Un point de vue didactique sur les questions d'évaluation en éducation.

Eric Roditi

Un point de vue didactique des questions d'évaluation en éducation

Éric Roditi
Université Paris Descartes, Laboratoire EDA
Faculté des sciences humaines et sociales - Sorbonne

La première section de ce chapitre propose quelques réflexions qui ne manquent pas d'apparaître dès que l'on commence à se pencher sur la question des évaluations des élèves en mathématiques. On peut s'interroger sur leur portée et leurs limites, comme sur leur utilisation ; que l'on pense aux évaluations proposées en classe dans le cadre de l'enseignement, aux examens ou aux évaluations nationales et internationales.

Deux parties organisent ce texte. La première rappelle les objectifs poursuivis par l'évaluation et signale quelques difficultés rencontrées pour les atteindre. La seconde revient sur les recherches en docimologie et en évaluation, notamment sur l'évolution de la façon dont les chercheurs considèrent les notes attribuées aux élèves d'une part, et l'activité même d'évaluation d'autre part.

I. L'ÉVALUATION, OBJECTIFS ET DIFFICULTÉS

L'évaluation effectuée par un acteur d'une situation constitue une prise d'information sur l'activité ou la production d'un ou des acteurs de cette situation, en référence à des normes ou à des objectifs. Celle que nous considérons dans cette section concerne les élèves d'une classe ou plus généralement du système scolaire, en mathématiques. Comme l'indique le *Dictionnaire des concepts fondamentaux des didactiques* (Reuter et al., 2007, p. 105), « *l'évaluation poursuit des buts de certification, de régulation (…), d'amélioration (…), ou enfin de prédiction de l'avenir (par exemple, en vue d'une orientation scolaire ou professionnelle)* ». Différentes questions se posent alors : sur quel acteur porte l'évaluation ? qu'est-ce qui est évalué de son activité ou de la production de son activité ? avec quels objectifs et quels résultats ?

De façon plus ou moins directe, les trois sous-systèmes du système didactique peuvent être évalués : les savoirs enseignés (évaluation du système éducatif et en particulier des programmes scolaires), les élèves d'une classe, ou leur professeur de mathématiques si ces élèves ont suivi l'enseignement du même professeur.

1. L'évaluation des savoirs enseignés

Les évaluations produites depuis 1987 par l'APMEP (Association des Professeurs de Mathématiques de l'Enseignement public), sous l'impulsion de Bodin (1994), avec le soutien de l'INRP (Institut national de Recherche pédagogique, devenu l'Institut français de l'Éducation depuis 2010) et en lien avec les IREM (Instituts de recherche sur l'enseignement des mathématiques) et l'ARDM (Association pour la recherche en didactique des mathématiques), portent sur les programmes d'enseignement des mathématiques. L'acronyme qui les désigne, EVAPM, témoigne de leur objectif puisqu'il signifie « évaluation des programmes de mathématiques ».

Ces évaluations permettent aux professeurs de mathématiques, mais aussi à tous ceux qui s'intéressent à leur enseignement, de mieux connaître l'état et l'évolution des apprentissages des élèves, et donc, d'une certaine manière, de la réalité de la transmission des notions qui figurent au programme d'enseignement.

Les analyses didactiques et statistiques effectuées sur les données récoltées ont aussi permis de mettre en évidence des relations entre les domaines enseignés, géométrique et numérique par exemple, des différences de réussite entre les redoublants et les non-redoublants, entre les élèves destinés à des orientations différentes, ou entre les filles et les garçons… D'une certaine manière, en évaluant les productions des élèves confrontés à différentes tâches mathématiques, c'est le *curriculum* qui est évalué, c'est-à-dire tout le système duquel le programme scolaire fait partie.

Avant d'aborder la question de l'évaluation des apprentissages des élèves, je souhaite revenir sur le fait que ces évaluations du *curriculum* sont indirectes puisque ce sont les évaluations des productions des élèves qui en sont à l'origine. Il faut donc, pour passer des dernières aux premières, mettre en œuvre des méthodes qui, dans certaines études, sont loin d'être indiscutables. Plus exactement, ce sont les interprétations des résultats obtenus grâce à ces méthodes qui doivent être discutées ; les conclusions des recherches en éducation n'étant pas indépendantes de supposés, ou de conséquences, idéologiques ou économiques. Voici un exemple pour illustrer, simplement et brièvement, l'intérêt que l'on peut trouver à ne pas se contenter des conclusions d'une enquête, et à se documenter tant sur les résultats que les méthodes utilisées pour les obtenir.

On entend souvent affirmer que le redoublement n'est pas efficace et que de nombreuses enquêtes fondées sur des données internationales ont permis d'aboutir à cette conclusion dont les systèmes éducatifs doivent tenir compte. Un rapport intitulé « *Les apports de la recherche sur l'impact du redoublement comme moyen de traiter les difficultés scolaires au cours de la scolarité obligatoire* » a été établi à la demande du Haut Conseil de l'évaluation de l'école ; il a été rédigé en décembre 2004 par Jean-Jacques Paul, directeur de l'IREDU (Institut de recherche sur l'éducation, Université de Bourgogne) avec la collaboration de Thierry Troncin, doctorant dans le même laboratoire de recherches. Dans une section consacrée à l'efficacité pédagogique du redoublement, on trouve, à la page 27 de ce rapport, un graphique (cf. figure 1) mettant en relation le taux de redoublement et la

performance en lecture des élèves de dix ans, cette performance étant reprise de l'enquête internationale PIRLS (*Progress in International Reading Literacy Study*) réalisée en 2011.

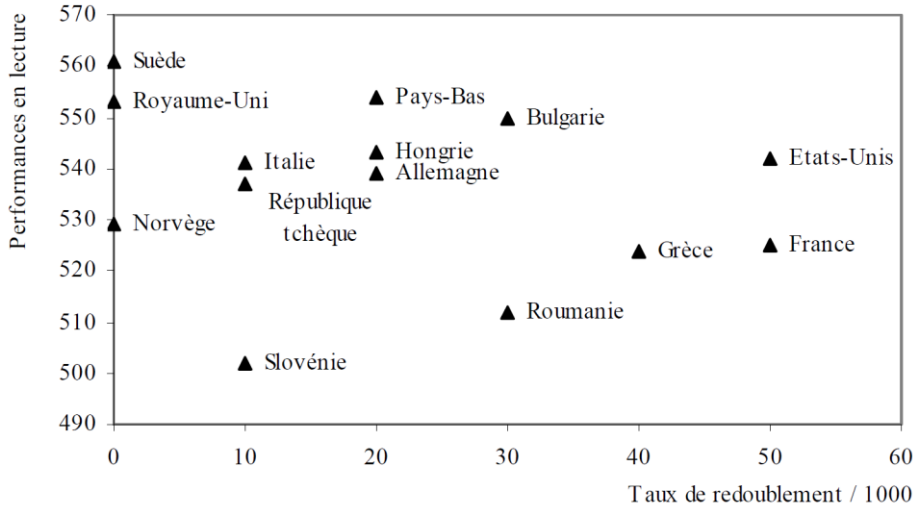

Figure 1. Performances en lecture à dix ans et taux de redoublement

Le rapport indique, à la page suivante, qu'à la lecture de ce graphique, « *il apparaît nettement que les pays les plus adeptes du redoublement n'ont pas de meilleurs résultats que ceux qui le pratiquent modérément ou qui le réfutent* ». Il ajoute que « *le redoublement n'apporte pas de valeur ajoutée quant au niveau moyen de la population d'élèves. La tendance serait plutôt en sens opposé.* » Deux analyses de cette conclusion peuvent être proposées. La première est linguistique : les auteurs écrivent que les pays les plus adeptes du redoublement n'ont pas de meilleurs résultats que ceux qui le pratiques modérément ou qui le réfutent ; force est de constater qu'ils auraient aussi bien pu (ou dû) écrire que les pays qui réfutent le redoublement ou qui le pratique modérément n'ont pas de meilleurs résultats que ceux qui en sont adeptes. La seconde est statistique : les auteurs écrivent que la tendance serait plutôt en sens opposé. Ici, la « *tendance* » doit se comprendre comme la valeur de coefficient directeur de la droite de régression du nuage de points, tendance dont le sens est indiqué par le signe du coefficient directeur. En utilisant les valeurs lues sur le graphique, on peut calculer le coefficient de la droite par la méthode des moindres carrés, il vaut – 0,2. Autrement dit, une augmentation de 10 pour mille du taux de redoublement, c'est-à-dire de 20% de l'étendue puisque ces taux varient de 0 à 50 sur le graphique), fait baisser le niveau moyen de performance en lecture de 2 unités sur une échelle où ces performance varient de 500 à 560, c'est-à-dire de 3% seulement environ ! On comprend mieux l'emploi du conditionnel dans la dernière phrase… Et cela d'autant plus que le coefficient de détermination associé à cette droite, coefficient qui permet de juger de la qualité de l'approximation du nuage par la droite, a une valeur extrêmement faible : $R^2 = 0,06$. Il ne s'agit pas ici de dévaluer l'ensemble de ce rapport, mais de souligner combien le passage de l'évaluation des performances des élèves à celle du système éducatif conduit à des difficultés méthodologiques souvent

extrêmement importantes. Mais cela constitue, d'autant plus, un bel enjeu pour la recherche en éducation !

Changeons d'échelle et revenons à un autre passage, fondamental pour le professeur qui travaille dans sa classe, celui qui conduit de l'évaluation des productions des élèves à celle de leurs apprentissages.

2. L'évaluation des apprentissages des élèves

Par l'évaluation des productions d'un élève, c'est en effet plutôt généralement l'élève, ou plus précisément ses apprentissages qu'on cherche à évaluer. Dans les travaux scientifiques, jusqu'aux années 1970, l'évaluation était pensée comme la mesure de la production de l'élève, c'est-à-dire la mesure de la qualité de cette production. Les recherches portaient alors sur des questions d'efficacité et d'équité dans la détermination de cette mesure, notamment lors des examens qui aboutissent à la délivrance de diplômes. La seconde partie de ce texte, développe les difficultés rencontrées pour parvenir à une telle mesure et les conséquences qui en ont découlé en éducation.

Depuis les années 1970, en lien avec le développement de recherches constructivistes en éducation, d'autres manières de concevoir l'évaluation se sont développées qui visent à mettre en rapport, non seulement les objectifs d'enseignement et les résultats atteints, mais aussi les moyens mis en œuvre pour les atteindre.

Différentes évaluations ont alors été distinguées suivant les objectifs de l'évaluateur, voire de l'évalué dans le cas de l'auto-évaluation. De nombreux chercheurs en éducation (Allal et al, 1979 ; De Ketele, 1986 ; Perrenoud, 1998) ont contribué à préciser la nature de ces évaluations, et à développer leur intérêt pour l'éducation et la formation (Barbier, 1985). Ainsi, l'évaluation *sommative* s'effectue en fin d'apprentissage afin d'estimer les connaissances acquises, une telle évaluation est dite aussi *certificative*, notamment lorsqu'elle conduit à la délivrance d'un titre ou d'un diplôme. L'évaluation *formative* poursuit plutôt un objectif de régulation : régulation de l'apprentissage pour l'élève, régulation de l'enseignement pour le professeur qui vise alors à soutenir le processus d'apprentissage. Hadji (1995) propose de distinguer l'évaluation formative, dont la visée est plutôt pédagogique, de l'évaluation *formatrice*, plus spécifique des approches didactiques qui accordent une importance majeure aux processus d'apprentissage de savoirs précis[1]. Si ces évaluations permettent au professeur de réguler son enseignement, l'évaluation *diagnostique*, en amont, lui permet de le programmer en fonction des acquis des élèves. Enfin, l'évaluation *pronostique* vise à anticiper les chances d'un élève de construire les apprentissages visés dans un enseignement ultérieur, et donc à orienter les élèves vers certains cursus parmi ceux qui leur sont proposés.

[1] La gestion, par les professeurs, des « incidents didactiques » (Roditi, 2005) qui surviennent en classe, constitue, par son ancrage didactique et par l'adaptation de l'enseignement à l'apprentissage des élèves qui en découle, un cas particulier d'évaluation formatrice.

De même qu'il avait été fait mention, à travers l'étude d'un exemple extrait d'un rapport de l'IREDU, de la difficulté du passage de l'évaluation des performances des élèves à celle des systèmes éducatifs, il convient à présent de mentionner les difficultés rencontrées pour évaluer les apprentissages des élèves à partir de leur performance, et cela malgré tous les apports des recherches sur l'évaluation qui viennent d'être mentionnés.

Il faut, en effet, bien garder à l'esprit que les productions des élèves lors des évaluations sont les produits de leurs activités, et que ces activités comprennent bien d'autres processus que celui de le la mise en œuvre de leurs connaissances pour réaliser les tâches qui leurs sont proposées. Prétendre cela conduit cependant à de nouvelles difficultés méthodologiques : comment prouver, comme le prétendent certains chercheurs en éducation, qu'un élève aurait été plus performant s'il avait eu une meilleure estime de lui-même ou un meilleur sentiment d'auto-efficacité (Bandura, 2003) ?

L'exemple qui suit contourne cette difficulté en raisonnant sur un grand nombre d'élèves et en jouant sur les raisons qui conduisent certains élèves à se sous-évaluer. Il vise à montrer que les filles sont moins performantes en géométrie qu'elles le seraient si les stéréotypes de genre associés aux mathématiques, et à la géométrie particulièrement, ne les conduisaient pas à minorer leur performance, non-consciemment. L'exemple vient d'une expérience de Steele (1997), chercheur à l'université de Standford, reprise avec des élèves de 6^{e} et 5^{e} par Huguet & Régnier (2009) d'Aix-Marseille université. Dans cette expérience, réalisée par les chercheurs français, les élèves sont séparés en deux groupes et doivent reproduire une figure complexe selon deux modalités : dans la première c'est une tâche de géométrie qui est proposée au premier groupe, dans la seconde c'est une tâche de dessin que les élèves du deuxième groupe doivent réaliser. Aucune différence significative n'est observée par les chercheurs entre les performances des garçons des deux groupes : la tâche de géométrie est réussie par un peu moins de 24% des garçons, et la tâche de dessin par environ 23% d'entre eux. Une différence significative est en revanche observée pour les filles. Elles sont proportionnellement moins nombreuses que les garçons à réussir la tâche de géométrie puisque seulement 21% d'entre elles reproduisent la figure de façon satisfaisante. Lorsque c'est une tâche de dessin qui est proposée, la réussite passe à plus de 25%, les filles deviennent meilleures que les garçons, alors que c'est bien exactement la même figure qui est à reproduire selon les deux modalités de présentation de la tâche !

Cette expérience intéresse bien sûr les chercheurs qui étudient les stéréotypes lies au sexe ; elle suggère aussi qu'une réserve doit toujours être présente lorsqu'on prétend évaluer les apprentissages des élèves alors qu'on ne fait qu'évaluer leurs performances. L'évaluation des apprentissages n'est pas une mince affaire, les discours diffusés par la presse spécialisée ou tenus par certains formateurs et inspecteurs tendent parfois à déprécier la qualité du travail des enseignants en matière d'évaluation. Les distinctions effectuées par les chercheurs entres les différents types d'évaluation ont produit un vocabulaire qui n'était pas familier des professeurs, cela ne doit néanmoins pas laisser penser qu'ils n'effectuaient pas, déjà, des évaluations diagnostiques, formatives, sommatives et pronostiques. Il faut se méfier, me

semble-t-il, de l'emprunt à Clémenceau qui déclarait que « *La Guerre est une affaire trop sérieuse pour être abandonnée aux généraux* » par des spécialistes en éducation qui déclarent que l'évaluation est une affaire trop sérieuse pour être laissée aux enseignants…

Si l'évaluation des apprentissages est utile aux professeurs pour programmer et adapter leur enseignement, elle est aussi utilisée pour évaluer les professeurs eux-mêmes, en s'appuyant sur une hypothèse selon laquelle les meilleurs d'entre eux sont ceux qui font, toutes choses égales par ailleurs, le mieux apprendre leurs élèves.

3. De l'évaluation des élèves à l'évaluation des professeurs

Dans le cadre de leur profession, les professeurs de l'enseignement secondaire sont évalués, administrativement par leur chef d'établissement, et pédagogiquement par des inspecteurs de leur académie spécialistes de leur discipline scolaire. Ces évaluations ne doivent pas être confondues avec les précédentes : même si certains prétendent qu'elles ont vocation à constituer des situations de formation (les évaluateurs prodiguant des conseils aux évalués devant leur permettre d'améliorer leur pratique professionnelle), elles n'en demeurent pas moins essentiellement un outil de gestion des ressources humaines et des carrières.

Les chercheurs en éducation, eux aussi, tentent d'évaluer l'efficacité des enseignants. Dans une note de synthèse intitulée « *Les recherches sur les effets école et les effets maître »,* Bressoux (1994) rappelle qu'au début du 20^{e} siècle, et particulièrement des années 1930 aux années 1950, les chercheurs ont essayé de mettre en lien des caractéristiques personnelles des maîtres et un enseignement de qualité, ces derniers devant être intelligents, sympathiques, vertueux, allègres, etc. L'inconsistance des résultats obtenus ont contraint les chercheurs à emprunter de nouvelles directions. L'une d'entre elle consiste à étudier les corrélations entre les comportements des enseignants et les apprentissages des élèves. Elle n'a pas donné davantage de résultats, sans doute parce que l'enseignement ne se décrit pas seulement par les comportements du maître, parce que les apprentissages des élèves ne résultent pas seulement de l'enseignement dispensé en classe, et, enfin, parce que ce que fait le maître en classe ne dépend pas que de lui.

Une autre voie a été empruntée, non plus pour évaluer les professeurs, mais plutôt l'effet de leur enseignement, grâce à un détour méthodologique : en évaluant les progrès de nombreux élèves de différentes classes, et après avoir neutralisé certaines variables comme les catégories socio-professionnelles des parents, on a pu repérer des classes où les élèves progressent particulièrement. Ces études ayant été effectué dans l'enseignement primaire où chaque classe possède un enseignant, on a pu faire l'hypothèse que bons résultats de la classe pouvaient être attribués à leur professeur. Il ne restait plus alors qu'à aller les observer pour comprendre ce qui fait un bon enseignant. Ces travaux n'ont pas donné davantage de résultats : des qualités ont pu être remarquées chez ces enseignants dont les classes étaient particulièrement performantes (les encouragements et les critiques, la structuration des situations proposées, les questions posées aux élèves, le temps laissé pour chercher à y répondre, le traitement de leurs erreurs, etc.) mais

on a retrouvé des qualités analogues chez d'autre enseignants dont les classes n'étaient pas aussi performantes ! Les chercheurs en ont déduit que, sans doute, la combinaison des qualités comptaient davantage que chacune des qualités prises isolément, et que ce qui constitue une qualité dans un contexte n'en est peut-être pas une dans un autre contexte…

Il faut donc le savoir : on ne sait toujours pas (au sens d'un savoir scientifique rigoureusement établi) ce qu'est un bon enseignant !

La recherche en éducation marque donc de façon importante le lexique utilisé pour parler de l'enseignement, y compris par les enseignants. Or la question de l'évaluation, on l'a vu dans l'ensemble des propos tenus précédemment, apparaît suffisamment difficile pour qu'il soit utile de les compléter par un développement concernant la recherche et ses évolutions, en docimologie et en évaluation.

II. LES RECHERCHES EN DOCIMOLOGIE ET EN ÉVALUATION

Si, comme indiqué au début de ce texte, l'évaluation constitue une prise d'information sur l'activité ou la production d'un acteur d'une situation, en référence à des normes ou à des objectifs, alors elle ne se réduit pas à l'attribution d'une valeur à cette activité ou à cette production, une valeur qu'on désigne par le terme « note » dans l'enseignement. Le terme de docimologie a été introduit par Henri Piéron pour désigner, dans le cadre des examens, l'étude modes de notation, et en particulier la variabilité inter-évaluateurs et intra-évaluateur. La recherche en évaluation ne se limite pas à cette étude des modes de notation, elle porte notamment sur l'activité de l'évaluateur que les chercheurs essaient de décrire, de comprendre, voire de modifier.

Après une présentation de quelques critères qui permettent de juger de la qualité d'une notation, cette section montrera l'évolution des recherches en docimologie et en évaluation.

1. Évaluer la qualité d'une notation ou d'une évaluation

L'étude de la qualité d'une notation s'oppose à une vision de l'évaluation selon laquelle, comme l'écrit De Ketele (1993), l'acte évaluatif serait un acte intuitif ne nécessitant pas de définir ses objectifs, ses critères, ni ses modalités de production, d'interprétation et d'utilisation des résultats. Avec cette vision, l'évaluation serait « *un acte syncrétique (...) dans la mesure où l'évaluateur est celui qui a mené l'apprentissage, (...) la personne la mieux placée pour bien connaître les performances de ses élèves ou de ses étudiants* ».

Deux critères principaux sont pris en considération pour juger de la qualité d'une notation : sa validité et sa fiabilité (ou fidélité). De nombreux écrits traitent de ces critères, mais on peut retenir, simplement, qu'une mesure est valide si elle mesure bien ce qu'on veut mesurer, et qu'elle est fiable si elle est indépendante d'autres variables que celles qu'on veut mesurer : du temps (constance), de l'épreuve (consistance), de l'évaluateur (concordance). Ainsi, par exemple, lorsqu'un professeur de mathématiques évalue

l'apprentissage du théorème de Pythagore, les notes obtenues par les élèves qui confondent encore le double d'un nombre et son carré, risquent de ne pas refléter seulement leur connaissance du théorème, ce qui pose un problème de validité. Autre exemple, classique : malgré toutes les conventions adoptées pour l'utilisation d'un même barème, deux professeurs différents n'attribuent pas toujours, loin s'en faut, la même note à une même copie, ce qui soulève un problème de fiabilité.

Les spécialistes utilisent des outils, comme le coefficient alpha de Cronbach, pour estimer la fiabilité, notamment dans le cas des recherches ou des évaluations à grande échelle, nationales ou internationales. Le propos n'est pas ici de définir ces outils, mais d'indiquer ce dont ils rendent compte. La situation suivante illustre bien ce à quoi correspond la fidélité d'une évaluation. Supposons que dix élèves soient évalués par la même épreuve et trois évaluateurs qui leur accordent chacun une note sur 10. On pourrait supposer tout aussi bien qu'ils sont évalués par trois épreuves et le même évaluateur ou par la même épreuve et le même évaluateur à trois dates différentes. Le graphique suivant (cf. figure 2) propose deux situations différentes quant aux résultats obtenus lors de cette évaluation : à gauche l'évaluation est plutôt fiable ; à droite, elle ne l'est pas.

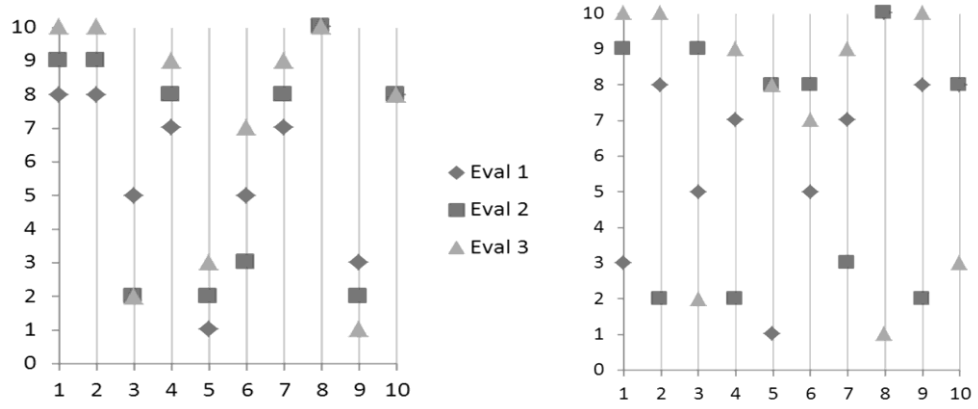

Figure 2. Fiabilité d'une évaluation : étude de la question de la concordance

Dans le graphique de gauche, on constate en effet que les trois évaluateurs sont exactement d'accord au sujet des élèves n°8 et n°10, et qu'ils sont plutôt d'accord quant aux autres élèves, excepté pour l'élèves n°6 où l'on constate une certaine disparité des notes. Dans le graphique de droite, en revanche, une forte disparité des notes accordées à chaque élève par les trois évaluateurs, excepté pour l'élève n°6. Cette disparité est même telle qu'à chaque fois qu'un évaluateur accorde une bonne note à un élève, on trouve un évaluateur qui lui en accorde une mauvaise.

Quel bilan peut-on tirer quant aux apports de ces recherches en docimologie ? c'est à cette question que répond la section suivante, en focalisant sur les aspects essentiels.

2. Docimologie : heurs et malheurs du modèle de la « vraie note »

L'étude de la qualité d'une notation a commencé avec les travaux de Piéron (1963) qui a montré les principaux biais présents dans les examens. L'évolution des travaux menés en docimologie est retracée dans la note de synthèse rédigée par De Ketele (1993) pour la *Revue française de pédagogie* qui vient d'être citée, ainsi que par Chevallard & Feldmann (1986) dans une brochure de l'IREM d'Aix-Marseille intitulée « *Pour une analyse didactique de l'évaluation* ».

Jusque dans les années 1980, ces chercheurs en docimologie, généralement psychologues, ont abordé le problème de la fidélité de la notation en s'inspirant du modèle de la mesure qui suppose l'existence d'une vraie mesure dont on ne peut obtenir que des approximations, souvent différentes, à chaque fois qu'on la détermine pratiquement. Ainsi, les docimologues postulaient l'existence de la « *vraie note* » dont la valeur serait celle vers laquelle converge la moyenne des notes lorsqu'on augmente indéfiniment le nombre de mesures. L'argument étant que la mesure subit des perturbations infimes, nombreuses et indépendantes, ce qui conduit à un modèle gaussien. Dans le cadre de travaux menés pour améliorer la correction des épreuves du baccalauréat, des recherches ont par exemple été réalisées par Laugier & Weinberg (1936) afin de déterminer le nombre de corrections nécessaires pour déterminer avec la précision souhaitée la note qu'il fallait attribuer à une copie. Il a fallu 13 corrections pour une copie de mathématiques, 78 pour une composition française, et 127 pour une dissertation de philosophie ! De tels résultats n'ont pas manqué de discréditer le modèle pour la pratique…

Du point de vue théorique, ce modèle a aussi été contesté : différentes travaux ont établi que les variations intra-correcteur et inter-correcteurs ne sont pas infimes, aléatoires et indépendantes. Un même professeur, de façon systématique, aura tendance à utiliser ce qu'il sait ou ce qu'il perçoit de l'évalué (effet de source), à évaluer plus sévèrement les premières copies (effet d'ordre), à sous-évaluer une copie qu'il corrige après plusieurs copies excellentes (effet d'ancrage), et il en sera de même pour une copie où les fautes apparaissent au début plutôt qu'à la fin. Tous les professeurs de mathématiques ont remarqué que deux collègues qui n'ont pas la même exigence quant à la rédaction des démonstrations géométriques n'attribuent pas les mêmes notes aux mêmes copies, et cela n'a rien d'aléatoire. Les chercheurs qui contestent le modèle de la mesure conçoivent plutôt l'évaluation comme « *une activité de comparaison entre une production scolaire à évaluer et un modèle de référence, comparaison qui est influencée par des <u>déterminants systématiques</u>* [2] *(…)* » (De Ketele, 1993, p. 61). Mais De Ketele conclut que, même si les auteurs les plus récents n'assimilent pas mesure et

[2] Souligné afin d'insister sur la différence entre une telle conception, et celle qui suppose que les variations sont infimes, nombreuses et indépendantes.

évaluation, il n'en reste pas moins que les travaux portent presque tous sur la « note » ou la « mesure ».

Ainsi, par exemple, dans la seconde partie du 20e siècle, un modèle statistique appelé « théorie de réponses aux items » est apparu, et qui est toujours en développement, pour faire face aux besoins de calibrage des questions d'évaluation utilisées dans les enquêtes à grande échelle. Ce modèle permet de déterminer la difficulté d'un item, ou son caractère discriminant, de manière indépendante des échantillons sur lesquels cet item est testé, et de manière indépendante des contenus scolaires sur lesquels il porte. C'est à la fois sa force… et sa faiblesse. Une faiblesse que pourrait compenser une approche didactique. Les savoirs que produisent les didacticiens sur l'apprentissage de contenus mathématiques précis pourraient permettre, en effet, de comprendre ce qui, du point de vue des savoirs en jeu et de leur enseignement, explique la difficulté ou le caractère discriminant d'une ques-tion d'évaluation.

3. Des recherches en évaluation en lien avec le contexte scolaire

Des chercheurs en évaluation ont emprunté d'autres voies que celle de la docimologie. Certains ont exploré la fonction de l'évaluation dans une école où, selon les sociologues français des années 1970, les inégalités culturelles et sociales sont légitimées. D'autres ont tenté de développer une conception de l'évaluation qui confronte les performances des élèves et les objectifs d'apprentissage, une conception qui a été beaucoup critiquée pour le découpage de l'enseignement en micro-objectifs auquel elle conduit. Dans les années 1980, se sont développées des conceptions de l'évaluation qui lui confèrent d'autres fonctions que celle de mesurer un état (celui des performances des élèves) en la plaçant au service de l'amélioration de l'enseignement, pour aider le professeur à l'adapter aux élèves (différencier) ou, plus généralement, pour l'aider à prendre les décisions adéquates. Aux deux critères de qualité d'une évaluation que sont la validité et la fiabilité, De Ketele (1993) propose alors d'ajouter un troisième critère, la pertinence, qui garantit que l'évaluation répond bien aux fonctions qui lui sont assignées. Dans cette lignée de travaux, certains auteurs conçoivent l'évaluation comme un processus de régulation, ce qui les conduit à mener des recherches en vue de le décrire et de le comprendre. D'une certaine manière, il y a, dans ces travaux et dans ceux des docimologues, comme une opposition entre deux modèles, entre deux façons de considérer les élèves : soit comme étant en concurrence, soit comme étant en développement.

Une telle approche possède un intérêt incontestable pour qui s'attache à mieux saisir ce qui se passe en classe. Chervallard & Fieldmann (1986) ont ainsi montré, dans la troisième partie de leur ouvrage, que les faits d'évaluations sont totalement imbriqués dans le fonctionnement didactique, et qu'ils permettent même d'en repérer les règles. Les notes attribuées aux élèves par les professeurs ne constituent pas des mesures de leur apprentissage, mais des leviers qui permettent une négociation entre le professeur et le groupe classe quant aux exigences que le professeur peut attendre et aux efforts que les élèves peuvent consentir pour atteindre l'objectif partagé d'apprentissage du savoir visé.

Ainsi, les « contrôles » répondent à d'autres finalités que de positionner les élèves sur une échelle de réussite. En rendant les notes, les professeurs renvoient un message plutôt qu'une valeur : la moyenne des notes constitue un signal pour chaque élève quant à l'adéquation entre les apprentissages réalisés et les attentes de l'enseignant ; la dispersion des notes de la classe traduit la cohésion du groupe, elle informe l'enseignant sur l'hétérogénéité des apprentissages ; la moyenne et la dispersion trimestrielles rendent compte du travail mené en classe pendant le trimestre. Pour en assurer la pertinence (au sens de De Ketele), le professeur ajuste l'outil d'évaluation : lorsqu'il compose le sujet (en choisissant les questions, en les décomposant éventuellement) ; pendant la passation (en aidant les élèves par des indications complémentaires ou des rappels de ce qui a été fait en classe), par la construction du barème (en valorisant plus ou moins les tâches techniques), et lorsqu'il attribue des points lors de la correction.

Il reste néanmoins que des notes sont attribuées aux élèves. Aussi, plutôt que d'étudier les défauts de fiabilité des évaluations et de tenter de les limiter, certains chercheurs ont choisi de se pencher sur l'activité d'évaluation, voire même sur l'activité de l'évaluateur. Cette activité est motivée (elle a des fonctions dans l'enseignement) et possède plusieurs destinataires, autres que l'évaluateur lui-même : l'élève, la classe, les parents, l'équipe pédagogique, etc. Les travaux de Noizet & Caverni (1978) avaient conduit à décrire le comportement de l'évaluateur en postulant l'existence d'un modèle de référence. D'après ces auteurs, l'évaluateur recherche des indices pour associer la copie à un modèle de référence, construit par l'évaluateur, sachant que ce modèle évolue pendant l'évaluation du fait de la connaissance croissante des productions des élèves évalués, et donc de la diversité des réponses qu'ils ont produites. On pourrait aussi supposer, dans la perspective d'une évaluation formative, que l'évaluateur compare les réponses des élèves à différentes réponses attendues[3] afin de repérer les besoins d'enseignement de chacun d'entre eux.

Des recherches récentes (Vantourout & Gosadoué, 2011) remettent en cause la place centrale accordée à ce processus de comparaison ainsi que que les termes même de cette comparaison. L'activité d'évaluation est alors analysée « *comme un processus de compréhension, au sens conféré à ce concept dans les travaux sur le compréhension de texte, c'est-à-dire comme un processus interactif et dynamique impliquant un lecteur (qui élabore une représentation du texte) et un texte (où figurent des indices dont la saisie dépend de la représentation, et qui, en retour, contribuent à son évolution)* ».

Alors que les travaux de didacticiens sont encore rares sur les questions d'évaluation, une telle proposition ouvre de riches perspectives pour les

[3] Une réponse attendue serait alors une réponse juste ou fausse à laquelle le professeur peut s'attendre.

didactiques. La compréhension dont il est question concerne, en effet, non seulement les savoirs disciplinaires dont l'apprentissage est évalué à travers les productions des élèves, mais aussi les savoirs relatifs à l'apprentissage, ces savoirs construits par les didacticiens dans le cadre de leurs recherches, et qui permettent d'identifier les conceptions des élèves à travers leurs réponses erronées.

BIBLIOGRAPHIE

Allal, L., Cardinet, J. & Perrenoud, P. (dir.). (1979). *L'évaluation formative dans un enseignement différencié*. Berne : Lang.

Bandura, A. (2003). *Auto-efficacité. Le sentiment d'efficacité personnelle*. Paris : De Boeck.

Barbier, J.-M. (1985). *L'évaluation en formation*. Paris : PUF.

Bodin, A. (1994). Un observatoire du système d'enseignement des mathématiques : EVAPM. *Vingt ans de didactique des mathématiques en France* (395-402). Grenoble : La pensée Sauvage.

Bressoux, P. (1994). Les recherches sur les effets-écoles et les effets-maîtres. *Revue française de pédagogie, 108*, 91-137.

Chevallard, Y. & Feldemann, S. (1986). *Pour une analyse didactique de l'évaluation*. Marseille : IREM d'Aix-Marseille.

De Ketele, J.-M. (dir.). (1986). *L'évaluation : approche descriptive ou prescriptive ?* Bruxelles : De Boeck.

De Ketele, J.-M. (1993). L'évaluation conjuguée en paradigmes. *Revue française de pédagogie, 103*, 59-80.

EVAPM. http://ctug48.univ-fcomte.fr/evapm/

Hadji, C. (1995). *L'évaluation : règles du jeu*. Paris : ESF.

Huguet, P., & Régner, I. (2009). Counter-stereotypic beliefs in math do not protect school girls from stereotype threat. *Journal of Experimental Social Psychology, 45*(4), 1024-1027.

Laugier, H. & Weinberg, D. (1936). Élaboration statistique des données numériques de l'enquête sur la correction des épreuves du baccalauréat, in *La correction des épreuves écrites dans les examens*. Paris : Maison du Livre.

Paul, J.-J. & Troncin, T. (2004). *Les apports de la recherche sur l'impact du redoublement comme moyen de traiter les difficultés scolaires au cours de la scolarité obligatoire*. Ministère de l'Éducation nationale : DEP.

Perrenoud, Ph. (1998). L'évaluation des élèves. De la fabrication de l'excellence à la régulation des apprentissages. Bruxelles : De Boeck.

Piéron, H. (1963). *Examens et docimologie*. Paris : PUF.

Reuter, Y. et al. (dir). (2007). *Dictionnaire des concepts fondamentaux des didactiques*. Bruxelles : De Boeck.

Roditi, E. (2005). *Les pratiques enseignantes en mathématiques. Entre contraintes et liberté pédagogique*. Paris : L'Harmattan.

Steele, C. M. (1997). A threat in the air : How stereotypes shape intellectual identity and performance. *American Psychologist, 52*(6), 613-629.

Vantourout, M. & Goasdoué, R. (2011). Correction de dissertations en SES. *Idées, 63*, p. 71-78.