques. Auguste Comte voit dans la sociologie la science des sciences qui, lorsqu'elle sera établie, fondera raisonnablement politique et religion ... vision totalisante-totalitaire de la science. Dans des domaines plus précis comme la criminologie, la psychologie, la médecine, les travaux expérimentalistes abondent : on mesure, on quantifie, on cherche à dégager des lois générales qui expliquent et gouvernent les phénomènes observés. On essaye de déterminer la normalité et l'anormalité, on recherche surtout des fondements biogénétiques ou organiques aux "pathologies". L'investigation est souvent de facture déterministe et positiviste. Si François Broussais recherchait des liaisons entre la forme du crâne, celle du cerveau et le développement des fonctions mentales (phrénologie), Césare Lombroso mène des travaux sur les caractéristiques du "criminel né". Ces travaux qui ne sont que des exemples d'une ambiance scientifique montrent le paradigme dominant : il y aurait, il y a, un fondement organique, biologique, génétique aux différences interindividuelles de caractère ou d'aptitude. Il est intéressant à cet égard de décrire quelques éléments liés plus précisément à l'avènement de la docimologie. Nous avons emprunté pour cela à une chronologie de la psychologie différentielle et génétique (13) proposée par Norbert Sillamy.

Cette chronologie allégée montre la constitution d'un réseau épistémologique qui établit des filiations directes ou discrètes entre :

1. la recherche sur le caractère inné ou héréditaire des aptitudes ;
2. l'élaboration de modèles différentiels ou hiérarchiques de "l'intelligence"
3. l'utilisation de la "mesure", de la notion de normalité et des traitements statistiques et mathématiques en sciences humaines ;
4. la mise au point de test "d'aptitudes" en tous genres et particulièrement des fameuses échelles métriques de l'intelligence" ;
5. la docimologie elle-même. (...)

Il serait sans doute abusif de conclure trop hâtivement à un amalgame de la docimologie aux travaux de psychologie différentielle, mais on ne peut pas ne pas remarquer certaines interactions évidentes entre différentes recherches. Si l'on consulte la notice biographique concernant sir Galton, on voit qu'il s'intéresse d'abord à la question de l'hérédité du génie (1869), au développement des facultés humaines et développe la "psychologie" différentielle".

En 1884, Galton ouvre le premier laboratoire d'anthropométrie dans l'enceinte d'une Exposition internationale de la Santé à Londres. Il classe par ordre croissant de grandeur, toutes les mesures relevées et obtient une courbe en ogive (appelée depuis ogive de Galton (15) avec un palier représentant la mesure la plus fréquemment rencontrée.

En découpant l'effectif, il va repérer des limites utilisées depuis en statistiques (quartiles, déciles, centiles). L'échelle de référence (ou étalonnage) dont il dispose grâce à ce procédé, lui permet de classer n'importe quel sujet, pourvu qu'il appartienne à la même population que l'échantillon.

Nous verrons plus loin que l'un des référents clés de la docimologie est la distribution dite "normale" des notes qui attesterait de la justesse d'une notation. Le paradigme épistémologique fondateur de la psychologie différentielle et de la docimologie apparaît donc identique et nous en montrerons les postulats et présupposés dangereux pour l'action pédagogique.

Avant de le faire, résumons les caractéristiques les plus importantes des représentations de la scientificité à l'époque de l'avènement de la docimologie et qui en constituent alors le paradigme dominant :

- croyance en un déterminisme strict s'exprimant dans une causalité linéaire ;
- croyance positiviste et scientiste où la mise en évidence des lois universelles est considérée comme un gage de maîtrise et de toute puissance pour l'homme ;
- croyance en un déterminisme biogénétique ou organique qui explique les différences interindividuelles et les légitime ;
- fascination des techniques de quantification, du langage mathématique et des statistiques en particulier.

Ce modèle "naturaliste" des sciences humaines dont le cadavre remue encore, a eu des effets fâcheux dans le domaine de l'évaluation et de l'action pédagogiques.

Les représentations de l'éducation et de l'éducabilité

Notre hypothèse est la suivante ; la docimologie laisse fonctionner ou conforte subtilement une conception où l'action pédagogique a des effets restreints et où l'éducation même des sujets apprenants est objet de doute.

Pour tester cette hypothèse, nous avons choisi de nous intéresser de très près à la courbe de Gauss : elle constitue en effet un analyseur commode des présupposés et postulats communément admis par les docimologistes.

Nous n'ignorons pas que certains auteurs n'ont pas ménagé leurs critiques à l'encontre de cette fameuse courbe de Gauss et qu'ils ont relevé quelquefois avec beaucoup de pertinence les abus ou les difficultés que pouvait entraîner la courbe. Il nous semble pourtant qu'ils sont restés dans un registre de critiques plutôt techniques.

Par exemple, Anna Bonboir parle du "mythe de la courbe de Gauss" et recommande de voir dans chaque cas précis les repères qui pourraient être utilisés à bon escient ; "S'il est normal que dans la population totale les variables psychologiques et biologiques suivent cette courbe modèle, la transposition au domaine de la pédagogie ne peut se faire automatiquement car on ne peut prétendre à l'isomorphisme des théories."

Gaston Mialaret (17) s'en prend au "culte de saint Gauss" et G. de Landsheere prend lui aussi ses distances en posant à la fin de son ouvrage la question de la légitimité de ce référent. (...)

Les présupposés de la courbe de Gauss : les charmes discrets de la symétrie.

Si le modèle gaussien a été adopté aussi facilement, c'est parce qu'il présente toute les apparences de l'évidence : symétrie harmonieuse, calculs simples, cohérence avec, ou légitimation des idées reçues sur la distribution des aptitudes, cohérence

encore, avec les impératifs de sélection sociale et de choix des "meilleurs", habitudes mentales et socioculturelles... De plus, il se vérifie expérimentalement sur des variables aléatoires et semble rendre compte d'une sorte de "loi de la nature".

Le modèle.

Sur des effectifs nombreux, on a constaté que les caractéristiques physiques (tailles, poids, mesures anthropométriques diverses) se distribuaient en fréquence selon une courbe gaussienne qui est en fait la transcription en effectifs cumulés de l'ogive de Galton.

(...) La généralisation de ce modèle probabiliste aboutit à la loi normale réduite dont nous rappelons ci-dessus quelques caractéristiques (18). Dans une distribution normale, le mode, le médian et la moyenne coïncident ; le sigma ou écart type délimite des zones dont les pourcentages sont connus et stables.

A partir de ces constats probabilistes liés à des caractéristiques physiques mesurables ou à d'autres paramètres naturels aléatoires, on a inféré qu'il devait en aller de même pour des traits psychologiques ou encore des dons et des aptitudes mentales censés être distribués eux aussi ordinalement... et normalement aléatoirement ! Ce postulat rejoint la représentation commune selon laquelle il y a des gens très intelligents (peu), des gens moyennement intelligents (beaucoup) et des gens peu intelligents (peu).

Au nom de cette même représentation, on va donc s'attendre, lorsque des candidats ont à affronter la même épreuve, à ce que :

- très peu obtiennent de très mauvaises notes ;
- beaucoup obtiennent des notes moyennes ;
- très peu obtiennent des notes excellentes.

On sait que la notation sous la forme habituelle (de 0 à 20) est une échelle ordinaire d'intervalles. La transcription en notes des aptitudes mentales va dans ce paradigme suivre donc également les caractéristiques de la loi normale réduite. D'où l'idée qu'une notation juste doit se conformer à la courbe de Gauss qui constitue pour la docimologie traditionnelle le référent mathématique suprême. Nous avons vu que certains auteurs contestaient cette manière de voir pour des raisons techniques. Nous pensons que ce qui est en cause est beaucoup plus fondamental encore.

Imaginons que l'on propose après apprentissage systématique à des élèves un exercice de mise en oeuvre d'une règle et que tous y parviennent... Supposons maintenant que l'on veuille noter un tel exercice... La "logique" voudrait que chacun ait 20/20, mais, dans ce cas, la notation est systématiquement aberrante. Un correcteur placé dans ce dilemme risque donc d'être amené à réintroduire une normalité dans sa notation, quitte à justifier a posteriori les différences de notes par des considérations et rationalisations diverses (temps mis à répondre, soin de l'écriture, etc). Comment faire autrement en effet ? pour qu'une notation soit juste selon Gauss, il faut qu'elle classe ordinalement et normalement les élèves, ce qui conduit en toute logique à proposer des travaux et exercices dont la nature permet de sélectionner, de distinguer, de classer, c'est-à-dire soit d'installer un niveau de difficulté sélectif au départ, soit noter sur des paramètres se révélant après coup discriminants. Par exemple, pour des candidats au concours d'entrée à Polytechnique, les mathématiques n'étant pas un bon outil de sélection, l'échec ou l'admission au concours peuvent se jouer sur des matières où les candidats atteignent des performances moins homogènes.

L'autre solution est de distinguer ordinalement les élèves par un système de notation très fin ; un peu comme pour les champions de ski quand les 1/1O° de seconde n'ont plus permis de les classer, on s'est mis à chronométrer au 1/1OO° de seconde.

Comme les enseignants utilisent très largement la notation même s'ils sont en dehors de tout contexte d'examen, on peut se demander si la représentation d'une "normalité" des notes ne contribue pas à faire dégénérer chroniquement des évaluations, pédagogiques dans leur intention, en opérations contaminées de sanction et de sélection. Si l'on peut admettre que le modèle gaussien de distribution des notes est une référence de justesse, ce n'est que dans la mesure où sont vérifiées deux caractéristiques :

1. L'épreuve objet de notation a été standardisée et étalonnée sur de grandes populations -un protocole de correction est fourni-, la fidélité a été vérifiée.

2. On se place dans une logique de repérage par rapport à un niveau général ou dans une logique de sélection, classement ou encore d'admission certificative.

Par contre, il nous paraît inadmissible épistémologiquement et déontologiquement d'admettre comme allant de soi la courbe de Gauss dans une perspective pédagogique pertinente et systématique, les acquisitions des apprenants ne se ventileront pas de manière aléatoire ! La "normalité" gaussienne apparaît là plutôt comme une monstruosité, irrecevable pour le pédagogue qu'elle réduirait à l'impuissance.

En fait les résultats de l'action pédagogique ne sont pas aléatoires et on a sans doute intérêt à les évaluer sur des comportements précisés et de manière binaire. L'emploi d'une échelle ordinaire comme la notation introduit une logique différentielle légitimée et mathématisée par la courbe de Gauss. Cette logique différentielle fait bon ménage avec une conception plus ou moins innéiste ou naturaliste des aptitudes et opère un renforcement de cette représentation.

La docimologie, en faisant ou laissant croire qu'une bonne évaluation était une bonne examination et devait se traduire par une bonne notation (conforme à la loi normale réduite), a contribué à installer de manière occulte et tenace une conception où l'éducabilité est restreinte, l'efficacité pédagogique faible et où l'idéologie des dons plus ou moins édulcorée se robore et perdure. Cette conception que nous croyons nocive et bloquante pour la recherche et l'action pédagogiques constitue peut-être la partie la plus importante du paradigme que la docimologie permet de dévoiler. (...)

(10) conclusion de G. Gastinel à la suite de l'enquête de la commission Carnégie, 1932, rapportée par H. Pièron

(12) Antoine Prost, *Les lycées et leurs études au seuil du XXI° siècle*, Paris, CNDP, 1983

(13) N. Sillamy, Extrait du *Dictionnaire de la psychologie* II, Paris, Bordas, 1980

(15) cette ogive de Galton transcrite en histogramme de fréquences devient la courbe de Gauss

(17) *Docimologie et éducation*, Discussion colloque AIPELF, Lyon, 1968

(18) Pour de plus amples renseignements, consulter J-M, Faverge, *Méthodes statistiques en psychologie appliquée*, Paris, PUF, 1975 et J-M Reuchlin, *Précis de statistiques,* Paris, PUF, 1976

PARISOT, J-C.,

Cardinet, J. "L'objectivité de l'évaluation", *Formation & technologies*, n°0, 1992, p. 17 / 25 :

(...) Le mode de correction ajoute sa part de variance à la mesure du résultat. Ce fut même le premier facteur perturbateur mis en évidence par les études docimologiques. La psychologie du notateur et les principaux effets qui modifient son jugement ont été réétudiés plus récemment (Noizet et Caverni, 1978 ; Bonniol, 1981), mettant en évidence les sources du manque de fidélité découvert précédemment (effets d'ancrage, de contraste, etc...), sans résoudre pourtant les problèmes pratiques correspondants. Faudrait-il ramener tous les examens à une forme unique, celle des questions à choix multiples ? Ce serait oublier que cette forme d'interrogation

introduit ses propres biais : pour beaucoup d'élèves, il est plus facile de reconnaître la bonne réponse que de la formuler entièrement (Cardinet et Dubosson, 1969).

4.4. Exigences des évaluateurs différentes

Dans le cadre de la classe, l'appréciation du résultat se traduit, traditionnellement, par la mise d'une note. Or, aucune fonction objective ne relie le nombre ou la qualité des réponses avec cette appréciation. Chevallard (1986) a bien démontré l'arbitraire du jugement ainsi porté, qui ressemble plus à un marchandage avec la classe qu'à une mesure. Selon le moment où elle se situe, l'interrogation servira à obliger les élèves à travailler davantage ou, au contraire, à remonter une moyenne de classe compromettante pour l'enseignant.

Si, pour un même évaluateur, les exigences évoluent, il va de soi qu'elles varient d'un enseignant à l'autre, comme le montre Boumard (1978) en analysant un conseil de classe. Elles varient surtout selon les écoles d'après les recherches de Duru et Mingat (1987), en raison de cultures locales spécifiques.

Elles varient enfin, selon le niveau socio-culturel du milieu social d'origine des élèves, des effets d'ancrage déplaçant les échelles d'évaluation pour les adapter à la performance moyenne observée par l'enseignant.

4.5. Significations de l'examen différentes

Une dernière source d'hétérogénéité a été mise en évidence récemment par les travaux de psychologie sociale de l'éducation. Perret-Clermont (1979) a montré qu'un enfant passait rapidement d'un stade de développement opératoire au stade supérieur dans un contexte d'examen où la dynamique de groupe social favorisait sa réflexion, par opposition au contexte de l'examen classique, mené par un adulte. Etait-ce fausser les conditions d'examen, comme auraient tendance à le dire les piagétiens orthodoxes ? C'était plutôt montrer que toute situation d'évaluation, quelle qu'elle soit, affecte inévitablement la probabilité de réussite et qu'il n'existe pas de situation de référence permettant une mesure objective. Schubauer-Léoni et Perret-Clermont (1985) ont pu observer les mêmes phénomènes à l'oeuvre lors d'interrogations relatives à la mathématique : les enfants donnent un sens à la question qui leur est posée, d'après ce qu'ils comprennent de la situation d'examen dans son ensemble. Carraher (1987) a montré également, que des enfants qui avaient résolu correctement des problèmes mathématiques dans une situation réelle se révélaient incapables de le faire à nouveau en situation de classe : ils ne mobilisaient plus les mêmes démarches de pensée et perdaient alors la référence à la réalité qui les avait guidés dans le premier cas. Laquelle est la compétence véritable de ces élèves ?

Il n'est pas plus facile de répondre pour d'autres situations où l'examen, apparemment objectif parce que le même pour tous, recèle en réalité un biais par rapport à un autre point de référence. Les épreuves de sélection scolaire qui sont encore passées dans certains cantons suisses à l'âge de onze ans favorisent massivement les filles qui ont atteint plus tôt que les garçons certains traits de maturité intellectuelle et affective. La différence serait sans doute moins grande quatre ans plus tard. La seconde mesure serait-elle plus "objective" que la première ? (...)

Noizet, G & Caverni, J-P., *Psychologie de l'évaluation scolaire*, Paris, PUF, 1978
Bonniol, J-J. "Déterminants et mécanismes des comportements d'évaluation d'épreuves scolaires", Thèse, Université de Bordeaux II, 1981(6)
Cardinet, J & Dubosson, J. "Comparaison des propriétés métriques de tests à réponse libre et à choix multiple, *Revue suisse de psychologie*, vol 28, 1969, p.12/27
Chevallard, Y. "Vers une analyse didactique des faits d'évaluation", De Ketele, J-M., *L'évaluation approche descriptive ou prescriptive ?*, Bruxelles, De Boeck, 1986, p. 31/ 59
Boumard, P. *Un conseil de classe très ordinaire*, Paris, stock, 1978

Duru, M. & Mingat, A., « Le redoublement au collège, en France ; éléments pour une évaluation », *Mesure et évaluation en éducation*, Vol 10, n°3, 1987, p. 5/28
Perret Clermont, A-N., *La construction de l'intelligence dans l'interaction sociale*, Berne, Peter Lang, 1979
Carraher, T.N. Mathematics as a personal and social activity. Communication au congrès international "Le fonctionnement de l'enfant à l'école", Université de Poitiers, 17-20 juin 1987

Cardinet, J.

BONNIOL, J-J. , GENTHON, M. & ROGER, M "L'évaluation en psychologie : approches théoriques et conditions méthodologiques", *AESE* n°6, 1986, p. 12/18 :

Si l'on établit un état des lieux concernant l'évaluation, le constat est celui d'une dynamique intéressante.

Au début était la mesure.

Ainsi l'ouvrage de Piéron (1963), celui de De Landsheere (1974) ou encore celui de Noizet et Caverni (1978) dans sa première partie sont centrés sur l'étude des résultats de l'évaluation ; ils décrivent les dysfonctionnements de l'évaluation, les divergences de notation et les diverses procédures empiriques grâce auxquelles on a tenté de réduire ces dysfonctionnements et ces divergences. L'avantage de ces recherches de la première période de la docimologie, dite descriptive était de constater l'ampleur des phénomènes de divergences et de poser des hypothèses interprétatives, sans que ces résultats soient réinjectés auprès des évaluateurs, les rapports entre chercheurs et évaluateurs étant inexistants. Ces recherches pouvaient permettre une éventuelle correction a posteriori des distorsions, mais leurs résultats ne concernaient que les effets et non les mécanismes responsables. La période suivante, celle de la docimologie expérimentale (4), étudiait le comportement, le fonctionnement de l'évaluateur avec des dispositifs expérimentaux. Elle a permis de mettre en évidence des dysfonctionnements de type "déformations perceptives". L'avantage de ces recherches était la généralisabilité des résultats. Mais l'objet de la recherche était alors restreint d'emblée aux rapports construits par le chercheur entre les variables de situation et les comportements des sujets.

La prééminence de la "structure" instrumentale de l'évaluation - recherche du meilleur instrument, de la meilleure situation - occultait les questions sur les modélisations sous-jacentes. Implicitement la référence de l'évaluation se construisait autour de l'objet approximativement défini, dans le cadre scolaire, par des contenus et des normes d'excellence.

Cette prégnance de l'objet et du rapport à l'objet comme résultat de toute évaluation, traduisait la non-perception du fonctionnement propre du sujet au-delà de ses aspects périphériques et l'extrême difficulté d'une "objectivation" de la complexité individuelle.

Ceci permettrait sans doute de sensibiliser les évaluateurs aux anomalies produites par leur fonctionnement, mais ne leur proposait pas d'indication en termes de remédiations. Il manquait des liens tissés entre problématiques pratiques et problématiques théoriques, entre les problématiques des évaluateurs et celles des chercheurs. Par ailleurs ces recherches ne permettraient pas d'expliquer les dysfonctionnements repérés dans des tâches d'évaluation d'épreuves scolaires où se sont à l'évidence des mécanismes cognitifs qui fonctionnent plus ou moins bien.(...)

Piéron, *Examens et docimologie*, PUF, 1963
De Landsheere, G. & V. *Evaluation et docimologie*, 1974
Noizet, G., Caverni J.P., *L'évaluation scolaire,* 1978

(4) cf. la seconde partie de l'ouvrage de Noizet et Caverni et la première partie de la thèse de Bonniol

BONNIOL, J-J. , GENTHON, M. & ROGER, M

I. 2.2.
La psychologie de l'évaluation ou doxologie

Champ d'étude :

L'intérêt docimologique se déplace, dans ce modèle local, vers le fonctionnement des évaluateurs correcteurs de copies, vers l'étude des comportements des examinateurs et des examinés en milieu scolaire. Sont mis à jour une série de mécanismes explicatifs des variations d'un correcteur à l'autre.

Le syntagme (le mouvement, l'école) de la psychologie scolaire est une charnière entre l'analyse critique du fonctionnement de l'acte de noter une copie ou un candidat (docimologie prescriptive et docimologie expérimentale) et la réflexion sur les conditions qu'on peut mettre en place d'abord dans la constitution des modes de passation (1) et puis dans la formation elle-même (2), pour que la réussite à l'examen soit optimale

Les dimensions sociales, fonctionnelles et organisationnelles, vont venir compléter les théorisations de l'évaluation. Le travail se déplace vers le rôle, la nature et la fonction des critères d'évaluation, au service notamment de l'acquisition des contenus de formation, au détriment de l'étude "interne" et de la modélisation du fonctionnement de l'évaluateur.

Alors, l'évaluateur est remis dans le processus de formation, il n'est plus réductible au seul rôle de correcteur : la correction elle-même est pensée en connection avec les autres dimensions de la situation de formation -- commence l'étude de l'influence du critère d'évaluation sur le fonctionnement des sujets de l'évaluation.

(1) ce sera le modèle de la métrie
(2) ce sera la matrice de l'évaluation comme gestion.

BONNIOL, J-J., "Influence de l'explicitation des critères utilisés sur le fonctionnement des mécanismes d'évaluation d'une production scolaire", *Bulletin de psychologie* XXXV, n°353, 1981 :

L'évaluation d'une production scolaire a pu être assimilée à une tâche de jugement perceptif (Bonniol, 1972, Amigues et al, 1975) au cours de laquelle l'évaluateur compare la production qu'il doit juger à un modèle de référence relativement stable, déjà constitué dans ses structures cognitives et qui peut évoluer au cours de la tâche ; constitution et évolution s'effectuent sous l'influence de déterminants divers. Certains de ces déterminants n'ont pas grand chose à voir avec la production elle-même et peuvent être considérés comme des biais de l'évaluation : il

en est ainsi du statut scolaire des élèves (Bonniol et al, 1972) de son statut ethnique (Amigues et al, 1975) ou de l'ordre de présentation de la série de devoirs (Bonniol, 1965) pour ne citer que quelques exemples des biais qui ont été mis en évidence expérimentalement.

Si le modèle de référence de l'évaluateur n'est pas suffisamment consistant pour résister à l'influence de facteurs qui le modifient sans qu'il en soit toujours conscient, il est vraisemblable que cela tient pour une large part au fait que les critères en fonction desquels la tâche d'évaluation devrait être accomplie ne sont pas toujours aussi explicites que chacun —et l'évaluateur lui-même— pourrait le penser.

La recherche ici présentée avait pour objectif de tester une première hypothèse selon laquelle les évaluateurs, lorsqu'ils disposent de critères explicites, devraient évaluer autrement que lorsque les critères sont relativement ambigus.(...)

2. Les conditions d'élaboration des critères

Ce problème de l'influence des critères utilisés sur l'évaluation d'une production scolaire est en effet pour le chercheur un problème délicat, au moins pour deux raisons : d'abord il ne peut pas se mettre à la place de l'enseignant pour déterminer les critères pertinents dans l'évaluation d'une tâche qui a été donnée à des élèves d'un certain niveau — qu'il ne connaît pas — déterminé éventuellement par un moment dans une progression — qu'il ne connaît pas. Mener une recherche sur les critères d'évaluation effectivement utilisés par les enseignants dans une situation concrète requiert alors une coopération qui ne peut se réduire au passage des uns dans les arcanes d'une situation expérimentale traditionnelle organisée par les autres.

D'autre part, le critère est un concept relativement ambigu à manipuler du fait de "son usage fréquent et parfois interchangeable" comme le dit Astin (1964) y compris chez le même évaluateur et en toute rigueur : on conçoit que le critère "style" n'ait pas la même signification quand il s'agit de devoirs d'élèves de niveaux différents dans le cursus scolaire par exemple. Il sera pris ici dans le sens d'une classe ordonnée d'événements qui est privilégiée parmi d'autres comme dimension du modèle de référence utilisé par l'évaluateur. Ainsi la couleur est un critère que l'on peut choisir de privilégier, ou dont on peut tenir compte parmi d'autres, dans l'évaluation d'un objet. *Un critère est donc toujours objectif, seul le choix qui en est fait est éventuellement subjectif, sans que cela signifie pour autant qu'il soit arbitraire***.** Cette définition reprend et complète celle d'Astin, pour qui le critère est une "classe d'événements", par l'idée d'"ordre", de "dimension" qui définit généralement le critère utilisé lors de l'évaluation. On peut remarquer d'autre part que le critère *peut être plus ou moins précis*, c'est-à-dire qu'il peut être énoncé sous forme conceptuelle, donc relativement abstraite ou en termes d'indicateurs comportementaux qui le concrétisent, sous forme d'exemples et de contre-exemples qui l'explicitent. (...)

La solution qui fut ici choisie était étrangère aux procédures habituelles de recherche, en ce que les sujets-évaluateurs ne devaient pas fonctionner pour le chercheur puisque la question était justement de *savoir comment le sujet opérationnalise les critères qu'il utilise, et d'abord à quelles fins il les opérationnalise*. Aussi fallait-il éviter de dévoyer les objectifs d'évaluation des sujets et les rapports entretenus entre ces objectifs d'évaluation et les objectifs pédagogiques ; ces deux types d'objectifs et leurs rapports devaient être précisés dans le seul intérêt de leurs élèves et de leur propre fonctionnement. De même les critères devaient être élaborés, précisés et assortis d'exemples et de contre-exemples à l'occasion des exercices proposés aux élèves dans la dynamique de leurs apprentissages, en fonction des objectifs définis par l'enseignant aux différents moments de l'année scolaire.

Il a donc paru opportun d'inciter les enseignants à définir opérationnellement les critères qu'ils jugeaient pertinents pour l'évaluation de différentes tâches, plutôt que de les définir nous-mêmes arbitrairement ou d'accepter une définition conceptuelle très générale telle que "style" ou "logique du plan". Qu'est-ce qu'un plan logique ? Il y a sans doute plusieurs critères utilisables pour l'apprécier, critères parmi lesquels des choix différents peuvent être effectués par les enseignants ; parler de "logique du plan" sans plus de précision, ce n'est pas parler d'un critère, mais d'une norme, d'un ensemble de critères implicites parmi lesquels le chercheur n'a aucune information sur ceux d'entre eux qui sont privilégiés — par tel enseignant, à tel moment—.

3. Les hypothèses sur l'influence de l'explicitation des critères

(...) Les hypothèses portent sur la possibilité que des critères soient explicités et utilisés à la place des normes implicites que les enseignants — et les bons élèves — intériorisent par osmose. Elles posent que les évaluations différeront alors de celles qui sont effectuées sur des exercices analogues sans que les critères aient été explicités. (...) il est alors nécessaire de disposer d'un indice d'explicitation des critères suffisamment consistant. L'idée simple était que, si les critères étaient véritablement explicités, ils seraient explicités pour les élèves qui pourraient alors les utiliser de manière pertinente en particulier pour auto-évaluer leur production. L'hypothèse était donc que les évaluations des enseignants et les auto-évaluations indépendantes des élèves effectuées sur les mêmes devoirs seraient systématiquement plus conformes les unes aux autres dans le cas où les critères auraient été explicités au préalable que dans le cas où ils ne l'auraient pas été. Les élèves sont donc placés ici en situation d'apprentissage de l'évaluation, ils sont considérés comme sujets évaluateurs (...)

Le facteur expérimental principal est donc le mode de connaissance des critères à deux modalités : "connaissance intuitive" et "connaissance explicite" ou le niveau d'explicitation des critères qui sont, ou ne sont pas, fournis aux élèves en même temps que le sujet du devoir à effectuer sur table.

Deux hypothèses sont rattachées à ce facteur si l'appropriation des critères d'évaluation doit permettre, en même temps qu'une auto-évaluation plus proche de l'évaluation ultérieure du professeur, une meilleure performance de l'élève : en effet les critères d'évaluation fournissent aux élèves, lorsqu'ils sont donnés en même temps que la consigne, des informations supplémentaires sur les comportements attendus, qui doivent leur permettre de mieux évaluer leurs résultats, mais qui pourront antérieurement être utilisés comme des signaux leur permettant de guider la réalisation de la tâche, donc probablement de mieux la réussir. (...)

4. Présentation de quelques résultats

4.1 - Résultats quantitatifs

(...) On constate d'abord que l'explicitation des critères joue le rôle attendu, quel que soit le type d'évaluation, dans chaque classe, en début comme en fin d'année : enseignants et élèves attribuent systématiquement des notes meilleures lorsque les critères sont explicites. (...) Il est remarquable que ces différences soient marquées quel que soit le type d'évaluation, qu'il s'agisse des évaluations de l'enseignant, de celles que les élèves pensent obtenir ou de celles que les élèves pensent mériter.

On constate ensuite que pour un même niveau d'explicitation des critères, (...) les professeurs attribuent des notes plus fortes en fin d'année en classe de seconde, (...) Ainsi les évaluations des professeurs sont en début d'année inférieures aux notes que les élèves pensent mériter en seconde, et en première lorsque les critères sont implicites ; en fin d'année au contraire, les évaluations des professeurs sont systématiquement supérieures aux notes que les élèves pensent mériter, et l'écart est

plus important quand les critères sont explicites. Ce phénomène peut être interprété comme l'indice d'une certaine "viscosité" du système de représentation de soi chez les élèves dont l'expérience scolaire et le proche passé d'échecs ne facilitent pas la transformation de l'image qu'ils ont de leur mérite. Cette "viscosité" est d'ailleurs tout à fait relative puisque les notes qu'ils pensent mériter en fin d'année sont nettement supérieures, lorsque les critères sont explicites, à ce qu'elles sont lorsque les critères sont implicites, surtout en première où l'écart dépasse deux points en moyenne entre deux séries de 192 notes.

Ces résultats sont confirmés dans l'ensemble par les comparaisons demandées, en particulier celles qui concernent les évaluations des enseignants dans chaque discipline (...), mais il y a des différences d'une discipline à l'autre. (...) Donc bien que l'explicitation des critères d'évaluation n'intervienne pas partout avec la même force, il est intéressant de souligner que l'on n'enregistre nulle part une tendance inverse à celle qui était attendue mais qu'au contraire tous les résultats concordent.(...)

L'hypothèse principale semble alors confirmée : s'il est possible que les critères d'évaluation explicités aux élèves soient autant d'éléments dont ceux-ci disposent pour guider la poursuite de leur travail et leur permettre de réaliser de meilleures performances, il est certain, en tout état de cause, qu'ils déterminent nettement les évaluations qui sont effectuées.

Puisque les élèves sont en mesure, lorsque les exercices possèdent des critères d'évaluation explicites, de répondre de façon plus adéquate aux comportements attendus, et puisque la connaissance de ces critères leur permet aussi de mieux comprendre et par conséquent de mieux utiliser les évaluations qui sont effectuées, il semble que cette méthode représente dans le déroulement de ses différentes étapes plus qu'une technique d'évaluation sommative convenable, peut être plus qu'un mode d'évaluation formatif, une véritable technique pédagogique, technique de guidage du travail scolaire dont l'effet bénéfique est plus ou moins rapide selon les disciplines, et plus ou moins transférable, mais systématique. Cette technique nécessite de la part des enseignants qui l'utilisent un travail d'élucidation et de clarification de leurs objectifs et de leurs outils d'évaluation. Elle s'avère par ailleurs d'autant plus efficace que les critères utilisés dans la discipline sont plus précis et que les élèves ont pu se familiariser davantage avec leur usage. (...)

4.2 - Analyse qualitative des auto-appréciations des élèves

(...) La première question est donc de savoir s'il existe un accroissement des références aux critères en cours d'apprentissage, et au cas où un tel accroissement existerait, s'il s'effectue préférentiellement sur la base d'une augmentation des références qui sont faites à des critères précis.(...) La question se pose alors de savoir si l'explicitation des critères d'évaluation influe sur l'accroissement des références aux critères précis en cours d'apprentissage, si l'explicitation des objectifs et des critères du devoir avant la réalisation de la tâche constitue une situation déterminante au point de modifier qualitativement les auto-évaluations des élèves ? (...)

Il est donc probable qu'entre le début et la fin de l'année, l'exercice d'une part et l'explicitation des critères d'autre part contribuent à accroître la lucidité et le réalisme des auto-évaluations portées ; la part qui revient à l'explicitation des critères est ici fortement soulignée ; elle joue comme un révélateur de la précision des jugements que les élèves peuvent porter et permet d'accélérer les processus qui lui permettent de se manifester.

Conjointement à cette augmentation du nombre des références à des critères précis, s'opère un abandon des critères flous et imprécis, sous l'effet de l'apprentissage (...) et en début d'année sous l'effet de l'explicitation des critères d'évaluation (...) En

fin d'apprentissage, et sous l'effet de l'explicitation, la précision des jugements des élèves a presque quadruplé.

(...) En conclusion, il n'y a qu'en Français que les résultats sont un peu différents de ceux qui étaient attendus. Dans les deux autres disciplines, on observe en cours d'apprentissage :

Pour les élèves de première un accroissement de la précision et une baisse simultanée des références à des critères imprécis, quel que soit le type d'exercice. L'apprentissage de l'autoévaluation et de la manipulation des critères permet une précision croissante, y compris quand les critères n'ont pas été explicités.

Pour les élèves de seconde, on constate également un accroissement des références à des critères précis mais qui n'est pas encore accompagné d'une diminution des références à des critères imprécis, sauf en Economie quand les critères sont explicités.

Le risque était que ces résultats soient à usage interne, qu'une fois franchies les portes de la classe, ces élèves se retrouvent aussi démunis qu'auparavant. Les résultats obtenus par les élèves de première G à l'épreuve anticipée de Français au baccalauréat permettent de penser qu'il n'est est rien.

5. Conclusion

Dans cette étude le problème était celui de l'importance relative des critères d'évaluation, comme source de variation constitutive du modèle de référence de l'évaluateur ; ces critères, dans l'ensemble des études analysées, étant des dimensions de la tâche, des axes du système de normes scolaires, par opposition aux dimensions de l'élève, aux axes du système de normes sociales générales. L'opposition ne signifie pas que les deux systèmes de normes soient antagonistes ; on sait au contraire que les normes scolaires fonctionnent selon les mêmes règles que les normes sociales qui les engendrent : la règle de l'implicite, la règle du nominal, la règle de l'osmose, dont le jeu subtil assure la sélection des "bons" élèves aussi sûrement que celui des normes sociales générales, sur le marché du travail ou sur le marché de l'argent, assure la sélection des bons ouvriers ou des bonnes entreprises.(...)

Si les résultats de la dernière recherche autorisent, sinon une généralisation, du moins une hypothèse, ils suggèrent que plusieurs phénomènes comportent des indices pertinents de l'opérationnalité des critères utilisés dans l'évaluation des productions scolaires : le premier phénomène serait qu'une annotation du professeur, référant un élément du devoir à un critère ou à plusieurs d'entre eux, suffirait à induire une autocorrection par l'élève ; le second phénomène serait une centration différente des évaluations chiffrées sur l'échelle de notation utilisée, et de plus en plus haute au fur et à mesure de l'apprentissage, de l'appropriation des critères par les élèves dont les résultats se distribueraient en fin d'apprentissage réussi selon une courbe en J, et non bimodale ou conforme à une distribution au hasard. (...) (Si les critères) jouent leur rôle spécifique, au lieu de préfigurer, dans l'organisation de l'évaluateur, les normes sociales auxquelles les individus doivent se conformer par divination confuse ou par osmose, l'hypothèse selon laquelle les autres sources de variation des évaluations auraient alors une influence très atténuée, n'est pas une hypothèse absurde, qu'il s'agisse des effets séquentiels ou des caractéristiques différentielles des élèves ; l'échelle utilisée serait elle-même déterminée, comme les critères, par les objectifs et le moment de l'évaluation. (...)

Bonniol, J-J., Les comportements d'estimation dans une tâche d'évaluation d'épreuves scolaires, étude de quelques-uns de leurs déterminants, Thèse de 3° cycle, Aix-en-Provence, 1972

Amigues, R., Bonniol, J-J., Caverni, J-P., Les comportements d'évaluation dans les systèmes éducatifs. Influence d'une catégorisation ethnique sur la notation de productions scolaires, *Journal International de Psychologie*, 1975, 10, 135-145
Bonniol, J-J., Les divergences de notation tenants aux effets d'ordre de correction, *Cahier de psychologie,* 1965, 8, 181-188
Astin, A-W., Criterion centered research, *0* 1964, 24,4,807/ 823

BONNIOL, J-J.,

I 2 La docimologie
I 2 2. La psychologie de l'évaluation,
point de vue des détracteurs

C'est la notion de critère d'évaluation qui fait écran aux conclusions de la psychologie de l'évaluation. Incomprise, confondue avec la notion d'objectif ou ignorée, la "dimension critère" va demander, pour être explicitée, l'apparition d'autres écoles de recherche de plus en plus éloignées du modèle de la docimologie.

ABRECHT, R. *L'évaluation formative, une analyse critique*, Bruxelles, De Boeck, 1991 :

(...) La notion de critère est pour nous moins claire. Tantôt il apparaît comme une norme (Nunziati, 1984, p. 24) , tantôt il se rapproche d'un point de vue sur l'apprentissage, tantôt encore il est assimilable à une information de guidage. Sous toutes ses formes, de la plus générale à la plus utilitaire, il nous paraît malgré tout demeurer empreint d'extériorité, par rapport à une démarche d'élève (ou, à tout le moins, nous semble quelque chose de trop "soufflé"). (...)

Nunziati, G. "Evaluation formative et réussite scolaire", *Collège* n°2, Bulletin de liaison pour la rénovation des collèges, MAFPEN Aix-Marseille, 1984, P 18/37

ABRECHT, R.

CARDINET, J. "L'évaluation en classe, mesure ou dialogue ?" dans *Hommage à Cardinet,* Fribourg, Delval, 1990, article publié dans *European Journal of psychology of education,* vol 11, N°2, 1987, p 133/144 :

(...) *L'ancien rôle de la psychologie :*

Depuis les travaux de Binet au début de ce siècle, les psychologues pensent être en mesure d'améliorer l'évaluation scolaire, en appliquant les méthodes psychométriques à la mesure des apprentissages. De fait, la docimologie d'un côté, les tests de connaissances scolaires d'un autre, se sont développés entre les deux guerres, parallèlement à la psychologie différentielle. Même la psychologie générale a contribué à mieux comprendre le comportement de jugement des notateurs, comme l'ont montré les travaux de Noizet et Caverni (1978), ou ceux de Bonniol (1981).

Une limite importante des épreuves construites comme des tests psychologiques, pourtant, c'est qu'elles s'appuient presque uniquement sur le classement des élèves les uns par rapport aux autres. Or la logique d'un examen est différente : il s'agit de voir si un niveau de compétence, défini dans l'absolu, est atteint ou non. Une épreuve pédagogique devrait aussi rendre compte des apprentissages réalisés, même s'ils ne différencient plus les élèves, du fait qu'ils sont acquis de façon générale. Négliger cette "mesure absolue" risque de transformer

l'évaluation scolaire en un concours, où les classes sociales privilégiées seront toujours gagnantes, parce que les acquisitions de la majorité des élèves sont passées sous silence. (...)

L'évaluation en classe n'est pas conçue comme une mesure des performances de l'élève.

Les enseignants ne cherchent pas un critère extérieur :

L'hypothèse de travail spontanée d'un psychologue est que les enseignants ont besoin d'épreuves objectives, par exemple pour déterminer si un élève maîtrise suffisamment bien un sujet pour commencer l'étude d'un autre. En fait, cela ne semble pas être le cas, si on laisse de côté le rôle défensif du test vis-à-vis des parents. (...)

En ce qui concerne la question de l'enseignement, les maîtres(ses) font d'abord confiance à leurs observations. (...)

L'évaluation en classe ne vise pas un trait latent :

La mesure en psychologie porte sur des caractéristiques du comportement des individus, qui peuvent varier en grandeur sans changer de nature, comme des temps de réaction, l'empan de la mémoire d'une série de chiffres, le nombre de problèmes résolus en un temps donné, etc. Chaque mesure détermine la position de l'individu étudié sur une échelle commune à tous les sujets, qui est censée représenter un trait commun à tous.

Il est probable que la note à un examen, ou à une évaluation sommative quelconque, pourrait aussi se référer à un trait latent, en situant chaque élève sur la même échelle. Dans le cadre de la classe, cependant, les enseignants voient mal comment exploiter une telle évaluation globale. Ce n'est pas le fait qu'un élève soit en avance ou en retard, qui permet de savoir comment l'aider. La généralité d'un tel jugement, qui serait un avantage pour établir un bilan, est nuisible pour la formulation d'un diagnostic ou d'une proposition d'action.

C'est donc en restant très près du comportement observé, en réagissant à l'erreur elle-même, ou à l'incompréhension particulière qu'elle révèle, que les enseignants exploitent l'instrument d'évaluation, et non en effectuant une inférence statistique pour estimer un niveau probable sur une échelle.(...)

Noizet, G., Caverni J.P., *L'évaluation scolaire,* 1978
Bonniol, J-J. "Déterminants et mécanismes des comportements d'évaluation d'épreuves scolaires", Thèse, Université de Bordeaux II, 1981

CARDINET, J.

I 2 La docimologie
I 2 2. La psychologie de l'évaluation,

figure de l'évaluateur produite

A cause de la confusion avec le docimologue expérimental, l'image de l'évaluateur produite par la psychologie scolaire est celle d'un agi, utilisateur de normes implicites. Victime des conditions de passation des examens, de la mise en circulation des

critères pour la validation, la certification des acquis, l'évaluateur est encore dissimulé par l'examinateur.

Parce qu'il a perdu confiance en ses impulsions, ses intuitions, qu'il est maintenant conscient de la gravité de son acte, il devient "métrologue", il vise à une technicité rationnelle, une prise de distance par des procédures qui paraissent toujours trop compliquées : il est entravé par le désir irrépressible d'être juste.

GUIGOU, J. "Evaluation et institution éducative", *Education permanente*, n°9 1971, p.39/56 :

(...) face à la question centrale pour toute action éducative du "quoi" et du "comment" évaluer, s'organisent plusieurs types de réponses pédagogiques. (...)

Les réponses instrumentalistes et technicistes

Suivant scrupuleusement les développements les plus récents de la docimologie, l'évaluation instrumentaliste ne cherche qu'à affiner les dispositifs et les appareils de mesure des résultats d'un apprentissage, sans jamais s'interroger sur la validité "scientifique" du champ sur lequel porte son expérimentation.

Tirant hâtivement des conclusions des dernières recherches docimologiques le "pédagogue-métrologue" qui souhaite introduire toujours plus de rationalité dans l'évaluation, isole arbitrairement la production -ou la non-production- de l'examiné sans prendre en compte l'ensemble des déterminations institutionnelles de cette production. La neutralisation qu'il réalise là -et qu'il nomme volontiers une "opération objective"- constitue déjà une action de légitimation de la sélection sociale instituée par l'école.

Traiter les résultats d'une épreuve de contrôle de connaissances comme on traite une série statistique quelconque, c'est masquer les différentes positions de chacun à l'égard du savoir, c'est dissimuler les segmentarités instituées par cette nouvelle technologie qui consacre l'arbitraire culturel et pédagogique établi.

Transposées mécaniquement du laboratoire dans la salle de classe, de telles "améliorations" dans le traitement des résultats d'un apprentissage relèvent du même modèle d'action que celui de l'ergonomie qui, dans l'entreprise, multiplie les raffinements dans la mesure du "rendement humain". Comme l'organisateur qui, dans l'atelier, modèle toutes les activités selon la grille du "job évaluation", le formateur qui, souhaitant moderniser sa pédagogie, ne s'attache qu'à perfectionner ses outils de mesure, risque de rencontrer les mêmes déboires que son homologue industriel. Car si l'on connaît l'impuissance de l'analyse ergonomique à saisir la réalité des rapports de l'homme au travail et notamment des rapports de pouvoir issus des formes néo-capitalistes d'organisation du travail (management) -en toute honnêteté épistémologique- on pense apprécier la réalité du travail pédagogique aux résultats d'une analyse docimologique bien menée. (...)

GUIGOU, J.

PARISOT, J-C., "Le paradigme docimologique : un frein aux recherches sur l'évaluation pédagogique?", *L'évaluation en question*, CEPEC, 2°ed, Paris, ESF, 1988, p.37/56 :

(...) L'attention exclusive portée pendant longtemps à l'examen nous semble produire des effets qui, pour certains, sont pervers. Cette orientation a pu faire penser ou laisser penser qu'évaluation et "examination" sont des pratiques identiques relevant des mêmes méthodologies et des même logiques. En conséquence, les pratiques habituelles d'évaluation dans le système scolaire, souvent inspirées de l'examen (copies notées, récitations écrites notées, interrogations orales, contrôle du niveau), n'ont pas été réinterrogées suffisamment. L'évaluation après coup, sous forme de

constat et de bilan, reste une pratique répandue et considérée souvent comme la forme la plus achevée de l'évaluation. Par exemple, le poids accordé dans les livrets scolaires aux devoirs en temps limité ou autres tests collectifs risque d'induire des attitudes peu souhaitables chez les élèves comme chez les enseignants. Les élèves n'apprennent pas pour apprendre mais pour réussir un test : les enseignants consacrent beaucoup de temps et d'énergie à des corrections classiques de copies, sans efficacité pédagogique réelle.

Le paradigme que révèle la docimologie semble bien encore traverser les pratiques et représentations collectives aujourd'hui. Si de nouvelles recherches qui ont -heureusement- acquis droit de cité essayent d'articuler pratique de l'évaluation et augmentation de l'efficacité pédagogique, il nous semble parfois qu'elles ont de la peine à s'imposer et que ce fait n'est pas sans lien avec la subsistance du paradigme lui-même. Le potentiel de recherche sur l'évaluation a été longtemps distrait sur une problématique courte et peut-être, en quelque manière, égaré sur une fausse piste. Tout se passe comme si l'inertie acquise contrariait encore la réorientation des recherches encore embarrassées d'habitudes mentales et méthodologiques surannées.(...)

Conclusion

(...) Pierre Dominice(19) écrit : "La docimologie a fourni des techniques mais elle n'a pas réussi à modifier la fonction politique et sociale de l'évaluation. La rationalisation à laquelle elle est parvenue peut ainsi être considéré comme technocratique, c'est-à-dire qu'elle déplace sur des spécialistes de la mesure le pouvoir de discrimination jusqu'alors exercé par chaque pédagogue à l'intérieur de sa classe. Des instruments, tels que les tests objectifs ou les épreuves standardisées, changent la forme de l'évaluation mais ne transforment nullement sa raison d'être ou sa signification."

Or, précisément, il s'agit bien de changer la raison d'être ou la signification de l'acte d'évaluer ou encore sa fonction. Pour ce faire, il faut déplacer le paradigme docimologique dont les différents aspects sont interactifs et solidaires, se renforcent et se légitiment réciproquement. Récapitulons.

La docimologie s'est presque exclusivement intéressée à l'examen, c'est-à-dire aux procédures de tri et de sélection qui ont des finalités sociales et économiques. Ce faisant, elle a renforcé et légitimé "scientifiquement" une conception de l'évaluation détachée de l'acte d'éducation proprement dit. Croyant en une répartition aléatoire des dons et des aptitudes "naturelles", elle a imaginé les systèmes de mesure et de classement plus ou moins sophistiqués qui prenaient leur inspiration première du côté de "l'objectivité psychométrique" issue d'un modèle différentiel de la psychologie. Par effet de conséquence, la question des examens est devenue affaire d'experts spécialistes statisticiens, ou concepteurs d'épreuves standardisées. Ce paradigme tenace traverse encore bon nombre de représentations et de pratiques et constitue encore un frein puissant au changement. (...)

(19) P. Dominice, *La formation, enjeu de l'évaluation*, Québec, Peter Lang, 1979

PARISOT, J-C.,

I. 2.
Lectures complémentaires à propos du modèle docimologique

ABERNOT, Y. "Caractéristiques et difficultés de l'évaluation", *Les méthodes d'évaluation scolaire*, Bordas, 1988.

BARBIER J-M. *L'évaluation en formation,* PUF, 1985, cf. p. 41/45

BONBOIR, A. *La docimologie*, coll sup, PUF L'éducateur n° 38, Paris, 1972

DAUVISIS, M-C., "Des titres et des nombres en quête de valeurs : de la docimologie à l'évaluation", Colloque AFIRSE, de Carcassonne, *Les évaluations*, PUM, p. 113/135

DE BAL, R. & DE LANDSCHEERE, G. *Construire des échelles d'évaluation descriptives*, Min. Education nationale, Bruxelles,, 1976

DE LANDSHEERE, G. *Evaluation continue et examens, Précis de docimologie*, Nathan-Labor, Paris-Bruxelles, 1972

Johnson, S. « Evaluation de la comparabilité des notations entre jurys d'examens », *Mesure et évaluation en éducation*, Vol 12, n°1, 1989, p.5/22

RANJARD, P. *Les enseignants persécutés*, Paris, Robert Jauze, 1984

REUCHLIN, M. " La docimologie, effort d'explicitation ", *Les amis de Sèvres*, n°2, 1968

I. 3.
L'évaluation dans le modèle de la métrie : psychométrie et édumétrie

Champ d'étude :

Ce sont les tests qui sont étudiés, pour contrôler ce qu'ils mesurent. Le travail porte sur la vérification des mesures.

Il s'agit de rechercher la fidélité des mesures obtenues par les tests, de trouver des méthodes de fabrication des tests pour que leurs résultats soient fiables et généralisables.

Deux mouvements de recherches se sont succédés, que nous traiterons ici ensemble : l'étude des tests mesurant l'intelligence (le modèle local de la psychométrie) qu'on peut faire correspondre, en suivant Pelletier (1971), aux objets de "La première période qui débute au tournant du siècle et se prolonge vers les années 1920-1930 et que l'on nomme Testing period. Cette période est caractérisée par l'intention de remplacer les mesures subjectives, individuelles et aléatoires en usage dans les établissement scolaires par des tests standardisés et objectifs. Ces tests sont surtout des tests d'intelligence et de rendement (achievement)." (1) .

Puis le modèle local de l'étude des tests mesurant l'acquisition des programmes (édumétrie), correspondant, selon Pelletier encore, à "Une deuxième période, mal distinguée de la première, (est) appelée Measurement period où le concept de mesure vient nuancer ce que le concept de test pouvait avoir de trop limité. Dans cette période, on continue à perfectionner les batteries de tests, mais on s'inquiète davantage de l'utilisation des résultats des tests et de la difficulté de réaliser des mesures objectives en ce domaine."

(1) "La notion d'évaluation", *Education permanente*, n°9, p. 7/19

DE KETELE J-M., "Une première lignée de modèles", *L'évaluation : approche descriptive ou prescriptive ?*, Bruxelles, De Boeck, 1986, p. 248/252 :

(...) *1.4. L'ère psychométrique*

S'enracinant dans les travaux de Binet en France et dans le développement des nombreux travaux américains dans le domaine de la mesure, la psychométrie est venue au secours de l'évaluation pédagogique. Deux soucis majeurs animent ce modèle : améliorer la fidélité des mesures des performances en décomposant la variance totale en ces différentes composantes (dont la variance erreur) et tenter de construire des épreuves valides en recourant à des techniques comme l'analyse factorielle qui permet d'approcher ce que mesure véritablement un test. Ce mouvement trouvera son couronnement dans la théorie de la généralisabilité des scores et des profiles de Cronbach, Gleser, Nanda et Rajaratnan (1972).

1.5. L'approche édumétrique

De nombreux travaux, dont ceux de Livingstone (1972), avaient tenté d'adapter le modèle psychométrique aux exigences de la pédagogique de la maîtrise. En effet, celle-ci n'a pas pour but de maximiser la variance entre les sujets (comme dans le modèle psychométrique), mais de discriminer les objectifs atteints de ceux qui ne le sont pas sur la base des tests critériés. Cardinet et Tourneur (1974, 1985) ont alors tenté une extension de la théorie de la généralisabilité de Cronbach. Ces auteurs partent de l'idée que les données observées lors des évaluations peuvent être organisées selon trois dimensions principales : les personnes examinées, les moments de prise d'information (comme par exemple un prétest est un posttest) et les conditions d'observation (comme par exemple différents niveaux de questions). Il en résulte trois types d'épreuves fondamentalement différents (psychométrique, édumétrique et diagnostique) selon les plans de mesures adoptés.

L'approche édumétrique consiste à transposer la matrice des données et à traiter les questions comme on traitait en psychométrie les personnes. Avec ce plan de mesure (les personnes comme facette de généralisation et les questions comme facette de différenciation), on peut estimer le degré de fidélité d'un test. Dans cette optique, un test sera d'autant plus fidèle que les résultats des personnes pour une question seront assez semblables, la variance à l'intérieur d'une question étant faible par rapport à la variance entre questions. La théorie de la généralisabilité est un outil extrêmement puissant pour le chercheur car il permet de générer une infinité de plans de mesure et, de ce fait, offre des possibilités nombreuses de répondre à certains problèmes posés par l'évaluation. (...)

Cronbach, L-J, Gleser, G-C., Nanda, H., Rajaratnam, N. *The dependability of behavioral measurements : Theory of generalizability for scores and profiles*, New York, John Wiley, 2

Cardinet, J., Tourneur, Y. *Une théorie des tests pédagogiques*, Neuchâtel, INRDP, 1974

DE KETELE J-M.

DE LANDSHEERE, G. *La recherche expérimentale en éducation*, Unesco, Delachaux & Niestlé,1982, p. 68/69 :

(...) *Les tests critériels*

D'aucuns estiment qu'ils constituent la première innovation, en ordre d'importance, de ces dernières années (90). On sait que le principe de ces tests a été défini pour la première fois par R. Glaser (91) en 1963.

Le principe général est aujourd'hui bien connu. Un domaine dans lequel les items seront choisis est soigneusement défini. Ce domaine, qui constitue le "critère", doit rester inviolé et ce souci prime sur les autres considérations psychométriques. En particulier, il faut obtenir que le test soit fidèle, de difficulté appropriée, etc., sans que jamais le critère ne soit compromis.

Les tests critériels (qu'il serait plus exact d'appeler tests centrés sur les objectifs ou référés au domaine) sont encore loin d'avoir atteint un développement optimum. Presque tout reste à faire dans le domaine affectif. Par ailleurs, on ne dispose pas encore d'une théorie solide pour déterminer quelle partie du domaine doit être maîtrisée pour que la performance soit considérée comme (tout à fait) satisfaisante. C'est le problème du standard souhaité.

Un standard peut être fixé, empiriquement, de façon subjective ou, au contraire, en cherchant un critère dans le niveau moyen ou maximum d'une population cible donnée. On réintroduit ainsi un biais normatif, mais il paraît inévitable dans la mesure où le développement psychologique s'opère lui-même selon des stades ou des étapes critiques les unes par rapport aux autres (92)

Comment la maîtrise va-t-elle être définie ? Adopter, par exemple, un critère absolu de 70°/° ou de 80°/° de réussite n'a guère de sens. Il est d'abord clair que le critère ne peut être le même pour les apprentissages instrumentaux de base et pour les autres. Où va donc être placée la note de césure, en deçà de laquelle on considérera que la maîtrise n'est pas atteinte? Comme l'observe J. Keeves, les problèmes de maîtrise sont aussi rendus complexes par la question du transfert et du taux de rétention : "Si la maîtrise est exigée après un temps assez long, différents taux de rétention parmi les individus peuvent impliquer différentes notes de césure selon ces individus ou certains changements dans l'ordre des choses enseignées ou dans le moment du testing. (C'est pourquoi) il semble nécessaire de réaliser des études comptant des observations de performances en classe, des évaluations par des juges du coût des erreurs de classification, et des mesures de la performance à l'aide de tests, ces aspects étant évalués aussi indépendamment les uns des autres qu'il est possible." (93) (...)

(90) voir Gagné, R. *Educational research and development : past and future*, In : Glaser, R., Cooley, W., eds, *research and development and school change : learning research and development*, New York, Halsted Press, 1978

(91) Glaser, R. *Instructional technology and the measurement of learning outcomes : some questions*, American psychologist, vol 18, 1963, p. 519-521

(92) sur ces problèmes voir : - Hively, W. et al. *Domain-referenced curriculum evaluation : a technical handbook and a case study from the Minnemast project*. Los Angeles, University of California, 1973 - Harris, C.W. ed Problems in *criterion-referenced measurement*. Los Angeles, University of california, 1974

(93) *Australian council for educational research*. Forty-ninth annual report, 1978-79, Hawtorn, Vic, 1979, p. 52

DE LANDSHEERE, G.

CARDINET, J. *Les modèles de l'évaluation scolaire*, Neuchâtel, IRDP, 1986 :

(...) *2. Le modèle psychométrique*

Tout un ensemble de théories et de techniques ont été développées depuis le début du siècle, d'abord dans le domaine des aptitudes, dans le prolongement des

travaux de Binet, puis dans celui de la mesure des acquisitions scolaires, au point que pour certains chercheurs anglosaxons, les seules méthodes d'évaluation pédagogique valables seraient de ce type.

2.1. Le souci de base : celui de la fidélité des mesures

Toute l'approche psychométrique repose sur l'idée qu'un score observé est la résultante d'effets indépendants, dus à la capacité de l'élève, au niveau de difficulté de l'épreuve, au degré de sévérité de l'expérimentateur, etc., et à l'interaction de ces effets principaux. Diverses techniques statistiques permettent d'analyser et de mesurer l'importance relative de chacun de ces effets. Si la proportion de la variance des scores observés qui est due à des sources de variation considérées comme acceptables dépasse 80%, on considère d'habitude la mesure comme satisfaisante. Si elle est inférieure à 80%, on doit chercher à améliorer le dispositif d'observation.

Cette possibilité d'évaluer la mesure est un outil puissant, qui permet de départager les idées valables des projets sans fondement, lorsqu'on parle de modifier le système des notes pour en améliorer la précision. On a pu montrer, par exemple qu'on ne gagnait rien à multiplier les échelons (en introduisant des demi-points ou des quarts de points), mais par contre qu'on avait avantage à analyser les divers aspects d'une production complexe, pour les juger séparément, et à totaliser ces différents jugements à l'aide d'un jeu de coefficients explicite.

Une application classique de l'approche psychométrique est la "modération" des notes données par un instituteur aux élèves de sa classe. Il est indiscutable qu'un maître qui suit jour après jour le travail de ses élèves peut situer leurs performances relatives dans une discipline avec beaucoup de sécurité. Par contre, il manque totalement de points de repère pour situer le niveau moyen de sa classe, et pour apprécier la dispersion des résultats qu'il observe. C'est sur ce point que des épreuves communes de niveau peuvent l'aider à ancrer son jugement par rapport à une échelle de performance générale, pour le rendre plus objectif.

2.2. Les développements psychométriques

Le niveau de sophistication des techniques psychométriques actuelles est difficile à imaginer par les non-spécialistes, qui risquent de rejeter pour des raisons futiles (objection à la forme de réponse à choix multiple, par exemple) des moyens d'action extraordinairement puissants.

Ainsi, il est difficile de proposer un meilleur moyen que l'analyse factorielle pour déterminer ce que mesure réellement un ensemble d'épreuves. Cette technique permet de résumer en quelques variables composites toute la variance commune à ces épreuves, et d'identifier la nature psychologique des facteurs communs ainsi déterminés.

On aurait tort également de négliger l'apport de la théorie de la généralisabilité à la construction d'épreuves critériées, mesurant l'écart entre la performance observée d'un élève et la performance attendue de lui, en fonction de l'objectif pédagogique visé.

Il faut citer enfin les apports possibles de la théorie des traits latents, qui permet de constituer des banques de questions aux propriétés bien contrôlées. On peut alors situer sur une échelle commune les résultats obtenus à des questions différentes, et ainsi comparer par exemple les moyennes relatives à des curriculums distincts, ou à des groupes d'âges successifs, passant des épreuves différentes, ou bien construire des instruments sur mesure, adaptés de façon optimale au niveau de chaque élève. Ces banques d'items sont déjà opérationnelles dans les pays anglo-saxons.(...)

CARDINET, J.

I, 3. L'évaluation confondue avec la métrie,

point de vue des détracteurs

Ce modèle de la métrie ne fit plus l'objet du consensus quand en corollaire se perdit la pérennité de la seule "intelligence géométrique" (1). L'évaluateur avait pris le point de vue du psychologue épistémique et ontogénétique (2), il travaillait dans l'Essence de l'homme, dans le paradigme (perdu) de la Nature Humaine (3), au détriment de l'intelligence sociale (1).

Préoccupé par les lois générales caractérisant les facultés de l'Homme cet évaluateur privilégie l'invisible et l'intemporel, il est au service d'une explication (4) des phénomènes. Technicien, il se donne comme tâche de construire les bons outils de mesure et l'opérationnel, la pratique évaluative n'est qu'un matériau opaque qu'il doit traverser.

(1) Oléron, P. *Savoirs et savoir-faire psychologiques chez l'enfant*, Liège, Mardaga, 1981
(2) dont Piaget (première manière) est la figure tutélaire
(3) Morin E. *Le paradigme perdu : la nature humaine*, Seuil, 1973
(4) Ardoino J & Berger, G. "Fondements de l'évaluation et démarche critique", *ACSE* n°6, 1989, p.3/11

CARDINET, J. *Les modèles de l'évaluation scolaire*, Neuchâtel, IRDP, 1986 :

(...) Les outils remarquables que le modèle psychométrique peut construire sont, comme les balances de précision, des instruments coûteux et fragiles. Il faut des milliers d'élèves, des centaines de questions, de gros ordinateurs et des techniciens très spécialisés pour mettre au point ces épreuves. Il faut donc que les investissements en recherche qu'ils nécessitent puissent être rentabilisés par un emploi suffisant de ces banques d'items. Or ceci paraît problématique dans les pays de langue française, vu l'étroitesse des marchés nationaux et l'instabilité des curriculums dont la transformation peut rapidement invalider les questions. (...)

CARDINET, J.

Le syntagme de l'édumétrie inaugure une longue période d'enchevêtrement des problématiques d'évaluation et d'apprentissage. C'est le début de la réduction de l'évaluation à la mesure de l'acquisition des programme de formation. L'évaluation consiste à vérifier l'intégration des contenus de formation et à communiquer cette vérification sous la forme d'une mesure objective. La valeur se quantifie, tel est le postulat. L'évaluateur devient le testeur des choses acquises dans le curriculum. L'évaluateur doit répondre à la question : "Qu'ont-ils appris ?". L'évaluation n'est qu'un acte de mesure de l'apprentissage.

Le "jugement de valeur" (qui n'est, somme toute, qu'une des exploitations possibles des résultats de l'évaluation) est présenté alors comme la caractéristique de l'évaluation (1). Le jugement n'est ici que cette vérification (des mesures de l'acquis) présentée comme indubitable, vraie, scientifique, puisque précédée de procédures sophistiquées de recueil et de traitement d'informations. De la

mesure comme procédure expérimentale (et donc scientifique), on passe, par la confusion entre évaluation et jugement, ainsi qu'entre échelle de valeur et échelle de mesure, à l'idéologie du contrôle (2).

(1) Barbier, J-M *L'évaluation en formation*, PUF, 1985
(2) Ardoino, J., "L'approche mutiréférentielle (plurielle) des situations éducatives et formatives", *Pratiques de formation*, avril 1993, p. 16/34

PELLETIER, L. "La notion d'évaluation", *Education permanente*, n°9, 1971 p.7/19 :

(...) Une troisième période appelée Evaluation period commence vers 1930. (...) Dans cette dernière période, on s'applique particulièrement à bien définir le concept d'évaluation et à le distinguer du concept de mesure. Le concept de mesure s'applique à la cueillette de données sur une ou plusieurs dimensions de l'objet étudié à l'aide d'un instrument approprié. Le concept d'évaluation par contre réfère au jugement subjectif ou à l'interprétation que l'on fait de la qualité ou de la valeur de l'objet étudié (1).

Ce qui différencie donc fondamentalement l'évaluation de la mesure, c'est le jugement de valeur (Abmann (2) , la référence à un critère permettant d'interpréter une mesure pour lui donner une valeur dans un contexte vital. Il arrive évidemment que des instruments de mesure soient assez connus pour qu'ils contiennent implicitement un jugement de valeur, mais les deux moments demeurent distincts. Ainsi, établir que tel individu a un quotient intellectuel de 140, n'est de fait qu'une mesure, mais identifier que cet individu de 140 a une intelligence très supérieure est une évaluation. (...)

(1) John A. Green *Introduction to measurement and evaluation*, New York, Dodd Mead & Co, 1970, p. 13
(2) Abmann Stanley & Glock M.D., Evaluating pupil growth *Principles of test and measurement*, Boston, Allyn and Bacon, 1967 cf. aussi Adams, Georgia Sachs, *Measurement and evaluation*, N.Y., Holt, Rinebart & Winston, 1964, p. 5

PELLETIER, L.

I, 3. L'évaluation confondue avec la métrie,

figure de l'évaluateur

La métrie a alimenté l'imaginaire de l'évaluateur en produisant l'image de l'arpenteur. L'évaluation a spatialisé ses objets, elle entre dans l'univers du comptable donc du contrôlable. L'évaluateur est un vérificateur de la mesure adéquate. Par l'excessive importance accordée au jugement qu'il devrait prononcer, l'évaluateur se donne comme celui qui doit effectuer la pesée des âmes, il devient le Grand Juge, gardien du sens.

ARDOINO, J. "Au filigrane d'un discours : la question du contrôle et de l'évaluation", préface de MORIN, M. *L'imaginaire dans l'éducation permanente*, Gauthier-Villars, 1976, p. I /XXXIX :

(...) I. Le mot contrôle(1) est, d'abord, un terme d'administration et de comptabilité. Etymologiquement, il vient de contre-rôle(2) et se trouve, ainsi, au fil des avatars du langage, le produit d'une superposition syllabique fréquente.

Les idées de copie conforme, de vérification(3) pour le compte d'un autre (à la limite pour soi-même), ou de rendre des comptes, s'y trouvent étroitement attachées, qu'il s'agisse de la tenue des comptes, de battre la monnaie, ou des actes de justice. On passera ainsi de l'idée de registre tenu en double pour la vérification (par un autre) à la procédure de vérification proprement dite (la preuve par neuf en arithmétique, la comptabilité en partie double, etc.). La vérification est alors l'opération matérielle ou mentale qui a pour but d'établir la conformité entre une assertion (déclaration, affirmation, négation, supposition, hypothèse) et ce à quoi elle est censée se rapporter.

Le fait de devoir rendre compte à quelque autorité, à laquelle on se réfère, ou qu'on révère, est également important (ce serait, selon le cas, un mandataire, un commanditaire, un chef, un Dieu, la norme, la loi ou la réglementation, l'état, la volonté générale, le consensus populaire, le bon sens, la conscience morale, les données de l'expérience, l'authenticité, etc.). Comme l'étymologie nous le laissait déjà pressentir, le concept de vérification nous renvoie, plus profondément, encore, à la notion de vérité.

(...) Parallèlement, les représentations que nous nous donnons de la vérité évoluent en fonction des contextes culturels et des systèmes de connaissance.

Primitivement la vérité est un "principe certain"(4) qui ne saurait être mis en doute parce qu'il se fonde sur une conformité indiscutable (mais supposée) entre la connaissance du sujet et la réalité de l'objet connu. Cette définition ontologique de la vérité correspond à une théorie explicite de la connaissance(5) (réalisme). Elle affirme déjà une position critique s'opposant à une conception plus archaïque de la connaissance fruit de la révélation(6) . (...)

On est ainsi passé de la notion d'une vérité indiscutable, qui ne saurait être mise en doute, ou tirant sa perfection même de son caractère indémontrable, parce que lié à la reconnaissance de l'évidence(7) à une vérité qui n'est que le produit perpétuellement élaboré et remis en cause d'une discussion critique interminable. Mais, bien entendu, cette discussion a toujours lieu selon des règles précises et dans le cadre d'un champ bien délimité. Que la vérité scientifique se fonde sur la référence à une réalité extérieure dont elle serait le reflet approximatif, ou sur le mythe méthodologique de l'objectivité positiviste, ou encore sur le principe de la cohérence interne du discours, conséquence des lois même de la pensée, il y a partout l'aveuglement spécifique, l'occultation de principe, que ces règles et ces limites constituent toujours pour le type d'analyse ou de recherche qu'elles fondent. (...)

Compte tenu de cet héritage, la dénotation (ou l'extension) du concept contrôle s'est considérablement élargie. Le développement considérable de l'Administration, de l'organisation et de la technologie, dans les civilisations industrielles avancées, contribuent à la généralisation du terme. On parle couramment, aujourd'hui, de contrôle policier, douanier, fiscal, sanitaire, financier, de contrôle du crédit(8) , de contrôle social(9), de contrôle des naissances(10), de contrôle d'une expérience, de contrôle de fabrication, de contrôle de gestion, de contrôle des connaissances, de contrôle aérien au sol, etc.; etc. Paradoxalement la connotation (ou la compréhension) s'est, de son côté, tout à la fois, précisée (jusqu'à devenir, dans certains milieux, synonyme d'oppression ou de répression) et enrichie. Se retrouvent associées dans une constellation plus large, des notions telles que : enquête (policière, judiciaire, administrative mais aussi étude d'opinion, sondage), inspection, examen (au double sens de contrôle des connaissances et d'examen critique) analyse, validation, et, bien sûr, comme nous le verrons plus loin, évaluation. Mais sous-jacent à cet ensemble de notions, à leurs différents emplois, comme aux nuances de leurs significations respectives, c'est toute une conception du contrôle, une

philosophie pratique, une théorie implicite(11), justement marquée par son besoin de cohérence, qui s'affirme à travers notre histoire.

Parce qu'originellement contrôler c'est avérer(12), la conformité à la norme est l'intentionnalité la plus ancienne, on pourrait dire archaïque, de toute opération de contrôle. Ainsi plusieurs caractéristiques du contrôle que, jusqu'ici, nous percevions plus ou moins confusément, à la fois ressenties comme éparses et cependant liées entre elles, sans toujours bien en comprendre la raison commune, peuvent s'éclairer à une lecture seconde plus perspicace.

- Le contrôle est normatif. Il implique, c'est-à-dire qu'il suppose, d'abord, et imprime, ensuite, le respect des règles supposées bonnes parce que transcendantes, indiscutables, naturelles, établies par la sagesse de l'usage, ou réglementaires, c'est-à-dire imposées par une puissance supérieure, ou tirant son autorité de la souveraineté populaire par le biais d'une délégation de pouvoir.

- Cette normativité est, avant tout, logique. Elle assume l'héritage aristotélicien : substantialiste et catégoriel fondé sur le principe de non-contradiction(13). C'est pourquoi le contrôle sera longtemps plus analytique que synthétique. Il cherchera à isoler des dimensions pour les mesurer (ou il opérera des prélèvements dans une optique probabiliste). Lorsqu'on étudiera, dans la perspective de certains économistes de l'entreprise ou d'une science de l'organisation, la relation privilégiée entre la conception du contrôle et les mécanismes de la prise de décision on pourra parler du modèle prégnant d'un déterminisme linéaire(14) ou, ce qui revient au même, d'une mono-rationalité de type cartésien(15). Mais dans cette conception très inspirée d'une vision du monde platonicienne la normativité prend inévitablement un caractère moral. Il existe nécessairement des affinités essentielles entre le Vrai et le Bien. Ce qui est normal est meilleur que ce qui est pathologique.

- Dans la pratique traditionnelle, le contrôle est dévolu à la hiérarchie. Comme Fayolle l'a mis en évidence le contrôle est l'une des fonctions du commandement. La qualité de l'opération de contrôle suppose donc un quasi-racisme, ou si l'on préfère un distinction de type essentialiste entre les contrôleurs et les contrôlés. Ainsi les chefs contrôleront la qualité du travail des subordonnés, les maîtres les connaissances des élèves, les parents la croissance et l'éducation des enfants, l'Etat, la police et les tribunaux l'ordre public, etc. L'inspecteur, l'examinateur, l'enquêteur exerceront ainsi leur mission de police sociale.

- Il en découle tout naturellement un caractère sanctionnant du contrôle. La recherche de la vérité dans sa conception normative, et même au-delà, se montre incapable de distinguer entre ses implications et ses conséquences logiques et morales. Le vrai et le faux se laissent mal séparer du bien et du mal. C'est pourquoi le contrôle est un procès nécessitant pour sa tenue en bon ordre la mise en oeuvre de procédures. Ce juridisme qu'on retrouve partout, implicite ou explicite, jusqu'à travers les jurys d'examens, entraîne à son tour le caractère contentieux(16) et formel qui, par la perte du sens, va prendre une signification essentiellement répressive. S'il y a bien, à partir du constat(17) d'un écart à la règle, recherche d'une réparation, c'est d'une compensation(18) (au sens juridique et moral du terme) qu'il s'agit.

-Ce type de contrôle est hors le temps. Dans la mesure où la vérité est, sinon éternelle, suffisamment établie pour être durable, le temps n'est pas une dimension prise en considération si ce n'est à une très large échelle qui n'intéresse pas l'opération de contrôle proprement dite mais ses fondements lointains, généralement perdus de vue. Quand on parle, malgré tout, du temps c'est de temps logique qu'il s'agit. Il y a, bien sûr, un avant et un après. L'ici et le maintenant sont analysés, contrôlés en fonction des vérités établies dans le passé. Mais le contrôle est toujours a priori ou a posteriori, quand il n'est pas encore temps, ou quand il est trop tard pour améliorer, éventuellement, les résultats par une réintégration de la lecture des écarts dans un

processus (et non plus un procès) dynamique(19). En ce sens la conception traditionnelle du contrôle que nous essayons de traduire ici consacre le triomphe de la vision du monde de Parménide sur celle d'Héraclite. Si malgré tous ses efforts pour s'en dégager l'éducation permanente n'y est pas parvenue, il faut aussi lire sa dénomination comme éducation à la permanence, à ce qui demeure inchangé(20) .

- Dans la mesure où il est normatif, hiérarchique, répressif, sanctionnant, policier, contentieux, tout contrôle est nécessairement politique. Celui-ci est résolument conservateur. La "fausse conscience"(21) qui le fonde aboutit à une "déchéance de la temporalité" dont l'établissement(22) seul tire son profit. Ainsi l'institution du contrôle a pour visée de permettre le maintien des institutions en place à travers une reproduction fidèle. A propos des conditions épistémologiques de production de la connaissance comme pour les conditions économiques et idéologiques de fonctionnement social, la fonction du contrôle est l'affirmation d'une cohérence, la victoire de l'ordre sur le désordre et l'incertitude. Qu'il s'agisse d'établir la conformité ou la comptabilité entre une supposition et des faits réels, entre une prévision, un diagnostic, un pronostic et l'échéance, la réalisation d'un événement, qu'il s'agisse de contrôler une situation mouvante, au sens de "l'avoir bien en main", l'intention est d'assurer le maintien de l'ordre ou de limiter les conséquences de l'entropie. Surtout quand il prend la forme lénifiante et manageriale, se voulant opératoire et lié à l'efficacité, ce type de contrôle aboutissant à un quadrillage organisationnel(23) est le discours même du pouvoir.(...)

II. Notamment dans les sciences humaines, avec le développement des démarches cliniques(24), les pratiques de l'évaluation(25) tendent à se multiplier voulant à l'évidence renier l'héritage sémantique, complexe et contradictoire, attaché à la notion de contrôle. Il en résultera une polysémie dont certains s'empresseront de souligner l'ambiguïté, alors que d'autres y verront justement la richesse d'un "carrefour sémantique"(26). Nous pouvons retenir des acceptions les plus générales du terme, comme de l'étymologie, que l'évaluation cherche à déterminer (précisément ou approximativement) la valeur, et que, lorsqu'il s'agit d'apprécier une quantité, c'est par une estimation, c'est-à-dire par une autre méthode que par la mesure directe. Effectivement dans un emploi déjà plus spécifique intéressant les sciences de l'éducation chez les auteurs nords-américains(27) l'évaluation se distingue des méthodes métriques (testing period and measurement period) auxquelles elle finit par s'opposer, celles-ci, plus analytiques, cherchant à isoler, pour les mesurer, des dimensions, des facteurs, des paramètres, celle là privilégiant une approche synthétique, voire interprétative, pour tenter de comprendre l'individu saisi dans sa totalité.(...)

1 Littré : -1 registre en double qu'on tient pour la vérification d'un autre. Autrefois, particulièrement, registre double qu'on tenait des expéditions, des actes de finance, de justice. --2 vérification administrative --3 (fig.) examen, censure --4 état nominatif des personnes qui appartiennent à un corps, etc...
2 XIV° siècle : registre (rôle) tenu en double pour la vérification d'un autre (1367). D'où contrôler proprement, porter sur le registre en double dit contrôle.
3 Le fait de vérifier, opération par laquelle on vérifie. On retrouve dans vérifier (1369), dérivé du bas latin verificare (de verus : vrai et facere : faire) les sens de enregistrer, homologuer. Le Robert : 1) examiner la valeur par une confrontation avec les faits ou par un contrôle interne--2) examiner (une chose) de manière à pouvoir établir si elle est conforme à ce qu'elle doit être, si elle fonctionne correctement--3) reconnaître une chose pour vraie par l'examen, l'expérience...
4 Littré : sens 5

5 "Les définitions" du vrai qui ont été produites dans l'histoire de la philosophie se réfèrent soit au rapport de l'idée de vérité à l'idée d'objet pensé (par exemple l'adaequatio rei et intellectus) soit au rapport de cette idée à celle du sujet pensant (conformité de l'esprit à ses lois) ; soit aux moyens de discerner le vrai du faux (consentement universel ; théorie de l'évidence ; théorie de la "convention" ; définition pragmatiste de la vérité par le succès ; critérium de la convergence intellectuelle, etc...). Elles constituent donc des hypothèses philosophiques sur la théorie de la connaissance, ou même, sur l'épistémologie proprement dite, et non des définitions à proprement parler." Lachelier in Lalande, Vocabulaire philosophique, op. cit.
6 Révélation : Littré ; sens 1 : "tirer comme de dessous un voile, faire savoir ce qui était inconnu et secret" ; sens 2 :"particulièrement, il se dit de l'inspiration par laquelle Dieu fait connaître".
7 "Toutes les vérités (axiomes) ne peuvent se démontrer ; et cependant ce sont les fondements et les principes de la géométrie ; mais, comme la cause qui les rend incapables de démonstration n'est pas leur obscurité mais au contraire leur extrême évidence, ce manque de preuve n'est pas un défaut, mais plutôt une perfection." Pascal, Pensées, XXIV 12.
8 "Peut être à la fois qualitatif, selon qu'on désire faire une sélection de ceux qui auront droit au crédit ou qu'on désire appliquer des mesures instructives à l'ensemble des branches de l'économie." Dictionnaire des sciences économiques, P.U.F., Paris, 1956, p. 296.
9 "Le maximum de liberté individuelle apparaît comme résultant du maximum de conformisme social... Le contrôle social est celui qui est exercé par l'entité que nous appelons société." Dictionnaire des sciences économiques, P.U.F., Paris, 1956, p.296.
10 "Plus connu, même en France, sous le terme de Birth contrôle devrait être dénommé dirigisme des naissances, et se rattache du point de vue économique aux théories malthusiennes et néo-malthusiennes..." Dictionnaire des sciences économiques, op. cit., p. 297.
11 Il faudrait parler ici d'"assumptions" au sens où en anglais D. Mac Gregor emploi ce terme. Cf. La dimension humaine de l'entreprise et Le manager professionnel, D. Mac Gregor, Gauthier Villars, 1970 et 1974.
12 Littré : ancien français : voir vrai, du latin verus ; avoir, donné la certitude qu'une chose est vraie ; adj. avéré : établi comme vrai.
13 Cf. J. Ardoino, Propos actuels sur l'éducation, op. cit., 1re partie (1963)
14 Idem (1963)
15 Cf. L. Sfez, Science des organisations et changement social, Projet, Paris, 1973 et également, Critique de la décision, Armand Colin, Paris, 1973.
16 S'oppose, ici, à gracieux, gratuit.
17 Dans cet univers formel et procédurier les formes les plus usuelles de contrôle sont l'accusé de réception (du type, je vous reçois bien ou mal, je vous reçois 1/5 ou 5/5-contrôle des transmissions), le collationnement (cf. le monologue d'Yves Montand, Le téléphone) et le constat (avec ses références à "l'huissier" et au "gendarme" - Procès verbal est aussi synonyme de constat). Cf. J. Ardoino, Information et communication (1961), Ed. d'Organisations, Paris, 1964, pp. 36 et ss.
18 Par exemple le vergeld germanique, la composition pécuniaire, la peine en droit pénal.
19 La prise en considération d'un monde scolaire illustre assez bien ces énoncés plus généraux.
20 Du latin : (permanere = durer) ce qui reste
21 Cf. J. Gabel, La fausse conscience, Ed. Minuit, Paris, 1962.
22 Au sens de établie, société établie.
23 Cf. Y. Stourdze, Organisation antiorganisation, Repères Mame, Paris 1973.
24 Chez les rogériens toutefois, où elle est pratiquement équivalente au jugement de valeur par autrui et s'oppose ainsi à compréhension, la notion d'évaluation sera longtemps proscrite. "... une évaluation faite par autrui ne saurait me servir de guide." C. Rogers in Le développement de la personne, Dunod, Paris 1966, p. 21
25 Evaluer : Littré : 1) estimer la valeur, le prix d'une chose... 2) fixer approximativement une quantité...

26 L. Pelletier, "La notion d'évaluation", *Education Permanente* n° 9, janv.-mars 1971, p.7/19.
27 Cf. notamment Greene and Jorgensein, Measurement and evaluation in the modernschool, New York 1962 ; Gronlund, Norman E., Measurement and evaluation in teaching, New York, Mac Millan 1965 ; Lindvall, Testing and evaluation, An introduction, Hartcourt & Brace 1961 ; John A. Green, Introduction to measurement and evaluation, New York, Dodd, Mead & Co 1970 ; Ahmann Stanley & Glock M. D., Evaluating pupil growth principles of test and measurement, Boston, Allyn and Bacon 1967 ; Adams, Georgia Sachs, Measurement and evaluation, New York, Holt, Rinehart & Winston, 1964.

ARDOINO, J.

VIAL M. *Instrumenter l'auto-évaluation - Contribution à la pensée complexe des faits d'éducation*, Thèse de l'Université de Provence en Sciences de l'éducation, Aix-Marseille I, 1991

(...) l'évaluateur lorsqu'il se situe dans la logique du bilan (...) définira l'évaluation comme "un temps d'arrêt où l'on s'interroge sur l'action conduite afin d'améliorer cette conduite" (HADJI, 1989, p 144). (...) L'arrêt définit le bilan, lequel débouche sur un jugement : l'évaluateur devient ce Juge qui "s'interroge sur la valeurs des transformations opérées, qu'il veut pouvoir apprécier" (p.144). Il est seul à bord, décideur de la validité, dirions-nous plus que de la valeur. Il est le garant de la conformité.

Mais l'Interprète que Hadji (1) appelle le Philosophe, est tout autant seul et dans le bilan : "il s'interroge sur le sens de ce qui s'est passé, qu'il veut pouvoir faire émerger" (p.144). Il " construit un système d'interprétation" (p.144). Il est en fait un Traducteur, "il lui faut- c'est là son impératif catégorique- traduire le monde tel qu'il est" (p.143) (Donc, il sait le monde ?) . Il construit un texte rendant intelligible la réalité, il se donne pour mission de "comprendre ce qui se passe et de dégager la signification du travail réalisé" (p.158). Il se situe "dans le cadre d'une gestion du probable"(p.144), ce qui est une bonne définition d'un espace de contrôle qui pour être convivial, n'en est pas moins encore du contrôle. Ce Philosophe qui est, en définitive, un gentil vérificateur de la présence du sens, "construit une grille pour juger un réseau de significations"(p.144). S'il veut comprendre, c'est pour prendre en compte. Ou bien, "faisant preuve de retenue et de modestie"(p.144), il risque de n'être qu'une des figures du modèle charismatique de l'acte éducatif où "le maître accède à la dignité d'éducateur lorsqu'il comprend et fait comprendre que le savoir qu'il transmet n'est pas l'objet réel de la communication établie avec l'élève", ce "maître d'humanité" dont l'idéal est Socrate qui "enseigne non pas ce qu'il sait mais ce qu'il est" parce qu'il réalise "la synthèse des dons, des vertus" dont l'élève doit s'imprégner." (FERRY, (2) 1970, p. 8/9). Bien sûr, ce maître est déjà moins du côté du contrôle que le Gourou qui ne vise qu'à se faire des adeptes, mais il travaille encore à remplir le vide.

Il nous semble qu'on confond là le Traducteur, l'Interprète avec le Philosophe, ou bien le philosophe aurait intérêt à céder sa place au Sage. En effet, cette parole interprétative, telle que la conçoit Hadji, est encore un plein qu'il va donner à l'autre. Le philosophe "se prononce sur la réalité" (143). Nous le trouvons bien bavard, ce philosophe-là : il s'autorise du Sens. Cet Interprète (qui plus est, il a l'air de vouloir donner un texte intelligible, la pythie au moins laissait aux hommes l'énigme à déchiffrer), n'est pas, nous semble-t-il, la meilleure figure du philosophe. Il pratique au mieux une maïeutique qui, chacun sait, risque de n'être que la construction guidée d'une réponse que le maître possède. Il risque de vouloir appareiller, bureaucratiser le rapport au savoir et faire perdre de vue le "défi de la complexité".(...)

L'évaluateur doit pouvoir passer du rôle de "Gardien du sens" (HADJI, 1989 p.143) (donc d'un sens déjà construit, par exemple par l'Institution) quand il est dans la logique du bilan, au rôle d'Initiateur (qui fait découvrir à l'autre le sens, qui lui permet de faire émerger du sens) dans la logique formative.(...)

(1) Hadji, C. *L'évaluation, règles du jeu*, Paris, ESF, 1989
(2) FERRY, G. *La pratique du travail en groupe*, Paris, Dunod, 1970

VIAL M.

I. 3. La métrie : lectures complémentaires

CARDINET, J. "L'apport de la théorie de la généralisabilité à l'évaluation sommative individualisée" *Evaluation scolaire et mesure,* De Boeck, Bruxelles, 1986, p.119/ 140

CARDINET, J "Remettre le quantitatif à sa place en évaluation scolaire", *Les nouvelles formes de la recherche en éducation,* colloque AFIRSE Alençon 24/26 mai 1990, Matrice Andsha, 1990, p. 58/66

AUGER, R. & DASSA, C. , "Les pratiques de mesure et d'évaluation des apprentissages et la validité des tests", Laveault, D. *Les pratiques d'évaluation en éducation*, Montréal, 1992, p. 151/165

Tourneur, Y Les domaines d'application de la théorie de la générabilité, *Mesure et évaluation*, vol 12, N°2/3, 1989, p. 53/68

I.
L'évaluation comme mesure : vision d'ensemble

Les modèles de l'explication causale, de la docimologie et de la métrie veulent, au travers de l'acte d'évaluation, armer l'évaluateur pour qu'il puisse affirmer les résultats de ses évaluations : ses jugements. On travaille donc à construire ses outils, des outils fiables qui permettront de "réellement" mesurer. Par là, on construit la figure emblématique de l'évaluateur objectif, externe, expert. "On donnait à l'évaluation une centration instrumentale"(1). Se propage l'image de l'évaluateur détenteur des attributs du juge, porteur de la parole vraie, parce qu'assuré de ses techniques scientifiques -- un remake du jugement dernier.

C'est pourquoi on peut regrouper ces trois modèles dans la matrice théorique de la mesure : on voit s'établir avec la force d'une réduction(2), d'une exclusive, la confusion entre évaluer et mesurer. L'évaluation ne serait qu'une méthodologie de la mesure. "On limitait l'évaluation à ces variables pour lesquelles la science de la mesure avait déjà élaboré avec succès des instruments. Les autres variables, on les qualifiait "d'indépendantes", ce qui revenait à dire qu'on ne pouvait

les mesurer, et partant, qu'elles n'avaient aucune utilité et, ultimement, aucune importance" (1).

Il ne s'agit pas, aujourd'hui, de "rejeter la mesure" mais d'éviter d'importer en évaluation, comme si cela allait de soi, son idéologie positiviste ou une méthodologie empruntée aux Sciences de la nature qui se présentent comme incontournables et seules garantes de la scientificité. "En bref, cette définition donne lieu à une évaluation dont l'objet est trop restreint et dont l'approche est trop mécaniste." (1)

1 Stufflebeam, D et all. *L'évaluation en éducation et la prise de décision*, NHP, Montréal, 1980
2 Morin, E., *La méthode* (quatre tomes)

DE KETELE, J-M. & ROEGIERS, X. "Le recueil d'informations, l'évaluation, le contrôle, la mesure, la recherche : serviteurs et maîtres", Colloque AFIRSE de Carcassonne 1991, *Les évaluations*, PUM, 1992, p. 142/ 161 :

(...) *La mesure*

1. Définition

Mesurer est un processus par lequel on assigne des nombres à des choses selon des règles déterminées.

Pour qu'il y ait mesure, plusieurs conditions doivent être réunies :

1. La variable à mesurer doit être clairement définie. Une variable est une quantité ou une qualité susceptible de fluctuation. Le quotient intellectuel est une variable car elle peut revêtir une infinité de fluctuations ou de modalités comprises entre 0 et plus ou moins 200 : il s'agit d'une *variable continue*. Par contre, le nombre d'accidents à un carrefour est une *variable discrète*, car elle ne peut revêtir qu'un nombre fini de modalité : en effet, on ne peut parler de deux accidents et demi.

2. *Les modalités de la variable* à mesurer doivent être clairement définies et celui qui mesure doit savoir comment faire correspondre un chiffre à une modalité observée.

Dans les sciences humaines, le problème de la mesure est un problème délicat auquel il faut accorder beaucoup d'attention.

Quand un physicien parle de mesure, il pense immédiatement au fait d'assigner des nombres aux observations de telle sorte que ceux-ci lui permettent par l'intermédiaire de manipulations et d'opérations, selon certaines règles, d'acquérir une nouvelle information. En physique, la mesure ne pose pas beaucoup de problèmes, car la relation entre les choses observées et les nombres qui leur sont assignés est très directe. Ainsi, par exemple, si vous prenez une latte de fer pesant 3 kg, en principe homogène, et que vous la coupez en deux, il vous suffit de connaître le poids de la moitié de la latte, de diviser 3 kg par deux. Il s'en suit une relation directe entre la division de la latte et la division du nombre.

En est-il de même en psychologie et en pédagogie ?

Absolument pas. Si le psychologue orienteur, le pédagogue ou le psychologue social agissaient comme le physicien avec les résultats d'un test d'intelligence ou une note d'examen ou de degré d'une attitude mesurée, ils s'exposeraient à des dangers lourds de conséquences.

Aussi importe-t-il de distinguer en sciences humaines différents types de mesures (nominales, ordinales, d'intervalles égaux et de rapport), et de bien connaître les opérations permises pour chacune d'elles.

2. Le processus de recueil d'informations par rapport au processus de mesure

Le recueil d'informations est un processus qui suppose un objectif organisateur, la définition d'une stratégie, la collecte proprement dite des informations et le codage des informations sélectionnées.

On peut identifier quatre catégories d'informations selon le type de codage qu'elles permettent :

1. Une première catégorie d'informations reprend l'information qui est naturellement codée sous la forme d'une mesure (ordinale le plus souvent). C'est le cas du recueil de données de type numériques : chiffre d'affaires, nombre de personnes employées dans tel service, nombre de jours d'absence, résultats obtenus par tel étudiant dans telle branche, âge, etc.

2. Une deuxième catégorie reprend les informations qui peuvent être codées directement sous forme d'une mesure (nominale ou ordinale) lorsque les variables à observer et leurs modalités on été définies au préalable.

Exemples :

- codage des niveaux hiérarchiques dans une entreprise : directeur général 1, cadre supérieur 2, cadre moyen 3, employé 4, personnel de maîtrise 5, ouvrier 6 (variable ordinale) ;
- codage des élèves selon leur pays d'origine : Algérie 1, Belgique 2, France 3, Maroc 4, Tunisie 5, Turquie 6, etc. (variable nominale).

3. Dans une troisième catégorie, les informations ne sont pas directement codées sous forme de mesures mais peuvent faire l'objet ultérieurement d'un processus de mesure. Les cas sont potentiellement les mêmes que dans la deuxième catégorie, mais les variables et leurs modalités sont définies après le recueil de l'information, ou après le recueil d'une partie de celui-ci.

4. La dernière catégorie reprend enfin les informations qu'il n'est pas possible de coder sous forme de mesure, ou que l'on ne désire pas coder sous forme de mesure : données spécifiques, avis, sentiments, historiques, etc.

Le processus de mesure vient donc se mettre naturellement au service du recueil d'informations lorsque ces dernières appartiennent à une des trois premières catégories.

3.. Le processus de mesure par rapport aux processus de contrôle et d'évaluation

En entreprise, le "contrôle de qualité" fait largement appel à la mesure, le plus souvent automatisée par ailleurs.

Dans le monde de l'enseignement, le contrôle des connaissances recourt le plus souvent aux notes chiffrées - que les enseignants utilisent abusivement comme si celles-ci avaient les propriétés d'une échelle à intervalles égaux - ou aux échelles d'appréciation, c'est-à-dire à des mesures ordinales comme par exemple "Excellent", "Très bien", "Bien", "Faible", "Insuffisant".

Le contrôle ne requiert cependant pas nécessairement la mesure. Je peux très bien me contenter, par exemple, d'observer la présence de tel indice et, en fonction du résultat de mon observation, décider d'apporter une correction ou au contraire de continuer le processus. (...)

DE KETELE, J-M. & ROEGIERS, X.

CARDINET, J. "L'élargissement de l'évaluation", dans *Hommage à Cardinet*, Fribourg, Delval, 1990, p.109/138 , article publié dans *Education et recherche*, vol 1, n°1, 1979, p. 15/34 :

(...) *Le but de l'évaluation*

La pédagogie expérimentale est née de soucis très concrets. L'enseignement de l'orthographe atteint-il des résultats satisfaisants ? Une autre façon d'aborder cet enseignement ne réussirait-elle pas mieux ? L'apparition de méthodes concurrentes suscitait nécessairement des conflits : le recours à l'expérience devait permettre de trancher objectivement en faveur de l'une ou de l'autre. Une méthodologie simple en découlait. Il suffisait de comparer les résultats d'un groupe de contrôle. Les progrès ultérieurs dans l'organisation des dispositifs expérimentaux ont contribué à affiner ces comparaisons ; le principe d'une mise en compétition de méthodes différentes est ainsi resté fondamental dans la théorie classique de l'expérimentation pédagogique.

Lorsqu'on veut examiner cette conception de façon critique, la position de l'expérimentateur paraît forte à première vue, parce que parfaitement logique. Si une méthode a de meilleurs résultats qu'une autre, il faut la préférer. Qui pourrait le nier ? Il est intéressant, pourtant, de souligner combien cette façon d'aborder le problème restreint le domaine d'étude et rend fragiles les conclusions que l'on peut en tirer.

Il est sous entendu, tout d'abord, que les acquisitions cognitives des élèves sont le critère essentiel sur lequel devra porter la comparaison. Que fait-on alors des autres objectifs éducatifs ? On admet aussi implicitement que l'intérêt des autres parties engagées dans le processus éducationnel (parents, maîtres, autorités) s'efface entièrement devant l'impératif de l'apprentissage des élèves. Peut-on oublier les besoins des enseignants, les désirs des parents, les soucis des administrateurs, les pressions de l'opinion publique, et même, tout simplement, les goûts des enfants ? L'expérimentateur avait réduit les dimensions du problème pour pouvoir le traiter plus facilement. Il reste à se demander si la réponse qu'il donnait alors avait encore un sens.

Cronbach (1964) fut l'un des premiers à critiquer cette conception de la recherche pédagogique et à demander un élargissement de son domaine d'étude. La mise en compétition de méthodes, par exemple, est incapable de donner des résultats généralisables, en raison de l'importance que prennent les facteurs du contexte, par rapport aux facteurs mêmes que l'on veut étudier (engagement affectif des maîtres ou des élèves, par exemple). Il est, de plus, dispendieux de recommencer des évaluations comparatives pour chaque nouvelle méthode ou à l'occasion de chaque nouvelle décision à prendre. Cronbach demande que l'on examine plutôt les résultats de l'enseignement par rapport à un large éventail d'objectifs éducatifs. On devrait connaître la performance des élèves d'abord dans chacune des dimensions pédagogiques et psychologiques visées par le nouvel enseignement. Par exemple, on devrait pousser l'enquête aux différents niveaux d'objectifs cognitifs, pour s'assurer de la capacité de généralisation des connaissances acquises ; on devrait s'intéresser aux transformations d'attitudes résultant de l'enseignement. Par ailleurs, on devrait également contrôler des dimensions que le nouveau curriculum ne vise pas explicitement ; pour s'assurer que le gain dans une direction n'est pas obtenu aux dépens d'une autre capacité aussi importante. Si l'on possède ces données pour chaque forme d'enseignement, on est en mesure de faire ensuite toutes les comparaisons utiles.

Ce premier élargissement fut le point de départ d'une évolution que Stake (animateur principal de la Conférence de Liège) a cherché à pousser à son terme extrême. Pour lui, poser le problème en termes d'objectifs éducatifs plus ou moins atteints est trop restrictif. La théorie des curriculums, et la pratique classique de l'évaluation, mettent en évidence la multiplicité des liens qui relient l'école aux divers secteurs de la société. Evaluer un nouveau curriculum implique, par conséquent, qu'on consulte tous les partenaires, qu'on parte des décisions à prendre par les responsables ; qu'on détermine les informations dont ils ont besoin ainsi que les questions que se posent les autres intéressés. Le chercheur qui s'astreint à cette enquête s'aperçoit de la

multitude des dimensions à prendre en compte : économiques, sociales et politiques, juridiques et historiques, personnelles, éthiques, etc...

Dans cette perspective, la finalité de l'évaluation est d'abord pratique, au service de l'enseignement. Ainsi, l'évaluation élargie, appelée "holistique" par Wulf (1975), a pour but premier de fournir aux divers partenaires du système scolaire les informations dont ils ont besoin pour améliorer le fonctionnement de ce système. Ceci implique d'estimer, dans toute la mesure où c'est pratiquement réalisable, l'ensemble des facteurs qui paraissent significatifs aux intéressés, sans simplifier indûment la complexité des variables qui interviennent. (...)

Conclusions.

On peut, en terminant, prendre un peu de recul et chercher à évaluer l'évaluation élargie elle-même. Du point de vue pratique, il ne fait pas de doute que cette nouvelle perspective rendra beaucoup de services. Elle semble directement issue d'une réflexion sur les besoins concrets des responsables et d'un dialogue avec eux. Elle s'emploie efficacement à les satisfaire. Elle découle d'une politique générale de participation de tous les milieux concernés au pilotage des innovations pédagogiques. Elle est ainsi un instrument réfléchi et cohérent de gestion des systèmes scolaires en évolution.

Du point de vue théorique, cette conception nouvelle pose beaucoup de problèmes, mais c'est peut-être son grand mérite. La pédagogie expérimentale semblait en effet trop fixée, depuis longtemps, sur un modèle à la fois rigoureux et irréaliste. Cette contestation va rouvrir un débat de fond, trop longtemps esquivé.

Le premier problème, en effet, que pose cette nouvelle perspective est celui de l'objectivité. Ses détracteurs n'ont pas de peine à souligner le danger que représente l'appel à l'opinion subjective des intéressés pour fonder un jugement pédagogique. L'expérience d'Hawthorne, comme les travaux de Festinger sur la modification des attitudes, montrent que le simple fait d'être participant dans une expérience modifie la perception et l'attitude de celui qui est interrogé. Il faudrait donc, plutôt que de s'en remettre aux témoins, détromper, par la mise en évidence de faits objectifs, ceux qui s'illusionnent.

Ceux qui contestent le modèle scientifique traditionnel répondent que l'objectivité dont on le pare est, en fait, également illusoire. D'une part, comme on l'a vu, il ne permet pas vraiment de déterminer des lois générales : soit on ne peut pas contrôler les variables essentielles, soit on les contrôle si bien qu'on les oublie (en laboratoire) et les résultats n'ont plus alors de portée pratique. Plus profondément encore, on peut souligner le caractère "construit" et artificiel des théories scientifiques, surtout de toutes celles qui se fondent sur un modèle mathématique ou statistique élaboré. Il devient impossible de les vulgariser sans les trahir. On ne sait plus bien, par exemple, quel rapport les modèles d'apprentissage, l'analyse factorielle des aptitudes, ou les statistiques multivariées entretiennent encore avec les problèmes concrets de la classe. N'est-il pas temps de revenir au vécu, de retrouver ainsi l'intégralité du réel, au lieu d'élever un échafaudage ambitieux qui finit par nous le cacher ? Il semble que l'intuition directe puisse nous permettre un autre ordre de connaissance, infiniment plus riche que celui de la construction scientifique.

Mentionner l'intuition, c'est relancer un débat philosophique bientôt centenaire sur la possibilité ou la signification d'une science de l'homme. Par le simple fait qu'il peut prendre conscience des déterminismes qui l'enserrent, l'homme n'échappe-t-il pas toujours plus ou moins à la loi qui prétend le régir ? Il semblerait qu'on en fasse l'expérience aujourd'hui dans les sciences de l'éducation. L'évolution de la réflexion en pédagogie suit en effet un tracé analogue à celui que l'on observe actuellement parmi les autres sciences du comportement. Après un point de départ positiviste, la

psychologie aussi est en crise. On voit aujourd'hui des psychopédagogues aussi fameux que Cronbach douter de la possibilité d'objectiver véritablement leur domaine (1974). Le behaviorisme est battu en brèche aux U.S.A. le modèle piagétien s'y impose. Il n'est donc pas étonnant qu'en pédagogie également on recherche de nouvelles voies. Le cadre conceptuel de la "compréhension", aidé des cadres structuraliste et fonctionnaliste, ne peut-il rendre mieux compte de la réalité humaine que le cadre positiviste ?

Il semble difficile de répondre de façon assurée à cette question. En tout cas, d'autres sciences humaines commencent à apporter leur contribution à la pédagogie. Elles éclairent une série d'aspects de la réalité, partiels sans doute, mais complémentaires. La face interne des phénomènes retrouve place, enfin, dans un univers scientifique qui semblait l'avoir bannie pour toujours. Ceux pour qui les points de vue subjectif et objectif sont irrémédiablement étrangers espèrent ainsi prendre une revanche contre l'objectivité desséchante de la psychologie "en troisième personne". Pour d'autres, on peut aussi déceler dans cette évolution la recherche d'une synthèse, bien lointaine encore, où les lois objectives porteraient sur des variables dont la signification nous serait accessible par l'intuition. Ce ne sont là, aujourd'hui, que des conjectures. Il est préférable de ne pas engager l'avenir et de constater simplement un élargissement, désormais bien marqué, de nos conceptions scientifiques. La pédagogie expérimentale se dégage d'une norme trop rigide, qui faisait que toute information, non confirmée par un test statistique, n'était pas estimée digne de foi. D'autres sciences existent qui ont, elles aussi, leurs exigences d'universalité , et qui semblent pourtant plus souples, mieux adaptées aux problèmes concrets de l'évaluation pédagogique que les plans expérimentaux de Fisher. Sans renoncer aux acquis du passé, il faut donc nous ouvrir à cet enrichissement de nos moyens d'action, qui ouvre des perspectives passionnantes sur les plans de la recherche théorique comme de l'application pratique.(...)

Cronbach, L-J., Evaluation of course improvement, in Hearth, R-W (ed) *New curricula*, New York Harper and Row, 1964

Wulf, C. Evaluation in Rahem praxisnaher curriculumentwicklung, in Isenegger, U. und Santini, B. *Begriff und Funktionen des curriculums*, Weinheim und Basel, 1975, p. 131/159

Cronbach, L-J. Beyond the two disciplines of scientific psychology, exposé présenté à l'APA, New Orleans, 1974, Ocasionals papers of the stanford evaluation Consortium, stanford University

CARDINET, J.

LANDELLE, J-J. "L'évaluation, une pratique scientifique?", *L'évaluation en questions* , Delorme, C., CEPEC, 2° ed, 1988, p.57/66 :

(...) *L'évaluation comme mesure*

Il y a toujours eu une forte propension, chez les chercheurs en sciences de l'éducation, à utiliser les outils statistiques pour tenter d'atteindre le niveau de formalisation des sciences "exactes". Une des tendances actuelles en matière d'évaluation tourne ainsi autour d'une *théorie de la généralisabilité*, laquelle est fortement liée à un aspect de *l'analyse des données*.. Une étude fine et détaillée de l'outillage utilisé dépasse largement le cadre de l'article ; disons qu'elle vise à déterminer le degré auquel on peut étendre des résultats obtenus dans un contexte, à un contexte différent. J. Cardinet et Y. Tourneur proposent une démarche en quatre étapes:

- *plan d'observation* : on précise l'ensemble des données observées ;
- *plan d'estimation* : on précise le domaine des observations possibles ;

- *plan de mesure* : on spécifie la population des objets d'étude admissibles et l'univers des conditions admissibles de mesure ;

- *plan d'optimisation* : on spécifie la population de différenciation optimale et l'univers de généralisation optimal.

Le tout est fondé sur *l'analyse de la variance*. Celle-ci permet de rechercher l'influence d'un ou plusieurs facteurs contrôlés sur des éléments de base ; l'influence de ces facteurs peut avoir un caractère aléatoire et les éléments de base peuvent être dispersés ; l'analyse de la variance permet de discerner les variations dues aux facteurs et celles dues à la dispersion des éléments.

Des réserves sur l'évaluation-mesure

Réduire l'évaluation à la seule utilisation de l'analyse des données, ou du moins contribuer à faire penser que seule une évaluation quantitative et empiriste, fondée sur le concept de mesure, est sérieuse, appelle quelques réserves :

D'un point de vue scientifique

Aveuglé par l'élégance toute mathématique d'un outil, le risque est grand de ne bâtir que sur du sable. Les problèmes soulevés par le recueil des données sont trop souvent minimisés : les données ne sont pas celles d'une situation mathématique, mais plutôt celles d'une situation expérimentale ; elles ne tombent pas d'elles-même, mais ont été prises, choisies, sélectionnées ; si l'on n'y prend pas garde, le travail conduit à des résultats artefactuels, quelles que soient l'exactitude des calculs faits et la puissance de l'ordinateur pour les mener à bien. La pratique scientifique actuelle invite à une très grande circonspection et demande de garder présente à l'esprit la question : "Quel est précisément le phénomène que le corpus des observations restitue ?".

Le corpus est en général lourd et relativement inaccessible ; aussi va-t-on donner une image organisée, par l'intermédiaire de données - ce qui renvoie à des modèles, le plus souvent implicites et porteurs d'une bonne partie des informations auxquelles le traitement aboutira. L'analyse des données ne fournira rien de plus que ce qu'elles contenaient. On peut penser que les évaluations pédagogiques sont trop souvent subjectives et intuitives et dénotent une complaisance pour l'attitude littéraire plus que scientifique ; on peut mettre l'accent sur l'interdépendance des faits pédagogiques, ce qui conduit naturellement à l'analyse des données, mais alors il faut penser à choisir parmi les différentes méthodes multidimensionnelles (analyse en composantes principales, méthodes des nuées dynamiques, analyse ascendante hiérarchique...) en se donnant un critère de minimisation de la déformation et une mesure de cette déformation, et ne plus se contenter d'une seule technique.

D'un point de vue philosophique

On pense que seules les régularités, les normalités définies par la statistique présentent de l'importance. L'incident, le dysfonctionnement, la crise ne sont pas pris comme révélateurs pertinents. Il y a quelque vanité à vouloir réduire la subjectivité de l'observateur, à rechercher la désaffectivisation, à faire l'événement pour ne retenir qu'une donnée dans le travail sur le pédagogique. Considérer l'individu comme un *n-uplé* et l'ensemble des individus comme une image de points n'est pas neutre et peut vite relever d'une idéologie de la manipulation. C'est vite oublier l'épaisseur de toute transaction pédagogique. Une évaluation conduite uniquement sur une telle base, réduit, banalise, dépersonnalise et ne laisse rien subsister de ce qui fait le risque de l'action pédagogique. Le savoir constitué apparaît comme neutre, pur et utilisable, indépendamment des acteurs et des fins éducatives poursuivies. Est-ce bien sérieux ?

D'un point de vue épistémique

On ne sait de quel lieu est prononcé le discours sur cette pratique statistique de l'évaluation ; le système annule à l'avance toute opposition et monopolise la parole. Il est irréfutable et inattaquable... ce qui conduit à penser, selon le critère de la falsifiabilité, à une idéologie. "... si à un moment ou à un autre, un savoir local prend le pouvoir global, alors vous pouvez être sûrs, à ce signe et à ce travail, qu'il vient de s'affirmer comme tout autre chose qu'un savoir.." écrit M. Serres (p.153). L'utilisation de l'ordinateur, sous prétexte d'aide, ne fera qu'enfermer l'évaluation dans une caverne pseudoscientifique où le regard cloué sur les voyants lumineux, les états sortant de l'imprimante, il sera réduit à une obscurité tout intérieure. Après l'homo sapiens, l'homo computans ! (...)

Cardinet, J. & Tourneur, Y. *Assurer la mesure,* Berne, Peter Lang, 1985
Serres, M. *Le passage du nord-ouest,* Paris, Minuit, 1980

LANDELLE, J-J

ARDOINO, J. & BERGER, G. "L'évaluation comme interprétation", *Pour* n°107, 1986, p. 120/127

Une distinction épistémologique

La distinction contrôle/évaluation ne recouvre pas seulement une séparation entre des pratiques différentes. Elle délimite, en fait, deux univers différents mais complémentairement nécessaires. Il est devenu aujourd'hui quasiment banal d'opérer de multiples distinctions entre des pratiques très diversifiées, baptisées à tort "évaluation". L'épais dictionnaire de l'évaluation et de la recherche sur l'éducation de G De Landsheere est un modèle du genre. Il repère (on pourrait presque dire récupère) toutes les patiques possibles, et les classe en fonction de leur spécificité et de leur degré de scientificité, de rigueur ou même de pureté.

Au fond, toute taxonomie fait l'économie d'une phénoménologie infiniment plus complexe. D'autres (et c'est très à la mode actuellement dans les sciences sociales) préfèrent utiliser des classifications qui reposent sur des problématiques fonctionnalistes. On distinguera alors d'un côté l'évaluation régulatrice et/ou formative, qui a pour fonction de réorganiser un système à l'aide de méthodes de feed-back ou de rétro-action, de l'évaluation aux fonctions terminales, baptisée évaluation certificative (Cardinet) ou sommative ; dont la fonction est de catégoriser, certifier et valider des pratiques, des comportements ou des connaissances.

(...) Pour reprendre l'exemple de la notation, on peut inventer un système de notation continue de 0 à 20 ou à 1000, cela ne change rien au problème. Il n'y aura évaluation qu'au moment où quelqu'un déclare qu'une note est ou non accepable, qu'elle signifie tel type de qualité, etc. Soit au moment où émerge le qualitatif dans la quantitatif. En ce sens, il n'existe d'évaluation que qualitative, dans la mesure où elle représente l'introduction de discontinuités de valeur dans des sytèmes continus. Une évaluation économique consiste,par exemple, à trouver un produit trop cher ou inintéressant. Ce jugement de valeur, qui peut aller du "pas du tout" au "passionnément", comme dans le "jeu de la marguerite", introduit des seuils et des ruptures, simples ou complexes, dans la continuité des valeurs monétaires. Il ne s'agit pas de dénier l'importance de la qualification, au contraire, mais de toujours garder à l'esprit que la chaîne quantitative n'est jamais qu'un descriptif organisé de la réalité et que l'évaluation consiste justement à briser la continuité de cette chaîne.

(...)

ARDOINO, J. & BERGER, G.

L'autre phénomène idéologique, qui permet de comprendre une certaine logique dans la succession (non linéaire) des modèles,

est cette centration, qui s'amorce déjà dans cette première matrice, de l'évaluation sur les situations de formation. Le consensus réduit l'évaluation au domaine scolaire.

Dès lors, l'évaluation désigne, dans ce système d'idées, deux choses

- une dimension de l'Homme (émettre des « jugements de valeur» après avoir effectué des mesures "scientifiques"),
- une dimension de la formation, de l'acte éducatif, du scolaire (vérifier les acquisitions).

BONNIOL, J-J. , GENTHON, M. & ROGER, M "L'évaluation en psychologie : approches théoriques et conditions méthodologiques", *AESE* n°6, 1986, p.12/18

(...) On peut repérer une baisse importante du nombre des travaux sur la mesure. Il ne s'agit pas de vouloir effacer ou gommer la mesure, mais l'aspect "mesure" de l'évaluation qui occupait une place prépondérante ne peut plus, loin de là, prétendre rendre compte à elle seule de tout le champ des préoccupations sur l'évaluation.

Un élément important dans cette évolution est sans doute le fait que les Sciences de l'Education ont pris de plein fouet le problème de l'échec scolaire. Le regard de la mesure s'est donc concentré sur l'échec ; mais les constats auxquels aboutissait ce regard ne permettait pas de traiter le problème. (...)

BONNIOL, J-J. , GENTHON, M. & ROGER, M