Les recherches sur la notation des élèves

Quelles conséquences en termes d'action sociale ?

De l'entre-deux-guerres à nos jours, l'évaluation des élèves a fait l'objet de très nombreuses recherches en raison de l'enjeu social que constitue la question de l'équité des classements scolaires. L'objet de l'article est de présenter trois grandes catégories d'investigations et de montrer qu'elles reposent sur des conceptions théoriques spécifiques de l'évaluation qui présentent chacune un intérêt indéniable en termes d'action sociale pour les chefs d'établissement et les professeurs.

Pierre MERLE
IUFM de Bretagne

L'expression « évaluation scolaire » renvoie à des situations multiples – notation des élèves dans le quotidien de la classe, évaluation et orientation en conseil de classe, notation aux examens – et à des questions les plus variées : niveau scolaire des élèves, exactitude de la mesure des compétences, expérience subjective des élèves à l'égard des notes… La polysémie de l'expression indique assez la complexité de l'objet d'étude.

Parmi l'ensemble considérable des recherches réalisées, nous retiendrons trois types d'investigations qui participent de façon centrale à la réflexion sur le thème de l'évaluation scolaire[1].

L'ALÉA DE LA NOTATION : L'APPROCHE DOCIMOLOGIQUE[2]

La justification de la première recherche docimologique importante, réalisée en l'occurrence sur les notes au baccalauréat, est apportée, en 1936, par H. Laugier et D. Weinberg [1] : *« Le baccalauréat occupe en France une place prépondérante comme régulateur de l'accès aux professions libérales »*. L'interrogation des auteurs est donc explicitement celle de la pertinence de la sélection des

NOTES

1. Le terme d'évaluation scolaire est ici défini dans le sens restrictif, mais usuel, de notation des élèves. Les questions relatives à l'évaluation formative ne sont donc pas abordées directement.

2. La docimologie – du grec *dokimé*, « épreuve », et *logos*, « science » – constitue historiquement la première approche de la question de la notation des élèves. Elle consiste en l'étude statistique des variations des notes attribuées à une même copie.

TABLEAU 1 – Écarts maximums, écarts moyens et écarts les plus fréquents lors d'une expérience de multicorrection menée sur cent copies du baccalauréat (session 1930)

Discipline	Écart maximum	Moyenne des écarts	Écarts les plus fréquents
Version latine	12 points	2,97	5
Composition française	13 points	3,29	6 et 7
Anglais	9 points	2,24	4
Mathématiques	9 points	2,05	4
Philosophie	12 points	3,36	5 et 7
Physique	8 points	1,88	4

Source : H. Laugier, D. Weinberg [1]

LECTURE – les écarts les plus fréquemment rencontrés lors de la multicorrection de copies de français sont 6 et 7 points.

élites qui renvoie à une préoccupation sociale majeure, celle de l'équité des procédures de notation et de sélection. Cette préoccupation, centrale dans l'institution scolaire, est évidemment toujours présente aujourd'hui, et même davantage que dans l'entre-deux-guerres, en raison de l'accroissement considérable de la scolarisation et du rôle grandissant du diplôme dans l'insertion professionnelle, comme en témoigne notamment la part décroissante des autodidactes parmi les cadres.

Un des résultats majeurs des recherches de Laugier et Weinberg a été de montrer l'incertitude de la notation au bac. À partir d'échantillons représentatifs de lots de cent copies rédigées dans six disciplines ayant fait l'objet d'un écrit au baccalauréat, l'enquête a consisté à soumettre celles-ci à cinq correcteurs. L'étude a eu pour objet de s'interroger sur les écarts de notes entre les six correcteurs (cinq corrections par copie auxquelles il faut ajouter la correction réalisée par le véritable correcteur du baccalauréat) et a donné lieu, par discipline, au calcul de trois écarts spécifiques : l'écart maximum entre les six notes par copie ; la moyenne des écarts ; les écarts revenant le plus fréquemment (tableau 1).

Ces données, particulièrement classiques, montrent l'incertitude attachée à la correction des copies quelle que soit la discipline considérée. Laugier et Weinberg en avaient conclu que pour obtenir la « note vraie », il fallait recourir à la moyenne de 13 correcteurs en mathématiques, 78 en composition française, 127 en philosophie… La fiabilité incertaine de la correction des copies dans les disciplines scientifiques et littéraires pose sans aucun doute la question de la pertinence du jugement professoral. Dans la recherche menée par Laugier et Weinberg, l'évaluation professorale relative à l'admission ou à l'ajournement des candidats est, en effet, loin d'être consensuelle : une grande part des candidats serait soit refusée soit admise selon l'examinateur qui assure la correction de leurs copies. Ainsi en français, seulement 21 % des candidats sont refusés unanimement (note inférieure à 10 pour les six correcteurs), 9 % seulement des candidats sont reçus par l'ensemble des correcteurs, et 70 % sont refusés ou admis selon le correcteur. En mathématiques, discipline où l'accord des correcteurs est le plus fréquent, 36 % des candidats sont pourtant refusés ou admissibles selon l'examinateur…

Le bac : une loterie ?

Ces données ne correspondent cependant pas aux situations réelles d'examen : la décision d'admission ou d'ajournement n'est pas décidée à partir des notes obtenues dans une seule discipline mais à partir de la combinaison de l'ensemble des notes obtenues par le candidat. Le principe de l'examen est en effet de reposer sur plusieurs matières et tient notamment à l'idée qu'il existe, entre les différentes épreuves, un phénomène de compensation de l'aléa de la correction. Cette idée est juste : la sévérité d'un examinateur dans telle discipline est, en moyenne, compensée par l'indulgence d'un autre correcteur. Cependant, ces mouvements de compensation ne suppriment pas totalement l'incertitude relative aux jugements professoraux. En se limitant aux totaux des notes qu'obtiendraient les candidats dans six jurys constitués de façon différente, il s'avère que 31 % des candidats reçus par certains jurys seraient refusés par d'autres. C'est la même proportion de candidats – environ un tiers – qui passent actuellement les épreuves orales de rattrapage du baccalauréat. Autrement dit, les élèves jugés ni « bons » ni « faibles » après les épreuves du premier groupe (leur moyenne générale est comprise entre 8 et 10) font, par le biais d'un oral avec livret scolaire, l'objet d'une procédure spéciale d'évaluation de leurs compétences qui tient compte de l'incertitude de leurs performances au baccalauréat

en prenant en considération leurs résultats obtenus au cours de l'année de terminale [2]. Le principe mis en œuvre est relativement simple : dans une situation d'incertitude, mieux vaut deux évaluations – celle du bac et celle de l'année – qu'une seule.

La prouesse de l'organisation du bac tient au fait que, malgré l'incertitude de la notation de chaque copie, l'obtention du bac ne s'apparente pas, contrairement à des interprétations un peu rapides, à une loterie (*Cf.* aussi les données réunies sur ce sujet par É. Chatel). La fiabilité du bac comme indicateur de compétences scolaires est attestée d'ailleurs par le fait que les conditions d'obtention (mention, oral de rattrapage) sont de bons prédicteurs de la réussite scolaire ultérieure [3]. Il faut indiquer enfin que la pertinence du bac comme mesure du niveau scolaire n'exige pas une multiplication excessive des épreuves : certaines disciplines constituent des mesures plus fiables que d'autres pour connaître la compétence des élèves, si bien que le nombre d'écrits peut être limité sans porter préjudice à la validité de l'examen [4].

La note : une croyance scolaire ?

Cependant, la recherche menée par Laugier et Weinberg ainsi que les autres études de multicorrection réalisées depuis, aboutissent, quelle que soit la discipline, à montrer que la notation d'une copie est marquée par une incertitude sensible. Outre les données précitées, on peut présenter un autre exemple. Une expérience de multicorrection portant sur une copie de mathématiques (54 correcteurs) et une copie de français (58 correcteurs) en classe de troisième a donné les résultats suivants : l'écart maximum des notes est de 13 points en mathématiques (de 4/20 à 17/20) et de 11 points pour la copie de français (de 4/20 à 15/20). Les écarts-types et les coefficients de variation (écart-type sur moyenne) des deux distributions des notes sont respectivement pour la copie de mathématiques de 2,45 et 0,20, et celle de français de 2,46 et 0,26 [5]. La distribution des notes d'une copie corrigée par plusieurs dizaines de correcteurs suit approximativement une loi normale, c'est-à-dire une courbe « en cloche ». L'existence d'une telle distribution des notes indique, d'un point de vue statistique, que les sources de variation dans la correction d'une copie sont nombreuses et indépendantes entre elles. Comme l'indiquaient Laugier et Weinberg dès 1936 : *« Dans la dispersion des notes, la diversité propre des correcteurs intervient pour une part plus importante que la diversité des copies »* (*op.cit.*, [1]).

Or toute l'organisation administrative et pédagogique des collèges et lycées (relevés de notes, bulletins trimestriels, dossier de l'élève, classement quand celui-ci est maintenu) repose sur l'idée qu'il est possible de mesurer, avec précision, la compétence scolaire des élèves. Il serait possible d'établir, non seulement, un classement ordinal des performances, mais également un classement cardinal (la copie notée 15/20 étant supposée formellement trois fois supérieure à la copie notée 5/20), et ce postulat autorise le calcul de la moyenne des deux notes. Cette conception de l'évaluation repose sur le principe suivant : chaque copie fait l'objet, quel que soit le correcteur, d'une lecture identique et d'une pondération équivalente des mêmes qualités et insuffisances.

En fait, l'approche docimologique a montré qu'une telle théorie de l'évaluation n'est que partiellement fondée : elle n'est pertinente que pour distinguer les « bons » élèves des élèves « faibles », mais est déficiente pour classer les élèves, les plus nombreux, dont les performances sont moyennes. Cette conception d'une évaluation fondée sur la « vraie » note, mesure exacte de la compétence scolaire, existe au niveau des principes, au nom des valeurs de justice défendues par l'institution scolaire : conformément à l'idéal démocratique de l'égalité des droits, la mesure des compétences individuelles est réalisée dans des conditions égales pour chacun. Toutefois, cette égalité formelle à laquelle chacun adhère communément ne permet pas, comme le montrent les études docimologiques, que la notation d'un devoir ne diffère pas selon l'examinateur et les modalités de correction. Autrement dit, la note « vraie », sur laquelle les notateurs s'accorderaient, ne peut être qu'une construction statistique établie à partir de la moyenne des notes données à une même copie par des correcteurs différents. Cette notion de « vraie note » doit être distinguée de la notion de note « juste » pour le correcteur, définissable par l'application consciencieuse par celui-ci de ses critères d'appréciation ou de l'interprétation qu'il fait des critères de notation que chacun est réglementairement tenu de respecter (au bac par exemple).

Conséquences en termes d'action sociale

L'approche docimologique a été, et demeure, d'un grand intérêt pour trois raisons essentielles :
– montrer que la notation d'une copie est un travail d'expert et qu'à ce titre elle impose toujours un travail d'interprétation qui ne se répète jamais à l'identique et ne peut se réduire à une mesure de type physique (en supposant que celle-ci serait toujours simple à réaliser !). À ce titre, la notation d'une copie doit toujours être comprise comme une mesure approximative de la compétence de l'élève ;

– montrer que l'évaluation des compétences des élèves est assurée de façon plus équitable par la multiplication des mesures (ce qui est le cas du bac, et de façon renforcée aux oraux du second groupe d'épreuves avec le recours au livret scolaire) ;
– montrer, indirectement, l'intérêt des mesures standardisées des connaissances afin de pouvoir établir des comparaisons des compétences des élèves entre classes, entre établissements et dans le temps.

Les premiers travaux docimologiques ont toutefois été limités dans leurs investigations par leur construction théorique de l'évaluation scolaire : penser que l'évaluation des copies pouvait aboutir à établir des mesures exactes et, en conséquence, se limiter à constater la non-vérification de cette hypothèse quel que soit d'ailleurs le type de correction, globale ou par critères [6]. Les premières études statistiques n'ont donc pas eu pour objet de rechercher les causes des écarts de notes entre professeurs.

ALÉA DE LA NOTATION ET INFORMATIONS SCOLAIRES ET EXTRA-SCOLAIRES

Les travaux principaux relevant de ce qu'il est habituel de rattacher à la psychologie de l'évaluation scolaire sont dus particulièrement à J.-P. Caverni et G. Noizet [7]. La théorie fondant ces recherches est de considérer que l'évaluateur ne note jamais une copie en soi. La notation est un travail de comparaison de copies qui ne peut pas s'exercer indépendamment d'informations scolaires et extra-scolaires spécifiques. Il importe donc de chercher à connaître les effets de celles-ci sur l'activité d'évaluation.

Les effets d'ordre de correction des copies

L'étude de l'effet de l'ordre de correction des copies, spécifiquement réalisée par J.-J. Bonniol (cité par Caverni et Noizet, *op. cit.*, [7]), constitue une première analyse des effets d'informations scolaires détenues par le professeur sur sa façon de noter. La procédure expérimentale mise en place a consisté à confronter la notation de 26 versions anglaises selon l'ordre de correction (du n° 1 au n° 26 et du n° 26 au n° 1). Les résultats de cette recherche montrent que l'ordre de correction des copies produit deux types d'effet. D'une part, les copies placées dans le premier tiers du paquet sont en moyenne notées de façon plus indulgente (corrigées dans l'ordre inverse, ces copies sont notées plus sévèrement). D'autre part, chaque copie est notée différemment selon sa position à l'égard des autres copies : la même copie est surévaluée lorsqu'elle vient après une copie faible et sous-évaluée lorsqu'elle vient après une copie forte. Dans la procédure de correction, les qualités et insuffisances de la (ou des) copie(s) antérieurement corrigée(s) influencent donc l'appréciation de la copie en cours d'évaluation (l'ajout d'une bonne copie au début d'un paquet provoque une évaluation plus sévère des copies ultérieures, et inversement). Ces données éclairent en partie les modalités d'évaluation des copies : le correcteur établit au début de sa correction un ensemble d'exigences à partir de la lecture des premières copies et ces exigences servent de modèle de référence dans la suite de son travail de correction.

Cette recherche et les précédentes apportent des connaissances sur l'incertitude de l'évaluation des copies dans des situations anonymes de correction. Or la notation des élèves, dans le cadre de la classe, a la particularité de se réaliser dans une situation où l'anonymat des élèves n'est ordinairement pas la règle : les professeurs demandent en effet généralement à leurs élèves de remplir des fiches de renseignements sur lesquelles figurent des informations sur leur cursus scolaire et leur origine sociale (notes obtenues l'année précédente, redoublement, âge, profession des parents…). En dehors d'une telle fiche, les professeurs peuvent de toute façon être amenés à connaître une partie de ces informations. Quelles incidences ces informations scolaires et extrascolaires peuvent-elles avoir sur l'évaluation des compétences scolaires des élèves ? Les chercheurs ont répondu à cette question en créant des situations fictives de correction qui permettent d'isoler l'effet de la possession de telle ou telle information sur la notation des devoirs des élèves.

Effet du niveau scolaire

Caverni, Fabre et Noizet ont ainsi isolé l'effet sur la notation d'un devoir de la connaissance par le correcteur du niveau scolaire de l'élève. Dans une première étape, l'expérience a consisté à faire corriger douze copies d'anglais de baccalauréat de niveau très différent (ces copies étant dactylographiées). Dans une seconde phase (ignorée initialement par les correcteurs), six copies, également dactylographiées, et de niveau cette fois à peu près équivalent, ont été soumises à la correction des mêmes professeurs. Sur chaque copie figurait une note attribuée antérieurement par le correcteur et censée avoir été obtenue par l'élève (ces notes apposées sur les copies étaient soit fortes soit faibles). L'association entre la copie à corriger et la note initialement mise était fictive, mais une telle association a orienté significativement l'évaluation des enseignants : les mêmes copies sont notées de façon sensiblement différentes selon que la

note antérieure, attribuée fictivement, est faible ou forte. Dans cette seconde phase de l'expérience, le correcteur est influencé, lors de l'évaluation de chaque copie, par l'évaluation qu'il a réalisée initialement. Les auteurs de cette recherche concluent sur ces données en indiquant que le modèle de référence de l'évaluateur est constitué *« pour une part déterminante, par l'évaluation précédemment produite par l'évaluateur lui même »*. De la même façon, les auteurs montrent que ce phénomène d'assimilation entre évaluations est également présent lorsque les professeurs disposent d'un dossier fictif de bonnes ou mauvaises notes associé à la copie à évaluer. La confiance que le correcteur accorde à des évaluations antérieures réalisées par d'autres enseignants produit des effets identiques à celle qu'il accorde à ses propres évaluations.

Effet du statut scolaire

Outre le niveau, le statut scolaire de l'élève influence également l'évaluation des performances des écrits. Il y a déjà plus d'un quart de siècle, Bonniol, Caverni et Noizet ont fait procéder à la correction de huit copies provenant pour moitié de sixièmes de type 1 (réputées) et de sixièmes de type 3 (peu réputées). L'étude aboutit à un résultat qui concorde avec les données précédentes : lorsque les huit copies sont attribuées à des élèves de sixième de type 1, elles sont sensiblement mieux notées que lorsqu'elles sont attribuées à des élèves de sixième de type 3. Dans le premier cas, les copies obtiennent plus de 10/20 alors que dans le second cas elles n'obtiennent pas la moyenne. Lors de la notation des élèves, les correcteurs sollicitent une représentation du niveau des élèves sensiblement dépendante de son statut scolaire, en l'occurrence du type de sixième qui détermine la frontière symbolique séparant les élus (une note supérieure à 10 autorise le passage dans l'année supérieure) des réprouvés (la note inférieure à la moyenne signale l'insuffisance scolaire).

Ces deux types d'études portant sur l'effet de la connaissance du niveau et du statut scolaires sur la notation montrent, de façon manifeste, un phénomène de dépendance entre évaluations : toutes informations qui d'une manière ou d'une autre ne s'accordent pas, sont dissonantes, entraînent de la part d'un correcteur un effort pour mieux les faire s'accorder. C'est ce résultat de recherche – le contexte scolaire influence la correcteur – qui justifie qu'aucune information ne figure sur les copies du bac, et plus généralement, sur les devoirs d'examens et de concours.

Effet de l'origine sociale

On rappelle qu'en début d'année, les professeurs font assez souvent remplir à leurs élèves une fiche sur laquelle ceux-ci indiquent généralement la profession de leurs parents. C'est pour cette raison qu'il est utile de connaître les effets de cette information sur l'élaboration du jugement du professeur. Dans cette perspective, R. Weiss (cité par G. de Landsherre, [8]) a sélectionné deux compositions françaises rédigées par des élèves de quatrième. Celles-ci ont été dactylographiées et soumises à la correction de deux groupes de 46 enseignants. Les deux copies ont été transmises au premier groupe de correcteurs en mentionnant que la première copie était rédigée par un enfant doué, fils d'un rédacteur d'un quotidien connu alors que la seconde copie était rédigée par un élève moyen qui aime lire les bandes dessinées et dont le père et la mère sont employés. Au second groupe de correcteurs, les informations extra-scolaires associées à chaque copie étaient inversées.

Les enseignants avaient, pour chaque copie, à noter séparément l'orthographe, le style, le « fond » et l'ensemble de la copie sur une échelle de un (très bien) à cinq (insuffisant). Pour les quatre aspects considérés, la copie qui bénéficie d'un préjugé favorable est significativement mieux notée. Lors de l'appréciation de la qualité orthographique, critère pour lequel existe une définition stricte de l'erreur, 16 % des correcteurs accordent la note « très bien » au travail fictivement attribué à l'élève doué, fils d'un rédacteur connu. Pour la même copie, attribuée à l'élève moyen fils d'employés, aucun correcteur n'accorde la note « très bien ». Ces données sont essentielles : elles montrent que les informations extra-scolaires détenues par les professeurs sont susceptibles d'influencer leurs évaluations. La recherche de Weiss présente toutefois l'inconvénient de ne pas distinguer les effets des deux variables introduites dans l'expérience (statut scolaire de l'élève et origine sociale).

Une étude de J.-P. Pourtois [9] a permis de dépasser les limites méthodologiques du travail de Weiss. L'expérience a consisté à faire corriger par plusieurs correcteurs quatre compositions de français (classe de sixième) associées fictivement à deux types d'informations : « copies des enfants issus de familles socialement défavorisées » ou « copies des enfants issus de familles socialement favorisées ». Les professeurs avaient à noter le « fond », la forme et l'orthographe. Sur ces trois critères, les copies bénéficiant d'un préjugé laudatif font l'objet d'une notation plus clémente (l'écart le plus sensible de notation ne porte pas sur le fond mais de nouveau sur l'orthographe). L'expérience montre clairement que la

variable « appartenance sociale » est susceptible d'agir sur l'appréciation portée sur une copie.

Effet de l'apparence physique et du sexe

Ces explications des aléas de l'évaluation scolaire seraient incomplètes sans la présentation de quelques recherches complémentaires relatives à l'aspect physique des élèves. G. Nilson et L. Nias [10] ont associé à un même livret scolaire des photos représentant des visages jugés plus ou moins attrayants. Des groupes de professeurs avaient à établir des pronostics de réussite scolaire à partir des dossiers qui leur étaient présentés et qui ne se distinguaient que par la photo d'identité qui y était associée. Les livrets associés à un visage agréable reçoivent en moyenne un pronostic de réussite plus favorable. De même les copies associées à des photos d'élèves au visage agréable font l'objet de notations plus clémentes. Les modèles de réussite scolaire sollicités implicitement par les correcteurs intègrent donc des images sociales de la réussite dans lesquelles le succès est associé à l'attrait physique des personnes.

L'aspect physique n'exerce pas des effets évaluatifs indépendamment du sexe. Les évaluations professorales intègrent aussi des attentes socialement différenciées selon cette variable. Ainsi, des enseignants corrigeant des devoirs de sciences physiques associés au hasard à des garçons ou des filles, attribuent en moyenne de meilleures notes aux devoirs attribués fictivement à des garçons. Les appréciations littérales des correcteurs mentionnent aussi plus souvent que la copie est empreinte de *« rigueur »* et d'*« esprit scientifique »* lorsqu'elle est, fictivement, attribuée à un garçon (M.-G. Spear, [11]).

L'intérêt de ce type d'investigation

Ces recherches montrent que les procédures ordinaires de jugement des élèves dans le quotidien de la classe sont empreintes de « biais » (d'erreurs) d'évaluation produits par la détention d'informations scolaires et extra-scolaires sur l'élève. En ce sens, les informations demandées par les professeurs dans les fiches de renseignements qu'ils font remplir par leurs élèves en début d'année favorisent ces biais d'évaluation. En termes d'action sociale, ces recherches montrent l'intérêt des échanges de copies entre professeurs au cours desquels les élèves bénéficient de corrections anonymes. Les élèves sont généralement demandeurs de ce type d'évaluation alors que les professeurs sont plus réticents : ils craignent parfois d'être déjugés par une correction assurée par leurs collègues. En fait, davantage pratiqués, ces échanges de copies favoriseraient l'harmonisation des critères d'évaluation des professeurs dont on sait qu'elle permet de diminuer les écarts d'appréciation selon le correcteur (Caverni et Noizet, *op. cit.*, [7]). Ces études éclairent aussi les débats relatifs à l'intérêt des examens ayant recours à l'anonymat des candidats par rapport à des évaluations fondées sur le seul contrôle continu.

L'intérêt de ces études est donc indéniable. Il faut regretter d'ailleurs qu'elles demeurent relativement peu connues alors même qu'elles permettraient aux professeurs de mieux comprendre les incertitudes attachées à leurs pratiques de notation. Ces approches présentent toutefois deux limites.

D'une part, les résultats de recherche ont été établis à partir de situations fictives de correction et ne permettent pas, à ce titre, de connaître les incertitudes effectives d'évaluation dans le quotidien de la classe. Sur ce point, les travaux de M. Duru-Bellat et A. Mingat [12] ont apporté, à partir de données recueillies sur 2 500 collégiens scolarisés en classe de cinquième, les précisions suivantes :

– les élèves « en retard », mais non redoublants, obtiennent un point de moins par année d'âge à niveau de connaissances données. Les élèves redoublants font également l'objet d'une notation plus sévère à compétence donnée ;

– les enfants de cadres supérieurs se distinguent des enfants des autres groupes sociaux. Pour un niveau donné de résultats aux tests, ils obtiennent en moyenne un demi-point de plus que les enfants des autres milieux. Ce demi-point, biais social d'évaluation, représente presque le quart de la différence moyenne qui sépare les enfants de cadres des enfants d'ouvriers ;

– notées par leurs professeurs, les filles obtiennent en moyenne des notes supérieures de 0,78 point à celles des garçons à compétences identiques aux tests. Il faut noter que cette surnotation des filles n'est pas constatée, en second cycle, quelle que soit la discipline. Elle est, par exemple, en partie vérifiée en français en classe de première [13]), mais n'est pas retrouvée en sciences économiques et sociales (*Cf.* l'article d'É. Chatel dans ce même numéro)[3].

NOTE

3. On notera aussi qu'aucun biais social d'évaluation au détriment des élèves d'origine étrangère n'a été constaté, ce qui peut s'expliquer par la connaissance par les correcteurs de l'existence de préjugés « raciaux » et, en conséquence, par une vigilance accrue de leur part lors de la correction des copies d'élèves d'origine étrangère.

D'autre part, cette approche psychologique sollicite un cadre théorique fondé sur les notions de représentation sociale et d'attente qui expliqueraient la partialité des notes : au cours de la correction, le professeur devient particulièrement attentif à ce à quoi il s'attend si bien que ses attentes sont, pour une part, confirmées. Il s'agit d'une illustration des « effets d'attentes », des « prophéties auto-réalisatrices », mis en évidence par L.F. Jacobson et R.A. Rosenthal à la fin des années soixante [14]. Autrement dit, les professeurs, « victimes » ou « complices » de préjugés sociaux, favoriseraient, par exemple, les élèves issus de milieux aisés ou les bons élèves. Une telle interprétation des données est sans doute un peu rapide. Elle néglige, en raison de la méthode d'investigation utilisée (des situations fictives de correction), une donnée indissociable de l'activité ordinaire d'évaluation : la notation est influencée par les types de relations maître-élèves qui s'établissent dans le quotidien de la classe.

L'ÉVALUATION COMME ARRANGEMENT CONTEXTUEL ET INTERPERSONNEL

Les limites d'une interprétation des incertitudes de l'évaluation scolaire en termes de préjugés sociaux (ceux-ci sont d'ailleurs loin d'expliquer toute l'incertitude de la notation) expliquent le développement d'une théorie de l'évaluation fondée sur l'observation des interactions à l'intérieur de la classe (*op. cit.*, [2]). La réflexion a émergé notamment à partir de la manière dont s'exerce l'autorité professorale. Il s'agit évidemment d'une question centrale puisque cette autorité est une condition indispensable à l'exercice des fonctions du maître, qu'il s'agisse de « faire cours » ou de procéder à l'évaluation des élèves. Or quelques données indiquent que cette autorité connaît une certaine perte d'influence notamment avec le développement de la scolarisation de masse [15]. L'autorité pédagogique ne peut en effet, dans le cadre d'une relation personnalisée en partie routinière, s'appuyer durablement ni sur le charisme du professeur, ni sur l'autorité réglementaire dont il dispose, ni enfin, sur un respect traditionnel de l'institution scolaire auprès d'un public scolaire pour lequel la scolarisation prolongée est souvent une nouveauté familiale [16]. Les enquêtes de la Direction de l'évaluation et de la prospective du Ministère de l'Éducation nationale, de la Recherche et de la Technologie ont d'ailleurs confirmé que cette question de la discipline est sensible pour les jeunes professeurs [17]. Les données et analyses disponibles tendent donc à montrer que les rapports de force symbolique qui seraient au fondement de l'autorité pédagogique semblent perdre de leur efficacité sociale dans les quartiers dits « populaires » notamment. Pas seulement dans ceux-ci d'ailleurs : le comportement utilitariste des lycéens « bourgeois » [18] montre que l'acceptation lycéenne de l'autorité professorale se justifie autant, voire davantage, par une perspective coût-avantage (ou effort-récompense) que par une intériorisation des règles de l'ordre scolaire.

Différents types d'arrangements évaluatifs

Dans cette perspective de recherche, la notation a évidemment toujours pour objet de classer les élèves selon leurs performances scolaires ; mais les incertitudes associées à la note sont expliquées par l'existence d'arrangements qui tiennent au contexte scolaire d'évaluation, appréhendé à quatre niveaux définis par des collectivités (établissement, classe), la relation duale maître-élève et la personne du maître. Ces niveaux d'arrangements sont distingués pour des raisons heuristiques, ils sont en fait en forte interdépendance mutuelle.

Primo, les arrangements au niveau de l'établissement. Dans une recherche précitée réalisée sur 17 établissements (*op. cit.*, [12]), on pouvait s'attendre à ce que les établissements dans lesquels les élèves obtiennent des résultats globalement faibles aux tests de compétence soient ceux dans lesquels les notes moyennes attribuées par les professeurs étaient également basses. Il n'en est rien. Les notations les plus indulgentes sont attribuées plutôt aux élèves qui ont obtenu les résultats les plus faibles aux tests de compétence, et inversement. Des approches ethnographiques aboutissent à des conclusions du même ordre. Ainsi dans les collèges « difficiles », une partie des enseignants sont amenés à des adaptations sensibles de leurs pratiques ordinaires d'évaluation, en évitant des contrôles qui aboutiraient à des notes jugées très faibles, ou par exemple en adoptant une notation de la participation orale qui n'a que pour seul objectif d'augmenter la moyenne des élèves [19].

Secundo, la classe. Des arrangements de ce type se réalisent lorsque le professeur, prenant en considération la bonne volonté de « ses » élèves, décide de supprimer de la moyenne trimestrielle les notes d'un contrôle peu réussi, ou d'ajouter un devoir « facile » en fin de trimestre, etc. Dans la situation inverse – agitation, travail non fait, chahut – le professeur peut avoir recours à une « interrogation surprise » ou donner un devoir « difficile », sorte de sanction

pour montrer aux élèves les conséquences de leur manque d'attention en cours.

Tertio, ces arrangements concernent également les élèves considérés individuellement : outre l'octroi d'une note de participation en cours, l'élève qui accepte de faire un travail supplémentaire, un exposé par exemple, ou de refaire un devoir « raté », pourra bénéficier d'une note supplémentaire ou d'une note se substituant à ce devoir, ou bien encore d'une note qui ne sera prise en compte dans la moyenne que si elle dépasse 10/20, etc. (*op. cit.*, [2]). La dispersion des comportements des professeurs est grande dans ce domaine ; ils peuvent être, en effet, plus ou moins sensibles au sentiment d'iniquité que les élèves faibles ressentent à l'égard de leurs notes lorsque ceux-ci ont le sentiment que la récompense que constitue la note obtenue n'est pas à la hauteur des efforts fournis [20]. Dans ce type de situation particulière, le professeur peut être amené, au nom d'un équilibre nécessaire entre travail et gratification scolaires, à noter davantage les progrès réalisés par l'élève que le niveau atteint et normalement visé à tel ou tel niveau de scolarité.

Quarto, ces arrangements individuels qui engagent de façon personnelle élève et maître sont indissociables, pour le maître, d'un arrangement par rapport à soi qui constitue une quatrième forme d'arrangement : la notation du professeur est orientée par sa propre histoire scolaire et par les diverses significations que celui-ci associe à son activité de notation (« juger de façon impartiale », aider, récompenser, sanctionner…). Précisons enfin qu'une part des professeurs ont une certaine conscience de ces arrangements, et c'est notamment pour cette raison qu'ils sont favorables au maintien du bac qui constitue, dans leur propos, une garantie d'équité scolaire.

Ces arrangements évaluatifs sont d'une grande variété et sont d'autant plus fréquents que l'enseignant est certes confronté à la question de l'ordre scolaire, à la gestion des relations maître-élèves dans la classe, mais simultanément à la question de la transmission du savoir et des rythmes d'apprentissage. Cette contrainte de l'action enseignante explique que les arrangements évaluatifs intègrent inévitablement une dimension didactique. La notation au demi-point près ne doit pas être comprise dans le cadre de la *« note vraie »* espérée par Laugier et Weinberg (*op. cit.*, [1]), mais d'abord comme le résultat d'une *« transaction »* ou d'un *« contrat »* de type didactique [21]. Le 9,5/20 n'exprime pas tant la précision de la mesure des performances qu'une sorte d'avertissement symbolique dont l'objet est de signaler à l'élève que celui-ci ne remplit pas totalement les exigences attendues spécifiques à sa classe et son établissement.

Arrangements et biais sociaux d'évaluation

Dans cette perspective, les notes scolaires deviennent un motif de la mobilisation des élèves, un moyen de paix scolaire dans et hors de la classe, et aussi une dimension majeure de l'autorité du professeur qui dispose, plus ou moins consciemment, d'une certaine marge de manœuvre lorsqu'il accorde ou non une reconnaissance scolaire par la note. Cette analyse offre un cadre théorique pertinent à l'analyse de l'incertitude de la notation et des biais sociaux d'évaluation précédemment présentés. La mise en parallèle des comportements en classe des élèves et des évaluations de leurs écrits permet en effet de comprendre au moins partiellement, les biais d'évaluation constatés. Ainsi les comportements scolaires des filles, davantage conformes aux attentes professorales [22], expliqueraient leur surévaluation scolaire à compétences équivalentes aux garçons. Et le comportement plus agité, voire contestataire, des garçons d'origine populaire [23] expliquerait leur notation en moyenne un peu plus sévère. Inversement, les élèves d'origine aisée, dont on sait qu'ils ont une meilleure maîtrise du « métier » d'élève et qu'ils prennent plus facilement la parole en classe [23] [24] bénéficient d'arrangements « individuels » en leur faveur. L'analyse des entretiens menés auprès des enseignants montre d'ailleurs que dans le quotidien de la classe se contruisent des images scolaires d'élèves plus ou moins favorables à cette reconnaissance professorale. Les deux exemples suivants (*op cit.*, [2]) présentent la façon dont des comportements d'élèves particuliers peuvent orienter la dynamique relationnelle des interactions maître-élèves :

« C'est vrai que tu as des élèves qui sont épouvantables aussi, ça arrive : j'avais une classe, il y avait deux mecs, un jour à l'intercours, j'étais seule avec eux parce que les autres étaient sortis, et il y en a un qui a dit en rigolant à son copain : " on se la coince ". Des mecs comme ça, aucune pitié. Je crois qu'il n'a pas mesuré la distance à laquelle je tenais. Pour lui dans sa tête, ça n'était même pas insultant, parce que c'est comme ça qu'il doit traiter les filles par ailleurs, mais je m'en fous, je n'ai pas non plus une pitié infinie pour ces pauvres petits enfants d'ouvriers » (professeur femme, classe de terminale).

« J'ai de bonnes élèves, j'ai trois petites gamines qui ont vraiment une tête d'ange, qui sont vraiment…, qui sont adorables, qui sont bosseuses, qui répondent, qui connaissent plein de trucs. Et c'est sûr que quand j'arrive à leur copie, j'ai un préjugé positif, et je le sens. Alors quelquefois, je me

dis, oh là là ! Est-ce que tu ne l'as pas surnotée ? » (professeur femme, classe de troisième).

Les biais d'évaluation et plus généralement les incertitudes de la notation renvoient, dans cette perspective, non à une interprétation en termes de préjugés sociaux des professeurs, mais à une question de police scolaire qui prend la forme de sanctions et de récompenses par la notation selon les comportements adoptés en classe (le zéro pour « mauvaise conduite » ou « travail non fait » en sont des illustrations emblématiques)[4]. En fait, l'analyse des pratiques pédagogiques et des interactions en classe montrent que celles-ci peuvent relever d'une explication en termes de préjugés sociaux lorsque le professeur déclare faire *« surtout attention aux bons élèves »*, et exprime de *« la répugnance pour les parents de milieux populaires »* [25]. Mais les pratiques déclarées par les maîtres peuvent aussi s'opposer à ces préjugés lorsque le professeur indique faire *« attention aux moins bons »*, et *« recherche le contact avec les parents de milieu populaire »* (*op. cit.*, [25]). Toute la question est évidemment de savoir, en l'absence d'investigation spécifique, si les pratiques en classe sont conformes aux intentions déclarées et si le modèle de l'évaluation comme arrangement fournit un cadre d'interprétation des données plus large et plus pertinent qu'une explication en termes de préjugés sociaux. Le modèle de l'évaluation comme arrangement présente toutefois un triple avantage : mieux rendre compte de la diversité des situations scolaires appréhendées notamment en termes de différence de sélectivité scolaire et sociale des établissements ; rendre plus facilement intelligible la réussite scolaire en milieu populaire (trajectoires sociales peu compatibles avec les explications en termes de préjugés sociaux ou « d'école reproductrice ») ; être conciliable avec la diversité des origines socio-professionnelles des enseignants [26], donnée peu compatible avec des représentations et attentes professorales qui seraient partagées par tous de façon identique.

Arrangements évaluatifs et performances scolaires

Ce serait se tromper sur la signification sociale de ces arrangements évaluatifs que de les juger négativement. Dans les situations les plus ordinaires et les plus fréquentes, ces arrangements constituent une façon de « tenir » les élèves et de favoriser leur mobilisation scolaire. Quelques professeurs emploient d'ailleurs le terme de « notes thérapeutiques » pour désigner cette pratique (*op. cit.*, [2]). Laxisme ? Nullement. G. Felouzis [27] a récemment montré que ces « indulgences calculées » favorisaient les progressions des élèves aussi bien en mathématiques qu'en français. Il s'agit d'une modalité spécifique des « effets d'attentes » : la bonne note et la reconnaissance parentale apportée par la réussite aux devoirs sont des sources d'encouragement, favorisent la mobilisation scolaire, redonnent du sens au travail et aux études et finalement créent les conditions d'une amélioration des compétences scolaires. *A contrario*, une notation sévère est plutôt source de découragement scolaire et aboutit le plus souvent à freiner le rythme moyen de progression des élèves. Ainsi, pour un fils d'employé, d'âge normal, ayant 12/20 aux épreuves communes de mathématiques de début d'année, son score de fin d'année sera de 13,2 en cas d'évaluations sévères de son professeur au premier trimestre, de 14,3 si l'évaluation en classe est dans la moyenne des évaluations professorales réalisées pour ce niveau d'élève, et de 15,6 en cas d'évaluations indulgentes du professeur. Le niveau de la note exerce des effets encore plus sensibles en français : le score en fin d'année aux tests standardisés du même élève moyen varie de 12,1 à 15,2 selon la notation du professeur au premier trimestre (*op. cit.*, [27]). Autrement dit, l'évaluation sommative est aussi formative, pas seulement en raison des conseils qui peuvent être apportés lors de la correction et dont on connaît l'effet bénéfique sur les acquis cognitifs des élèves [28] [29], mais aussi par le niveau même de la note, plus ou moins source de mobilisation scolaire.

Si certains types d'arrangements évaluatifs semblent exercer des effet positifs sur le niveau des performances scolaires, ces arrangements ne sont pas observés cependant quel que soit l'établissement d'exercice. Ainsi, dans certains « grands lycées » en centre-ville, des moyennes généralement assez basses sont attribuées à des lycéens de niveau « moyen » ou « juste » afin d'assurer, *via* le redoublement ou le changement d'établissement, un taux de réussite au bac proche de 100 %. La publication des palmarès des établissements par la presse (les indicateurs bruts de réussite au bac) favorise la concurrence entre établissements, et incite les chefs d'établissement à une vigilance accrue lors du passage en classes de première et terminale. Ce contexte scolaire particulier suscite l'élitisme et, par ricochet, des notations plus

NOTE

4. On notera que l'arrêté du 5 juillet 1890 relatif au régime disciplinaire des lycées et collèges de garçons, en vigueur jusque dans les années quatre vingt, définit « la mauvaise note », comme premier type de punition autorisée (art. 2).

sévères au détriment des élèves faibles ou moyens[5]… Bref, le projet d'être un « bon établissement », au sens très réducteur, mais commun, donné par la publication des taux bruts de réussite au bac, peut être contradictoire avec des arrangements évaluatifs, internes à la classe, et favorables aux progrès des élèves.

■■■

Les relations maître-élèves qui se nouent au moment de l'évaluation ne se distinguent guère fondamentalement des autres types de relations sociales, et leur analyse pose les mêmes problèmes sociologiques. En ce sens, comparer, comme le faisait Durkheim au début du siècle, la classe à « une petite société » [30] reste toujours aussi judicieux. C'est pour cette raison que les confrontations théoriques qui agitent l'analyse sociologique se retrouvent au sein de la sociologie de l'école, et au sein d'une activité telle que l'évaluation des élèves. De façon simplifiée, l'évaluation des élèves peut être analysée en tant que modalité de reproduction (Cf. la thèse du même nom), elle peut aussi faire l'objet d'une approche fonctionnaliste et stratégique en termes de création et maintien de zones d'incertitude par les professeurs et les élèves dans le cadre de l'organisation scolaire [31], ou l'objet d'une analyse interactionniste focalisée sur les processus et les dynamiques interpersonnels propres au quotidien des relations maître-élèves [32] [33]. Par rapport à ces orientations théoriques, le terme « d'arrangement évaluatif » est une notion théorique « molle ». Elle penche plutôt vers une conception interactionniste de l'activité d'évaluation en accordant une attention privilégiée à l'*« ici et maintenant »*. Mais le terme d'arrangement, qui évoque clairement « l'ordre négocié » des interactionnistes ne signifie pas que les acteurs sociaux participent de façon identique à la « définition de la situation », et le recours à ce terme ne postule pas également ignorance des inégalités structurelles de statut et de pouvoir à l'intérieur de l'institution scolaire. L'usage du terme, équivoque, met l'accent sur la diversité des situations concrètes et sur les processus qui les constituent, ainsi que sur les dynamiques contradictoires et parfois conflictuelles qui animent l'institution.■

NOTE

5. On rappelle que le taux brut de réussite au bac est un indicateur très fruste qui ne permet pas de mesurer l'efficacité d'un établissement. La Direction de l'évaluation et de la Prospective du ministère de l'Éducation nationale, de la Recherche et de la Technologie a mis au point des indicateurs qui, prenant en compte les caractéristiques sociodémographiques des élèves accueillis en classe de seconde (origine sociale, âge moyen) et le taux de sélection au cours de la scolarité, montrent que des lycées peuvent être efficaces avec des résultats moyens au bac et que des établissements pourtant jugés « excellents » ne sont pas forcément les plus efficaces (MEN-DEP, 1996).

À LIRE

[1] **H. LAUGIER, D. WEINBERG,** *Commission française pour l'enquête Carnegie sur les examens et concours. La correction des épreuves écrites au baccalauréat*, Paris, Maison du Livre, 1936.

[2] **P. MERLE,** L'évaluation des élèves. *Enquête sur le jugement professoral*, Paris, P.U.F., 1996.

[3] **M. COMTE, X. POULARD,** « Trois ans après l'entrée à l'université : parcours suivis et diplômes obtenus. L'exemple des bacheliers 1991 de la région Rhône-Alpes », revue *Éducation & Formations*, n° 50, Direction de l'évaluation et de la prospective, Ministère de l'Éducation nationale, de la Recherche et de la Technologie, juin 1997, 33-39.

[4] **J.-P. JAROUSSE, A. MINGAT, D. OGET,** « Réflexions pour un changement de l'organisation des épreuves du bac », *Savoir. Éducation formation*, à paraître.

[5] Revue *Les amis de Sèvres,* n° 2, 1968, 10-11.

[6] **A. BONBOIR,** *La docilomologie,* Paris, P.U.F., 1972.

[7] **J.-P CAVERNI, G. NOIZET,** *Psychologie de l'évaluation scolaire*, Paris, P.U.F., 1978

[8] **G. de LANDSHERRE,** Évaluation continue et examens. Précis de docimologie, Éditions Labor, 1976.

[9] **J.-P. POURTOIS,** « Le niveau d'expectation de l'examinateur est-il influencé par l'appartenance sociale de l'enfant ? », *Revue française de pédagogie,* n° 44, 1978, 34-37.

[10] **L. NIAS, G.NILSON**, *Le charme a ses raisons,* Paris, Tchou, 1977.

[11] **M. G. SPEAR**, « Sex Biais in Science Teachers'Rating of Work and Pupils Characteristics », *European Journal of Science Education,* 6, 4, 1989, 369-377.

[12] **M. DURU-BELLAT, A. MINGAT**, *Pour une approche analytique du fonctionnement du système éducatif,* Paris, P.U.F., 1993.

[13] **P. MERLE,** L'adhésion des lycéennes de terminale C au modèle de l'excellence scolaire, *Sociétés contemporaines,* n° 16, 1993, 7-26.

[14] **L. F. JACOBSON, R. A. ROSENTHAL**, *Pygmalion in the Classroom. Teachers Expectations and Pupil Intellectual Development,* New-York, Holt, Rinehart and Winston, 1968. Trad. fr. : *Pygmalion à l'école*, Paris, Casterman, 1972.

[15] **J. TESTANIÈRE,** « Chahut traditionnel et chahut anomique dans l'enseignement du second degré », *Revue française de sociologie,* VIII, 1967, 17-33.

[16] **M. HIRSCHHORN,** *L'ère des enseignants,* Paris, P.U.F., 1993.

[17] **N. ESQUIEU, S. PÉAN,** « Les débuts dans le métier des nouveaux professeurs. Bilan des deux premières années d'exercice », *Note d'Information,* 97.25, Direction de l'évaluation et de la prospective, Ministère de l'Éducation nationale, de la Recherche et de la Technologie, juin 1997.

[18] **F. DUBET,** *Les lycéens,* Paris, Seuil, 1991.

[19] **A. VAN ZANTEN,** *Les carrières enseignantes dans les collèges difficiles,* in Bourdon J., et Thélot C. (éds.) Recherches et politiques éducatives, Séminaire DEP – IREDU, Éditions du CNRS, 1998, à paraître.

[20] **P. MERLE,** *Équité et notation* : l'expérience subjective des lycéens (à paraître).

[21] **Y. CHEVALLARD,** « Vers une analyse didactique des faits d'évaluation », *in* J.-M. De Kettle, *L'évaluation : approche descriptive ou prescriptive ?* Bruxelles, De Boeck, 1991, 31-59.

[22] **M. DURU-BELLAT,** « Filles et garçons à l'école, approches sociologiques et psycho-sociales », *Revue française de pédagogie,* 110, 1995, 75-109.

[23] **G. FELOUZIS,** *Le collège au quotidien*, Paris, P.U.F., 1994.

[24] **R. SIROTA,** *L'école primaire au quotidien,* Paris, P.U.F, 1988.

[25] **L. DEMAILLY,** « Contribution à une sociologie des pratiques pédagogiques des enseignants », *Revue française de sociologie,* XXVI, 1, 1985, 96-119.

[26] **C. THÉLOT,** *L'évaluation du système éducatif*, Paris, Nathan, 1993.

[27] **G. FELOUZIS,** *L'efficacité des enseignants. Sociologie de la relation pédagogique,* Paris, P.U.F., 1997.

[28] **S. SCHMITT-ROLLAND, M. THAUREL-RICHARD,** « Pratiques pédagogiques de l'enseignement du français en sixième et progrès des élèves, » *Note d'Information*, 96.39, Direction de l'évaluation et de la prospective, Ministère de l'Éducation nationale, de la Recherche et de la Technologie, septembre 1996.

[29] **M. THAUREL-RICHARD, R. VERDON,** « Pratiques pédagogiques de l'enseignement des mathématiques en sixième et progrès des élèves », *Note d'Information,* 96-44, Direction de l'évaluation et de la prospective, Ministère de l'Éducation nationale, de la Recherche et de la Technologie, octobre 1996.

[30] **É. DURKHEIM,** L'éducation morale, Paris, P.U.F., 1992.

[31] **M. CROZIER, E. FRIEDBERG,** *L'acteur et le système. Les contraintes de l'action collective,* Paris, Seuil, 1977.

[32] **P. BERGER, T. LUCKMANN,** *La construction sociale de la réalité,* Paris, Méridiens Klincksieck, 1994.

[33] **P. WOODS,** *L'ethnographie de l'école,* Paris, Colin, 1990.

[34] « Trois indicateurs de performances des lycées », *Les dossiers d'Éducation et Formations,* n° 66, Direction de l'évaluation et de la prospective, Ministère de l'Éducation nationale, de la Recherche et de la Technologie, mars 1996.

Épreuves du certificat d'études primaires en 1995

Étude de quelques facteurs ayant pu agir sur les résultats des élèves

Grâce à un fonds archivé de copies d'épreuves du certificat d'études primaires datant des années 20, on a pu comparer les résultats obtenus en dictée et rédaction par les élèves de l'époque avec ceux d'élèves d'aujourd'hui – de niveau équivalent – en faisant passer à ces derniers les mêmes épreuves. De cette enquête, réalisée en 1995, il ressort que les compétences en rédaction des élèves sont aujourd'hui au moins aussi bonnes, pour ne pas dire meilleures, que par le passé, tandis que le niveau en orthographe semble avoir beaucoup baissé.

Cette constatation doit être notablement nuancée. L'analyse révèle en effet plusieurs éléments explicatifs de ces différences manifestes, et les ramène à de plus justes proportions. Elle éclaire aussi, pour les élèves de 1995, les écarts observés, parfois illogiques a priori, entre leurs notes de contrôle continu et celles qu'ils obtiennent à des épreuves proposées il y a soixante-dix ans et plus.

Anne-Marie CHARTIER
Service d'Histoire de l'Éducation, INRP

En 1995, la Direction de l'évaluation et de la prospective (DEP) du Ministère de l'Éducation nationale a construit une situation permettant à une population d'élèves représentative de la France métropolitaine (2 876 sujets) de passer trois épreuves proposées au certificat d'études primaires (CEP) dans les années 1923, 1924 et 1925, pour lesquelles nous possédons des copies de l'époque[1]. Il s'agissait d'une dictée suivie de questions, d'une rédaction et de deux problèmes. Les élèves passaient alors le certificat d'études en fin de parcours scolaire, à une époque où la scolarité s'achevait à 13 ans. Compte tenu de l'âge des élèves et des programmes actuels, on a choisi de faire composer des classes de sixième, cinquième et, pour une petite part, de quatrième. Une première présentation globale des résultats a été faite dans une publication parue en 1996[2]. Cependant, beaucoup de données n'avaient pu alors être exploitées et interprétées en détail. Nous avons décidé de revenir sur certaines d'entre elles pour mieux comprendre les informations apportées par cette comparaison à 70 ans de distance.

Les deux épreuves que nous avons cherché à étudier de plus près sont la dictée et la rédaction.

NOTES

1. On avait également constitué un échantillon représentatif d'élèves de la Somme qui était le département dans lequel avaient été recueillies les copies du certificat des années 1920. On a pu ainsi vérifier que les résultats de ce département n'étaient pas déviants par rapport au reste de la France.
2. *Connaissances en français et en calcul des élèves des années 20 et d'aujourd'hui,* les dossiers d'Éducation et formations, n° 62, MEN-Direction de l'évaluation et de la prospective, février 1996.

En effet, il s'agit des deux exercices qui existent encore aujourd'hui, l'un inchangé (la dictée), l'autre sous une nouvelle appellation (expression écrite). En revanche, les questions de dictée, qui mêlaient les questions sur la compréhension du texte, la langue et la grammaire, ne se retrouvent plus sous cette forme aujourd'hui. Les problèmes d'arithmétique, avec leur libellé réaliste, mettant en scène des situations concrètes, ne ressemblent plus aux exercices donnés en classe de sixième, cinquième et quatrième en mathématiques.

Les conclusions de l'enquête sont très nettes : les élèves d'aujourd'hui obtiennent des résultats beaucoup moins bons en dictée que dans les années 1920, alors qu'en rédaction, ils sont aussi bons, voire meilleurs. Si les notes en rédaction sont comparables, les textes d'aujourd'hui sont bien plus longs qu'autrefois en moyenne. Comment interpréter ces résultats ? Deux hypothèses sont possibles. On peut penser que le niveau d'orthographe des élèves s'est effondré tandis que leur compétence à écrire s'est plutôt renforcée. On peut aussi se dire que ce sont les exercices qui sont en cause. La dictée, exercice typique de l'école primaire, ne serait pas suffisamment pratiquée dans l'enseignement secondaire et les élèves échoueraient, faute d'entraînement. En revanche, de la rédaction à l'expression écrite, il y aurait changement d'appellation plutôt que de forme, si bien que les élèves s'y adapteraient plus facilement.

Pour essayer de répondre à ces questions, nous avons comparé systématiquement les notes que les élèves d'aujourd'hui ont obtenues aux épreuves du certificat avec celles qu'ils ont obtenues la même année en contrôle continu, en orthographe et en expression écrite. Nous avons essayé d'analyser les facteurs qui influent sur la réussite ou l'échec dans les deux épreuves du certificat, pour revenir, en fin d'analyse, sur la comparaison avec les élèves des années 1920[3].

L'ÉPREUVE DE DICTÉE

Décalage entre les notes de dictée à l'épreuve du CEP et au contrôle continu d'orthographe en 1995

Nous possédons pour chaque élève la note d'orthographe obtenue en classe en contrôle continu en 1995. Nous l'avons comparée à la note obtenue à la dictée du CEP (graphique 1).

Au contrôle continu, la moyenne est de 10/20. La courbe obtenue, de type « gaussien », est très écrasée, indiquant une très forte dispersion des notes des élèves, autour de la moyenne. Cette note est elle-même une moyenne, obtenue par cumul de plusieurs notes d'évaluation pouvant se compenser entre elles. Elle reflète les performances ordinaires des élèves, qui sont réparties presque régulièrement sur toute l'échelle de notation. Les élèves qui ont 10/20 de moyenne représentent 7 % de l'échantillon, c'est-à-dire ni plus ni moins que ceux qui ont 8, 9, ou 11, 12 et 13.

NOTE

3. Les analyses qui ont permis la rédaction de cette étude ont été réalisées en 1996 par Floriane Aurousseau (MEN-DEP-Département de l'évaluation des élèves et des étudiants).

GRAPHIQUE 1 – Distribution des notes de dictée à l'épreuve du CEP et au contrôle continu d'orthographe en 1995

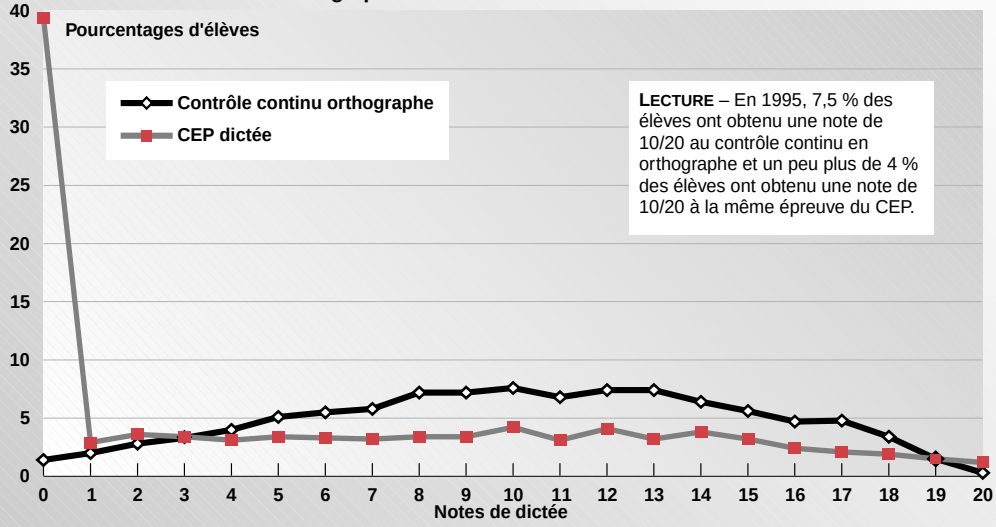


TABLEAU 1 – Liaison entre les résultats d'orthographe au contrôle continu et de dictée aux épreuves du CEP (élèves de l'échantillon France entière en 1995)

Note contrôle continu	Note CEP								
	0 à 2,4	2,5 à 4,9	5 à 7,4	7,5 à 9,9	10 à 12,4	12,5 à 14,9	15 à 17,4	17,5 à 20	**Ensemble**
0 à 2,4	90,9 %	5,1 %	2,3 %	1,7 %	0,0 %	0,0 %	0,0 %	0,0 %	**100,0 %**
2,5 à 4,9	85,8 %	5,6 %	4,1 %	1,9 %	2,6 %	0,0 %	0,0 %	0,0 %	**100,0 %**
5 à 7,4	78,6 %	3,1 %	8,7 %	4,3 %	3,1 %	1,2 %	1,0 %	0,0 %	**100,0 %**
7,5 à 9,9	50,8 %	12,6 %	12,6 %	8,3 %	10,0 %	3,7 %	2,0 %	0,0 %	**100,0 %**
10 à 12,4	38,0 %	10,2 %	13,0 %	10,9 %	12,9 %	7,1 %	5,8 %	1,5 %	**100,0 %**
12,5 à 14,9	21,1 %	6,7 %	12,7 %	12,1 %	18,1 %	13,2 %	11,4 %	4,7 %	**100,0 %**
15 à 17,4	11,8 %	4,7 %	5,5 %	8,8 %	17,9 %	19,0 %	19,3 %	0,1 %	**100,0 %**
17,5 à 20	3,4 %	0,7 %	2,7 %	4,7 %	9,5 %	16,9 %	24,3 %	37,8 %	**100,0 %**

LECTURE – 90,9 % des élèves qui avaient une note de contrôle continu comprise entre 0 et 2,4 sur 20 ont eu au CEP une note comprise entre 0 et 2,4 sur 20.

Comment cette note de contrôle continu en orthographe a-t-elle été obtenue ? D'après les professeurs interrogés, c'est une moyenne d'exercices variés. Elle peut provenir de dictées, qui sont soit des dictées de contrôle, notées comme celle du CEP, soit des dictées plus ou moins préparées. D'autres notes proviennent d'exercices d'orthographe plus ponctuels, plus courts, qu'il est plus facile de réussir (exercices d'orthographe grammaticale, exercices portant sur une règle lexicale particulière, etc.). La note de contrôle continu reflète donc aussi bien des compétences acquises que des savoirs en cours d'acquisition, dans des épreuves de performance complexes (dictées non préparées) et dans des exercices ponctuels (application d'une règle).

Pour la dictée du CEP, la courbe obtenue est en L, avec près de 40 % des élèves obtenant la note zéro. Les notes différentes de zéro se répartissent sur une courbe pratiquement plate, puis en pente légèrement descendante entre 12 et 20. La moyenne est à 5,7 sur 20, la médiane est autour de 2,5. Si la notation pouvait faire usage de notes négatives, on retrouverait probablement une courbe de type gaussien déportée sur la gauche (certains élèves, ayant fait plus de 10 fautes d'accord, auraient alors des notes inférieures à - 20).

Si l'on suit le barème de notation qui est généralement en usage[4] (en notation sur 20 : deux points pour les fautes de grammaire, deux points pour les fautes de langue, un point pour les fautes d'usage, sauf si la prononciation est changée, un demi-point pour les fautes d'accent), l'épreuve de dictée dépasse donc largement le niveau des élèves[5]. Les élèves n'ayant pas obtenu 0/20 se partagent en deux moitiés presque égales, ayant obtenu soit plus de 10/20 (30 % de l'effectif, 858 sujets) soit moins (presque un tiers a zéro sur vingt).

Dans les années 1920, 50 % des élèves seulement étaient présentés au certificat d'études (ceux que l'on considérait comme susceptibles d'être reçus). On pourrait donc croire que les élèves ayant eu zéro à leur dictée auraient fait partie du lot qu'on n'aurait pas présenté, puisqu'ils auraient été considérés comme éliminés d'avance. En fait, ceci serait vrai si la corrélation entre note au contrôle et note à la dictée était très forte. Qu'en est-il exactement ?

Ce n'est pas ce que l'on peut voir en regardant les résultats des élèves ayant eu la moyenne au contrôle continu (tableau 1). Parmi ceux qui ont entre 10 et 12,5 au contrôle continu, 27,3 % seulement ont la moyenne en dictée, soit moins d'un sur trois et pratiquement 30 % ont eu zéro (38 % ont eu entre 0 et 2,4). Certes, la probabilité d'avoir la moyenne en dictée augmente parmi les élèves qui ont une bonne note au contrôle continu. Parmi ceux qui ont entre 12,5 et 15 de moyenne, 50 % ont plus de 10/20 en dictée. On monte à 70 % pour ceux qui ont entre 15 et 17,5 et à 90 % pour ceux qui ont entre 17,5 et 20. Il y a donc bien un rapport entre la compétence repérée par la note au contrôle continu et la performance en dictée. Cependant, même parmi les bons ou très bons élèves, une petite minorité a une très mauvaise note : environ 15 % de ceux qui ont plus de 12,5 au contrôle continu ont eu 0 en dictée.

On peut, bien sûr, expliquer ce fait par les aléas d'une épreuve unique, où les difficultés sont imprévisibles et où les fautes d'inattention coûtent cher. En effet, les fautes d'orthographe ont deux origines : soit l'élève se trompe parce qu'il ne connaît pas l'orthographe du mot (orthographe d'usage ou règle d'accord), soit il la connaît mais ne l'écrit pas

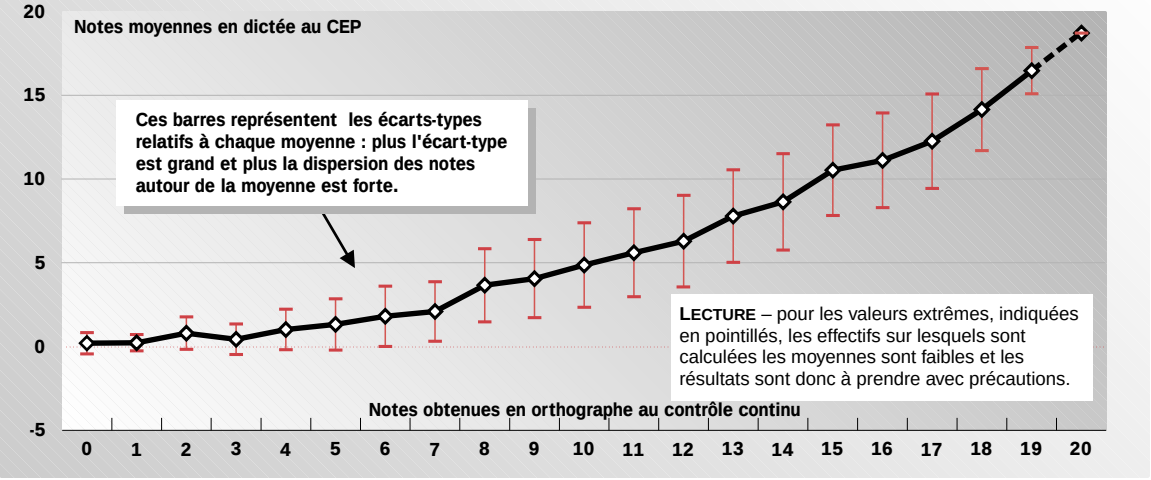
NOTES

4. C'est celui du brevet des collèges.

5. Rappelons qu'au certificat d'études, le barème n'était pas celui-là (la notation se faisait sur 5 et on enlevait un point par faute grave. Comme le zéro était une note éliminatoire, les correcteurs de 1920 étaient relativement indulgents). Pour permettre la comparaison, toutes les dictées des années 1920 ont été recorrigées par les évaluateurs de 1995.

THÈME

GRAPHIQUE 2 – Croisement entre la note obtenue en orthographe au contrôle continu et celle obtenue à la dictée du CEP



(il ne retrouve pas la bonne écriture ou bien oublie d'appliquer la règle). Par exemple, alors que la règle d'accord du sujet et du verbe est bien respectée quand les deux mots sont voisins, elle l'est beaucoup moins quand ils sont séparés (par un groupe de mots, une proposition, etc.), ce qui brouille l'automatisme. Ainsi, la vigilance nécessaire à la réussite de l'épreuve[6] n'est plus acquise aujourd'hui, même par les meilleurs élèves. Ce qui est plus difficile à expliquer, ce n'est pas que des élèves, même bons ou excellents, puissent faire une mauvaise performance en dictée, c'est que de mauvais élèves, voire de très mauvais (ayant eu moins de 7,5/20 au contrôle continu) puissent obtenir une note en dictée meilleure qu'au contrôle continu (ce qui est néanmoins beaucoup plus rare, entre 9 % et 10 % des élèves). Des phénomènes de ce type peuvent s'expliquer si l'élève a particulièrement « fait attention », faisant remonter sa note (alors qu'il ne peut, dans une dictée, acquérir ou manifester des savoirs qui lui manquent en contrôle continu). Ce fait peut aussi provenir de la façon dont les professeurs construisent la note de contrôle continu, tantôt de façon très exigeante (en ne comptant par exemple que des dictées de contrôle) tantôt plus généreuse (en cumulant les notes obtenues aux dictées et aux exercices d'orthographe). Cette marge d'aléas est mise en évidence par le graphique indiquant l'importance des écarts-types et donc la dispersion des notes en contrôle continu qui correspondent à chaque note obtenue en dictée (graphique 2).

On voit donc qu'il existe :

– une probabilité de réussite à la dictée en corrélation avec la réussite au contrôle continu ;

– un décalage dans les notations, explicable par le fait qu'on ait d'un côté une note unique et de l'autre des épreuves agrégées de type varié ;

– une marge importante d'aléas, imputable aux fautes d'inattention autant qu'à la méconnaissance de l'orthographe et aux variations entre les échelles de notes utilisées d'un enseignant à l'autre.

NOTE

6. Elle était courante dans les années 1920 pour la population présentée au certificat, puisque 24 % des candidats faisaient 0 ou 1 faute. Ils sont cinq fois moins nombreux aujourd'hui (en ne retenant qu'une demi-classe d'âge, la meilleure, pour avoir une population comparable à celle qui passait jadis le CEP).

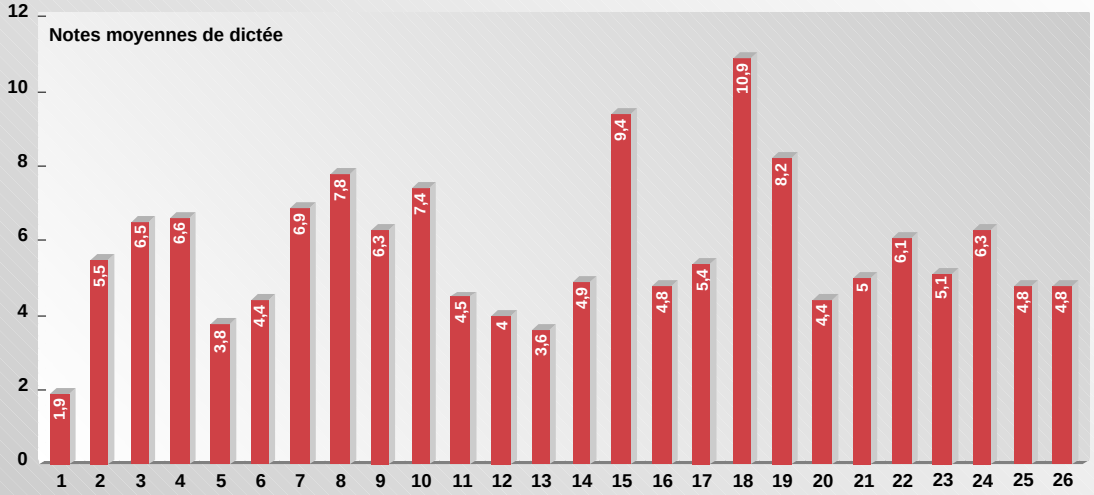
La nature de la dictée influe sur les résultats

Les épreuves étaient constituées par 26 dictées différentes dont le texte est donné dans la publication citée plus haut. Selon les cas, les résultats sont très contrastés, puisque la moyenne va de 1,9/20 à 10,9/20 (graphique 3). Certaines dictées étaient donc bien plus difficiles que d'autres. Plusieurs éléments permettent de hiérarchiser ces difficultés.

➔ *La longueur de la dictée*

Un des premiers facteurs auquel on pense est la longueur du texte. Rappelons que la longueur habituelle d'une dictée en fin de CM2 est de 80 mots. La longueur moyenne d'une dictée au brevet des collèges en fin de troisième est aujourd'hui de 160-170 mots. On sait que l'attention des élèves se relâche au fur et à mesure de la dictée et il n'est pas rare que les fautes s'accumulent dans les dernières lignes. Là encore l'entraînement à écrire joue un rôle essentiel. D'après les témoignages sur l'école des années 20 (comme les

GRAPHIQUE 3 – Notes moyennes de dictée au CEP selon le sujet



cahiers d'élève), les élèves se préparant au certificat avaient droit à une dictée de façon quasi quotidienne et étaient très entraînés à faire des copies sans faute (copies d'exercices, de résumés, de textes corrigés à mettre au net). Aujourd'hui, les élèves du collège font moins d'une dictée par semaine, recopient moins de textes et leurs cahiers ou leur classeurs ne sont pas systématiquement corrigés, contrairement aux cahiers du jour dans l'école primaire. En revanche, il existe d'autres entraînements à l'écriture dans les activités du collège (prises de notes, activités rédactionnelles dans toutes les matières).

La dictée la plus courte comporte 86 mots, la plus longue 140, la moyenne s'établissant à 111 mots. 3 dictées sur 5 de moins de 100 mots sont parmi les mieux réussies et 6 dictées sur 8 de plus de 115 mots ont obtenu une moyenne inférieure ou égale à 5/20. La dictée la mieux réussie est ainsi une poésie de François Coppée. On aurait pu penser que la forme versifiée serait un obstacle. En fait ce n'est pas le cas, mais la dictée ne comporte que 86 mots. De la même façon, une autre poésie (*L'alouette*) obtient la note moyenne de 7,8, ce qui la met au quatrième rang des dictées les mieux réussies : elle ne comporte que 90 mots.

Cependant la longueur ne suffit pas à hiérarchiser la difficulté des dictées. La dictée la moins bien réussie comporte 110 mots, elle se situe donc à la moyenne. Une autre dictée de moins de 100 mots (*Le facteur*) a une note moyenne de 3,8. En revanche, un texte de Victor Hugo (*La marche*) comporte 125 mots et la dictée portant sur ce texte est bien réussie (8,2 de moyenne, elle se classe au troisième rang). Il faut donc examiner d'autres facteurs.

➔ *Le lexique et le registre d'écriture*

Une autre difficulté est celle des mots qui composent le texte et dont l'orthographe peut ou non être connue. Cependant, des mots à orthographe inconnue peuvent être bien écrits s'ils sont réguliers (ce qui peut se déduire de leur audition en faisant fonctionner les règles de transcriptions habituelles du français) ; en revanche, des mots utilisés fréquemment (comme les mots « monsieur » et « femme ») peuvent poser des problèmes à cause de l'irrégularité de leur orthographe. Selon que le vocabulaire employé est rare ou familier, régulier ou irrégulier, les fautes seront plus ou moins probables. Pour pouvoir faire une analyse des dictées en fonction de ces paramètres, il faudrait disposer d'une échelle des fréquences des mots, non pas dans la langue mais en situation scolaire (rappelons, par exemple, que les mots « dictée », « orthographe » qui sont rarement utilisés dans la langue usuelle, sont d'un usage très fréquent à l'école). Faute de disposer encore d'un tel indicateur[7], nous pouvons constater « intuitivement » que le lexique des différents textes est plus ou moins « difficile », mais sans pouvoir justifier précisément cette impression. Le fait d'avoir eu recours à des dictées des années 1920 accroît cette impression : des termes qui devaient être familiers aux enfants d'une école rurale (« sillon », « moisson », « alouette », « grillon », « coquelicot », « arôme », « abattage », « ébrancheur », « fagot », « serpe ») sont sans doute

NOTE

7. Il est en cours d'élaboration à l'INRP (équipe de Bernard Léthé).

moins présents aujourd'hui dans les textes fréquentés à l'école, mais ils ne posent pas tous des problèmes d'écriture.

La question se pose différemment quand le sens des mots est inconnu, ou incertain, ou plus encore quand le registre textuel n'est plus du tout familier aux enfants. Même quand un texte est composé de mots simples, son sens peut être difficile à saisir. Ainsi, alors que la plupart des dictées sont la description d'une scène ou de brefs portraits, certaines se situent dans un autre style discursif : réflexion sur le pays natal (Bernardin de Saint-Pierre), sur la nécessité d'avoir une belle écriture, sur ce que symbolise le fait d'avoir un toit, sur la beauté de la langue française (Charles Bigot). Ces quatre textes se sont révélés tous les quatre particulièrement difficiles (moyennes respectives : 1,9/20, 3,6/20, 4/20, 5/20) et l'on peut penser que les élèves, saisissant mal de quoi il s'agissait, ont pu être troublés et de ce fait, multiplier les erreurs. Deux autres textes du même registre (le portrait d'un paresseux par Fénelon et un texte exposant quelle personnalité humaine symbolise chaque animal dans les fables), très faciles à comprendre, ont été bien mieux réussis (6,6 et 9,4).

➔ *La grammaire*

Les fautes de grammaire sont parmi les plus graves, celles qui enlèvent le plus de points. Elle pèsent donc particulièrement dans les résultats obtenus, par « construction » pourrait-on dire. Pour hiérarchiser les dictées en fonction de ces difficultés, nous avons cherché à dénombrer les erreurs potentielles qu'elles contenaient. Pour cela, nous avons dénombré les marques inaudibles à l'oral, au féminin (« claire » et « amusée » mais pas « douce » ni « laborieuse ») ou au pluriel pour les adjectifs et les noms (« petits », « pattes »). Pour les verbes, nous avons dénombré également les terminaisons non marquées à l'oral (« marchait », « marchais » ou « marchent », mais pas « marche » ou « marcha »). Si l'on ajoute les accords de participe passé et les confusions possibles entre participe passé et infinitif, on obtient un score qui montre l'existence d'une relation entre les deux : les cinq dictées les mieux réussies ont moins de 20 difficultés grammaticales (au sens où nous venons de le définir), sept sur onze des dictées les moins bien réussies ont de 27 à 33 difficultés. On dénombre dans la dictée courte *Le facteur* (98 mots) 31 difficultés, ce qui peut expliquer qu'elle ait été particulièrement mal réussie (moyenne : 3,8), alors que la dictée longue de Victor Hugo (125 mots, moyenne : 8,2) n'en comporte que 18.

➔ *Comparaison des résultats des meilleurs en fonction du sujet*

Si l'on veut comparer les résultats de l'année 1995 avec ceux des années 1920, il ne faut s'intéresser qu'aux résultats de la « meilleure moitié ». Comment réussissent ces élèves ? Leurs résultats sont évidemment bien meilleurs (près de 2 points de plus en moyenne, ce qui fait que leurs résultats vont de 4 à 13 et non de 2 à 11, selon le sujet des dictées). Cependant, on n'observe pas seulement un accroissement des notes, mais aussi une distorsion des résultats. 7 dictées sur 26 obtiennent des résultats significativement meilleurs que les résultats attendus (soit davantage que les deux points de plus). Cela signifie que pour la meilleure moitié des élèves, l'ordre de difficultés des dictées n'est pas le même que pour l'ensemble de la population. Par exemple, la dictée *Le facteur* (moyenne : 3,8 pour l'ensemble) a une moyenne de près de 10 dans la meilleure moitié. Tout se passe comme s'il existait des seuils de compétence, qui font que certaines difficultés, rédhibitoires pour les mauvais élèves, ne posent pas problème aux meilleurs. S'agit-il de fautes de lexique ou de grammaire ? Pour le moment les éléments dont nous disposons ne permettent pas de trancher.

Les causes des erreurs proviennent de multiples facteurs que seule une étude exhaustive pourrait élucider complètement. Pour l'heure, faute de pouvoir hiérarchiser complètement les dictées, on peut seulement constater que chaque facteur contribue pour une part aux résultats obtenus, sans pouvoir dire de quel poids pèse chacun d'eux. La longueur des dictées et la fréquence des difficultés grammaticales constituent deux premiers indicateurs efficaces mais ceux-ci devraient être affinés.

Influence de la classe et de l'âge sur les résultats en dictée

Le niveau de classe, et l'âge peuvent influer sur la réussite, en jouant soit sur l'apprentissage (par capitalisation des acquis anciens ajoutés aux nouvelles acquisitions), soit sur la vigilance (une plus grande maturité peut accroître le temps pendant lequel l'attention reste mobilisée sans fatigue et faciliter l'autocontrôle au moment où l'on écrit et/ou l'on se relit).

La moyenne des notes obtenues en contrôle continu varie peu, que l'on soit en sixième, cinquième ou quatrième, par construction, pourrait-on dire. En effet, chaque classe établit sa moyenne autour de 10/20, quels que soient les exercices demandés. En revanche, les résultats des élèves à la dictée du CEP s'améliorent nettement quand on passe d'une classe à l'autre. La moyenne des élèves est de 4,9 en

sixième, de 6,6 en cinquième et de 8,8 en quatrième. Si l'on s'en tient aux résultats des élèves « à l'heure », la moyenne passe de 6,2 en sixième à 8,2 en cinquième et à 9,9 en quatrième. En classe de quatrième, la moyenne en dictée, pour les élèves « à l'heure », rejoint pratiquement la moyenne en contrôle continu. Pourtant, on ne fait pas plus de dictées en quatrième qu'en sixième. Est-ce que cela signifie que la « maturité » orthographique des élèves est plus tardive aujourd'hui qu'hier ? De fait, si les savoirs orthographiques s'améliorent sans faire l'objet d'un entraînement accéléré, c'est soit par effet direct (le travail fait en orthographe se capitalise et porte des fruits), soit par effet indirect (des activités non spécifiquement centrées sur l'orthographe, comme la lecture ou la production d'écrit dans toutes les matières, ont des effets sur les compétences en orthographe). On peut également penser que la longueur des dictées de CEP déborde les capacités des enfants en sixième mais que ce n'est plus le cas en quatrième : la difficulté pour les élèves de sixième serait celle du contrôle de l'attention dans une tâche trop longue. Là encore, seule une analyse plus poussée pourrait trancher entre ces diverses hypothèses. Quelques indices nous permettront de proposer plus loin une piste d'interprétation.

Un deuxième résultat intéressant concerne les résultats des élèves qui, dans chaque classe, sont en avance ou en retard. Les élèves en avance, dans toutes les classes, ont la moyenne en dictée (avec 10,1 en sixième, 11,6 en cinquième et 10,8 en quatrième), ce qui signifie que la compétence orthographique fait bien partie des capacités qui permettent de repérer un très bon élève, celui dont la maturité intellectuelle est jugée suffisante pour qu'il soit en avance d'une classe. Inversement, l'échec orthographique est une caractéristique des élèves ayant redoublé une ou deux classes. Contrairement à ce que l'on pourrait croire, le résultat à une dictée est donc un indicateur (statistique) fiable de la performance scolaire globale.

Par ailleurs, les performances des élèves en échec, loin de rester stationnaires, s'améliorent au fil des années : parmi les élèves ayant redoublé, ceux de sixième ont une moyenne de 1,9 sur 20, en cinquième de 3,3 et en quatrième de 5,9. Leurs mauvais résultats relatifs ne doivent pas fait perdre de vue que l'écart avec les meilleurs élèves ne s'élargit pas mais au contraire, se resserre.

Ces résultats sont confirmés quand on compare les réussites de la totalité de la population avec les résultats de la meilleure moitié. Les notes moyennes de ces « bons » élèves s'améliorent de 1,5 points ; elles sont de 8,7 en sixième, 8,9 en cinquième, et 10,2 en quatrième. Les progrès faits d'année en année sont donc beaucoup moins sensibles pour les meilleurs que les progrès fait par l'ensemble du groupe (respectivement 4,9, 6,6 et 8,8 pour un gain de 3,9 points). Entre la sixième et la cinquième, ces progrès sont même non significatifs. Ceci confirme que les élèves les moins bons sont ceux qui progressent le plus en trois ans et sont donc ceux qui tirent le plus profit des apprentissages (sans qu'on puisse dire ce qui produit ce progrès). En revanche, pour les bons élèves (la meilleure moitié ou les élèves en avance), les progrès entre sixième et quatrième semblent plus lents. On doit prendre ces constats avec précaution, car l'échantillon d'élèves de quatrième est très restreint. Il serait intéressant que d'autres recherches puissent confirmer ou infirmer ces données, en travaillant sur la totalité du collège, pour savoir s'il y a réellement un niveau de saturation des résultats autour de la quatrième. Tout se passe en effet comme si les bons élèves arrêtaient plus vite de progresser que les autres.

Comparaisons avec les performances des élèves des années 1920

Toute la question est de savoir si les performances moyennes désastreuses à cet exercice des années 1920 sont dues à un effondrement des compétences orthographiques « en général » ou à l'échec devant cet exercice spécifique. Pour le savoir, il faudrait pouvoir comparer les performances en orthographe des élèves d'hier et d'aujourd'hui, mais en dehors de la dictée, dans une activité d'écriture libre. Deux biais nous sont offerts pour avancer dans cette voie.

La première possibilité est d'évaluer leurs performances dans une activité de copie libre, non évaluée. Une telle activité permet en effet de séparer les facteurs « connaissance orthographique » et « vigilance orthographique », puisqu'elle ne met en jeu que le second. En effet, l'élève qui copie un texte sans faute est soit celui qui se dicte à lui-même avec sûreté ce qu'il a lu au tableau parce que ses connaissances orthographiques sont fiables, soit celui qui a des doutes au fur et à mesure qu'il écrit et s'autocorrige alors par comparaison avec le modèle. En revanche, celui qui parsème sa copie de fautes ne se pose pas les bonnes questions au moment où il écrit et ne fait pas des vérifications efficaces. C'est donc bien sa vigilance orthographique qui est en défaut. On peut s'attendre à ce que les élèves des années 1920 dépassent largement ceux d'aujourd'hui sur ce terrain. Or, nous avons un texte de ce type, avec le sujet de rédaction qui, en 1920 et 1995, était inscrit au tableau et recopié sur la copie. Comment se comportent les élèves d'hier et d'aujourd'hui ? (tableau 2)

De fait, les résultats ne sont pas défavorables aux élèves d'aujourd'hui. 63,8 % des sujets ont copié le sujet sans erreur en 1920 et 70,8 % en 1995. L'écart

TABLEAU 2 – Copie du sujet dans les années 20 et en 1995
(Sous-échantillon des meilleurs : variations selon les centres de passage du CEP)

Numéro de centre	Numéro de sujet	Sujet copié sans erreur (parmi les élèves ayant copié le sujet)		Sujet copié avec erreur(s) (parmi les élèves ayant copié le sujet)		Proportion d'élèves n'ayant pas copié le sujet	
		Années 20	1995	Années 20	1995	Années 20	1995
1	25	60,1 %	62,8 %	39,9 %	36,2 %	0,0 %	12,8 %
2	12	81,2 %	76,1 %	18,8 %	23,9 %	0,0 %	0,0 %
3	7	45,7 %	77,4 %	54,3 %	22,6 %	0,0 %	10,6 %
4	15	91,0%	90,6 %	9,0 %	9,4 %	4,3 %	0,0 %
6	8	59,7%	82,1 %	40,3 %	17,9 %	0,0 %	0,0 %
9	11	51,1%	90,0 %	48,9 %	10,0 %	0,0 %	33,3 %
10	26	94,4%	86,3 %	5,6 %	13,7 %	0,0 %	0,0 %
11	16	61,9%	89,8 %	38,1 %	10,2 %	0,0 %	0,0 %
13	19	77,4%	72,9 %	22,6 %	27,1 %	0,0 %	0,0 %
15	6	60,0%	72,7 %	40,0 %	27,3 %	0,0 %	0,0 %
17	14	5,7%	80,8 %	94,3 %	19,2 %	0,0 %	0,0 %
18	23	58,5%	76,2 %	41,5 %	23,8 %	0,0 %	0,0 %
20	10	36,6%	71,0 %	63,4 %	29,0 %	0,0 %	31,1 %
21	18	43,1%	86,6 %	56,9 %	13,4 %	0,0 %	2,9 %
22	4	84,4%	92,3 %	15,6 %	7,7 %	0,0 %	0,0 %
24	9	77,5%	76,5 %	22,5 %	23,5 %	0,0 %	0,0 %
25	21	63,3%	85,3 %	36,7 %	14,7 %	0,0 %	2,9 %
27	17	82,9%	79,3 %	17,1 %	20,7 %	0,0 %	9,4 %
30	3	92,5%	90,9 %	7,5 %	9,1 %	7,0 %	0,0 %
31	13	79,8%	60,9 %	20,2 %	39,1 %	0,0 %	25,6 %
35	2	68,6%	84,4 %	31,4 %	15,6 %	0,0 %	0,0 %
36	1	75,0%	72,2 %	25,0 %	27,7 %	0,0 %	28,0 %
37	22	67,2%	67,7 %	32,7 %	32,3 %	7,9 %	11,9 %
38	24	72,3%	41,2 %	27,7 %	58,8 %	0,0 %	5,6 %
39	20	47,2%	64,0 %	52,8 %	36,0 %	0,0 %	7,4 %
40	5	28,6%	71,9 %	71,4 %	28,1 %	0,0 %	0,0 %

en faveur de 1995 (qui concerne la totalité de la population et non la meilleure moitié) peut s'expliquer par le fait que le sujet a toujours été inscrit au tableau en 1995, alors que ce n'est manifestement pas le cas pour quelques sujets dans les années 20. Par exemple, 5,7 % des élèves ont écrit le sujet 14 sans erreur en 1920, 80,8 % en 1995. Il s'agit d'un sujet d'une ligne *(Racontez l'une des promenades scolaires que vous avez faites. Dites ce que vous aviez observé)* qui a très probablement été dicté lors du passage du certificat d'études et non recopié. Si l'on supprime ces cas douteux, on obtient une différence de performance qui est tout de même en faveur des élèves d'aujourd'hui.

Aujourd'hui comme hier, environ 7 élèves sur 10 sont donc capables (et 3 sur 10 incapables) de vérifier l'exactitude de leur copie devant le modèle. On aurait pu s'attendre à ce que les élèves des années 1920, candidats sélectionnés au CEP, soient quasiment tous des copieurs modèles, même dans cette situation non évaluée, mais qui avait lieu un jour d'examen. Ce n'est manifestement pas le cas, ce qui veut dire que leur vigilance est dans cette situation bien moins forte qu'en dictée. C'est donc à force d'attention, et non parce qu'ils auraient fini par acquérir une orthographe « naturelle » qu'ils réussissent en dictée. C'est la pression exercée par la note éliminatoire et non le savoir orthographique en lui-même, qui produit les bons résultats observés. Dès qu'ils se retrouvent dans une situation non évaluée, l'attention d'un certain nombre d'entre eux se relâche et leurs résultats chutent. En revanche, si les résultats des élèves d'aujourd'hui sont très supérieurs à leurs résultats en dictée, c'est qu'ils sont capables de s'autocorriger efficacement dans cette situation : seule une minorité (30 %) est incapable d'avoir les doutes qui conduisent à vérifier et s'autocorriger. Ceci signifie que les élèves d'aujourd'hui qui n'ont pas des connaissances sûres en orthographe font néanmoins preuve d'une assez bonne vigilance orthographique dans cette situation.(tableau 3)

La seconde voie est obtenue en comptabilisant le nombre moyen de fautes dans une situation d'écriture libre, dans laquelle l'attention de l'élève est mobilisée sur une autre tâche (le contenu de ce qu'il produit, ou, quand il recopie son brouillon au net, de ce qu'il a produit). L'orthographe n'est pas évaluée formellement en rédaction, mais il existe une pression diffuse de l'institution, hier comme aujourd'hui,

TABLEAU 3 – Nombre de fautes par dix lignes en rédaction (comptabilisées par l'évaluateur)

Proportion d'élèves	Nombre de fautes par dix lignes	
	Années 20	Meilleure moitié –1995
25%	de 0 à 0,8	de 0 à 2,1
25%	de 0,9 à 1,8	de 2,2 à 3,8
25%	de 1,9 à 3,4	de 3,9 à 6,3
24%	de 3,5 à 9,9	de 6,4 à 17,2
1%	de 10 à 50	de 17,3 à 30,6

LECTURE – 25 % des élèves des années 20 ont fait de 0 à 0,8 faute par dix lignes en rédaction et 25 % de la meilleure moitié des élèves de 1995 en ont fait de 0 à 2,1.

pour que les élèves fassent attention à ne pas laisser dans leur devoir des fautes qui impressionnent défavorablement le correcteur.

Les élèves d'aujourd'hui, qui font cinq fois plus de fautes en dictée que les élèves des années 20, font « seulement » deux fois plus de fautes en rédaction. On peut interpréter ce fait soit pour y conforter l'écart déjà constaté en dictée, soit pour trouver que les « bons élèves en dictée » des années 20 ont une orthographe moins assurée qu'on aurait pu le penser, puisque dans une activité de moindre contrôle, leur avance diminue sérieusement. Pour les élèves de l'année 1995, les performances en rédaction ne sont pas très au-dessous des performances en dictée, alors que les élèves des années 20 ont des capacités de vigilance et d'autocontrôle en dictée qui ne se transfèrent pas dans la production d'écrit. Ceci est d'autant plus marqué que les élèves de 1995 utilisent un lexique plus varié et plus riche que les élèves des années 20, ce qui accroît les risques d'erreurs orthographiques.

Les savoir-faire orthographiques des élèves des années 1920 et d'aujourd'hui, tels qu'ils sont évalués dans l'épreuve de dictée, relèvent à la fois des connaissances et de la vigilance attentionnelle. La supériorité écrasante des élèves d'hier sur les élèves d'aujourd'hui tient pour une part à leur meilleure connaissance de l'orthographe française mais plus encore à leur entraînement à s'autocorriger dans une épreuve de dictée. Dès que cette vigilance n'est plus aussi sollicitée (copie non notée ou écriture libre), leur résultats se dégradent davantage que ceux des élèves d'aujourd'hui et l'écart entre les deux populations (qui demeure net) est beaucoup moins spectaculaire.

Les résultats des années 20 : une supériorité à relativiser

On peut tirer cinq conclusions de cette comparaison entre résultats en dictée et résultats au contrôle continu d'orthographe :

– les résultats des élèves dans l'épreuve de dictée sont globalement mauvais par rapport aux critères des années 1920. Les élèves d'aujourd'hui y font cinq fois plus de fautes que ceux d'hier ;

– il existe, pour les élèves d'aujourd'hui, une hiérarchie des résultats qui tient à la fois à la longueur du texte, aux difficultés grammaticales et à d'autres difficultés (lexicales, registre textuel) sans qu'on puisse déterminer le poids de ces différents facteurs dans l'analyse. Cette hiérarchie de difficultés n'est pas la même pour les meilleurs et les moins bons élèves ;

– l'échec et la réussite en dictée sont corrélés avec l'échec et la réussite globale de l'élève ;

– les élèves qui font des progrès significatifs entre sixième et quatrième sont les élèves moins bons. On peut penser que des progrès proviennent moins de l'acquisition de nouveaux savoirs que de meilleures capacités d'autocorrection, c'est-à-dire des progrès dans l'intériorisation d'une « conscience orthographique » ;

– la supériorité des élèves des années 20 sur les élèves d'aujourd'hui provient d'une bien meilleure capacité d'autocontrôle en dictée, capacité qui ne se maintient pas dans les épreuves d'écriture non évaluées (que ce soit en copie ou en rédaction). Tout se passe comme si leur capacité d'autocontrôle était liée à l'exercice pour lequel ils recevaient un véritable entraînement.

Les dictées des années 1920 témoignent d'une compétence élevée en orthographe, mais on ne peut tirer de ces résultats l'indice d'une compétence orthographique « générale », valant pour toutes les situations d'écriture. Cela semble être bien moins être le cas pour les élèves d'aujourd'hui : du fait que l'entraînement à la dictée n'est plus aussi intensif, l'écart entre les résultats en dictée et dans les autres activités d'écriture est bien moindre. Pourtant, malgré cette absence d'entraînement intensif, les résultats s'améliorent au fil du temps, surtout pour les moins bons élèves. C'est donc qu'il existe d'autres voies d'apprentissage, qui vont du renforcement des savoirs (systématique dans les leçons d'orthographe, ou occasionnel dans les activités d'écriture corrigées ou autocorrigées), à l'acquisition d'une meilleure « conscience orthographique » (qui permet de veiller efficacement aux accords et de « savoir qu'on ne sait pas » quand on a à écrire un mot non familier). En tout état de cause, une dictée de certificat d'études est aujourd'hui du niveau de la classe de quatrième pour

les élèves à l'heure (la note moyenne en dictée est la même qu'au contrôle continu) et de la classe de sixième pour les élèves en avance. Pour améliorer les performances des élèves d'aujourd'hui, il faudrait jouer sur les savoirs à acquérir (pour le lexique) ou à consolider (les règles de grammaire sont supposées connues). Il faudrait jouer tout autant sur les capacités d'autocontrôle. Le poids de ce facteur dans la réussite en dictée permettrait d'expliquer pourquoi la compétence orthographique est corrélée à la réussite globale. En effet, la vigilance attentionnelle conditionne les capacités d'auto-évaluation, exige concentration, retour sur la tâche, distanciation par rapport à elle, dissociation entre contenu et forme, etc. Or, c'est une compétence transversale, dont on peut supposer qu'elle s'investit dans d'autres activités scolaires et qu'elle constitue un atout important pour rendre tous les apprentissages efficaces, dans toutes les matières, une fois les savoirs compris et acquis. Alors que les savoirs orthographiques n'ont pas en eux-mêmes de corrélation avec le niveau général d'un élève, la compétence orthographique, par ce biais, peut devenir un indicateur de la réussite globale.

LA RÉDACTION ET L'EXPRESSION ÉCRITE

Entre les notes au CEP et au contrôle continu en 1995, un moindre écart que pour la dictée

La note moyenne des élèves au contrôle continu est de 10,6/20. Cette note est la moyenne des résultats obtenus aux divers devoirs d'expression écrite, elle provient donc d'un ensemble beaucoup moins composite que la note moyenne en orthographe. L'écart avec la note moyenne à la rédaction de CEP (9,4/20) est net mais moins fort qu'en orthographe (1,2 point contre 4,3 points). Les distributions des résultats aux deux épreuves ont une forme gaussienne très comparable, au décalage vers la gauche près (graphique 4). L'échelle des notes utilisée (de 0 à 19) est la même dans les deux cas. On peut donc dire que l'épreuve de CEP est un peu au-dessus du niveau des expressions écrites habituellement exigées des élèves, mais qu'elle ne les met pas en grave difficulté.

Une assez faible corrélation entre les résultats au contrôle continu et au CEP

Si les distributions aux deux épreuves sont proches, cela ne signifie pas pour autant que la corrélation entre elles soit forte. En effet, la performance d'un élève peut varier d'un sujet à l'autre. On sait aussi que la notation en expression écrite (ou en rédaction) n'est pas déductible d'un barème aussi rigoureux que pour la dictée ou l'épreuve de grammaire. Un même élève, qui aurait fait deux devoirs « de même valeur », peut avoir deux notes relativement différentes aux deux épreuves, même s'il est noté par le même évaluateur. De plus, le même évaluateur, notant le même devoir deux fois de suite, peut être amené à donner deux notes différentes, selon la copie corrigée auparavant, son degré de fatigue, etc. Ces phénomènes docimologiques sont aujourd'hui bien connus. C'est ce qui se vérifie ici : le coefficient de corrélation est de .06, ce qui indique qu'une corrélation existe mais qu'elle est relativement faible.

GRAPHIQUE 4 – Distribution des notes au CEP et au contrôle continu en 1995

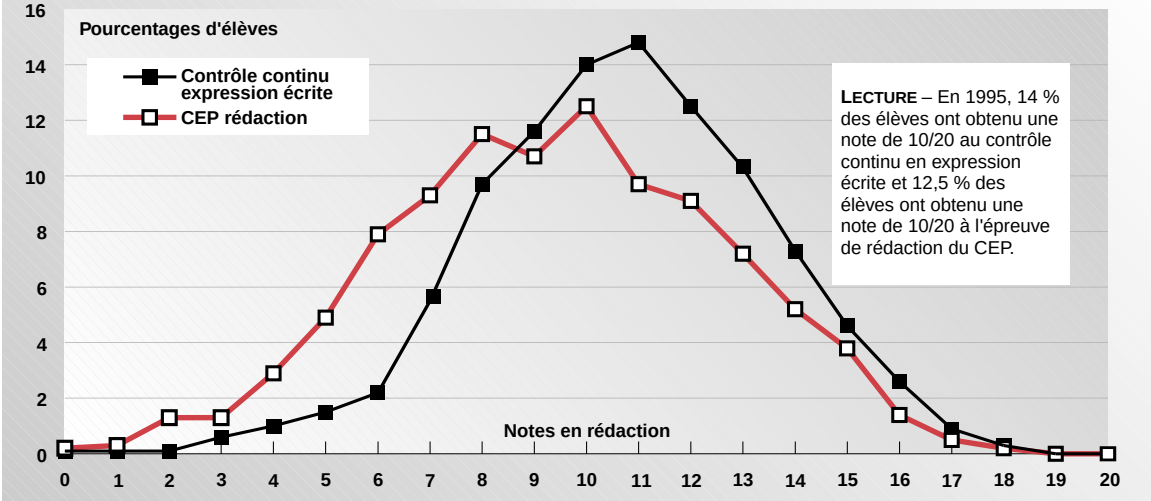


TABLEAU 4 – Liaison entre les résultats d'expression écrite au contrôle continu et les résultats de rédaction aux épreuves du CEP (Élèves de l'échantillon France entière en 1995)

Note contrôle continu	Note CEP								
	0 à 2,4	2,5 à 4,9	5 à 7,4	7,5 à 9,9	10 à 12,4	12,5 à 14,9	15 à 17,4	17,5 à 20	Ensemble
0 à 2,4	20,0 %	40,0 %	30,0 %	10 %	0,0 %	0,0 %	0,0 %	0,0 %	100,0 %
2,5 à 4,9	12,1 %	22,4 %	48,3 %	15,5 %	1,7 %	0,0 %	0,0 %	0,0 %	100,0 %
5 à 7,4	6,3 %	16,0 %	48,1 %	16,8 %	10,2 %	2,3 %	0,4 %	0,0 %	100,0 %
7,5 à 9,9	1,4 %	4,2 %	32,7 %	36,9 %	21,3 %	3,2 %	0,3 %	0,0 %	100,0 %
10 à 12,4	1,3 %	2,4 %	17,0 %	22,9%	41,2 %	12,5 %	2,8 %	0,0 %	100,0 %
12,5 à 14,9	0,5 %	1,0 %	7,9 %	15,1 %	38,1 %	26,4 %	10,2 %	0,7 %	100,0 %
15 à 17,4	0,0 %	0,6 %	1,8 %	7,6 %	18,2 %	34,7 %	35,3 %	1,8 %	100,0 %
17,5 à 20	0,0 %	0,0 %	0,0 %	10 %	20,0 %	30,0 %	30,0 %	10,0 %	100,0 %

LECTURE – 22,4 % des élèves qui avaient une note de contrôle continu comprise entre 2,5 et 4,9 sur 20 ont eu au CEP une note comprise entre 2,5 et 4,9 sur 20.

L'examen du tableau de contingence (tableau 4) permet de voir que les deux notes augmentent en même temps (pourcentages élevés concentrés dans la diagonale du tableau) mais que la dispersion autour de la note moyenne reste forte.

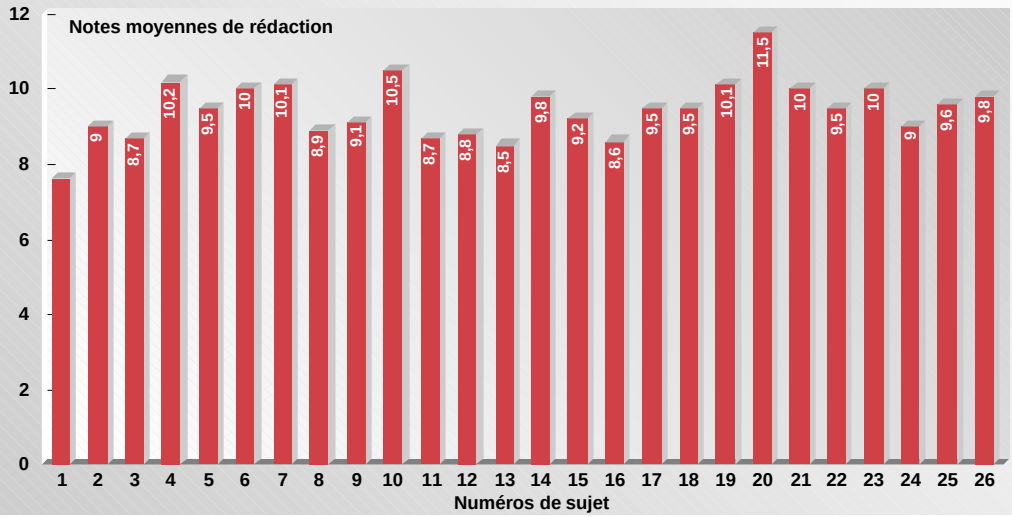
Le sujet proposé influe sur les résultats

Selon le sujet de rédaction, la note moyenne varie de 7,6 (sujet n° **1** : *Les grandes vacances. Leur utilité. À quoi allez-vous les employer ?*) à 11,5 (sujet n° **20** : *Par un mercredi pluvieux, vous observez à travers les vitres ce qui se passe dans la rue. Racontez ce que vous avez vu, dites quelles impressions vous a laissées le spectacle et les réflexions qu'il vous a inspirées*). À quoi attribuer ces différences ? (graphique 5)

➔ *Une réussite variant selon le type de sujet*

On peut penser que la difficulté des rédactions varie selon le type d'écrit demandé : récit, description, réflexion, etc. Si l'on regarde les huit meilleures rédactions (celles qui ont plus de 10 de moyenne), on voit que tous les sujets sont des récits explicitement signalés comme tels (*Racontez*). Ils portent sur des événements ou situations vécus ou imaginés, mais le scénario est souvent tracé dans le libellé (comme c'est le cas dans la rédaction sur le mercredi, la mieux réussie). En revanche, les sept moins bonnes rédactions (moyenne inférieure à 9/20) sont de genres variés. On trouve aussi des récits mais assortis de considérations générales (*En approchant d'un buisson, Jean voit un oiseau qui s'envole, il découvre un nid et… Racontez le reste et dites vos réflexions*) ou ne se présentant pas comme tel (cas des grandes vacances qu'il faut anticiper et dont il

GRAPHIQUE 5 – Notes moyennes de dictée au CEP selon le sujet



THÈME

faut dire l'utilité « en général »). On trouve également les descriptions assorties de considérations (*Décrivez votre habitation. Quels sentiments vous inspire-t-elle ?*) ou une lettre de remerciements. Les genres les moins familiers (la lettre) ou demandant de mêler deux registres d'écriture seraient moins bien réussis.

Pourtant, ce n'est pas toujours le cas : le devoir qui s'est avéré le plus facile (sur le mercredi) demande de mêler récit et impressions et d'autres récits sont très mal réussis (un acte de courage). D'autres facteurs doivent donc contribuer à cette hiérarchisation des difficultés.

➔ *Les sujets à scénario donnent de meilleures rédactions*

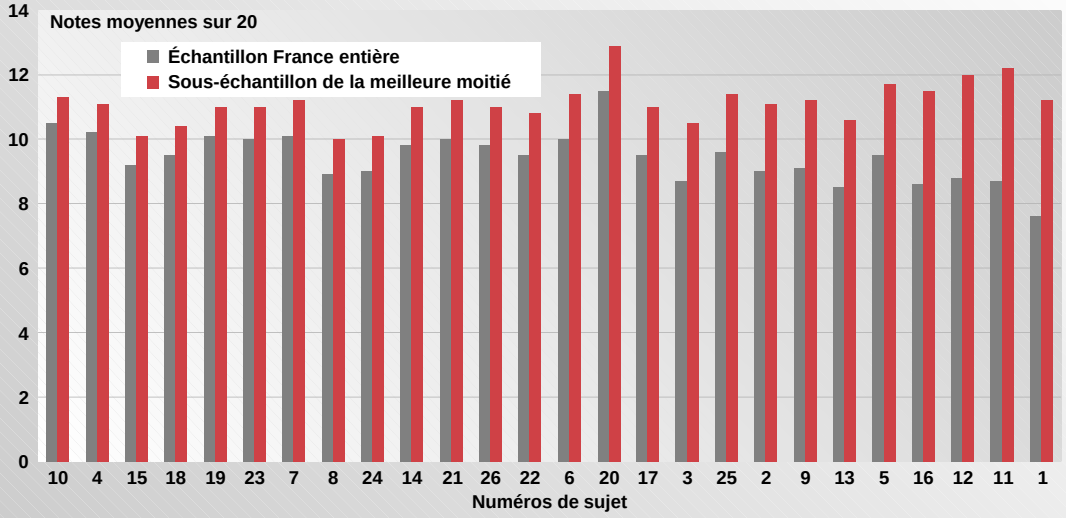
Voici les libellés des sept sujets les moins bien réussis (moins de 9/20 en moyenne) :
– *Les grandes vacances. Leur utilité. À quoi allez-vous les employer ?* (Sujet n° **1**, moyenne : 7,6)
– *Un acte de courage. Vous avez certainement et plus d'une fois fait preuve de courage. Choisissez un acte de courage parmi ceux que vous avez accomplis. Racontez-le simplement et avec précision.* (Sujet n° **13**, moyenne : 8,5)
– *Vous avez comme voisins de classe un bon et un mauvais camarades. Montrez comment ils se conduisent en classe, en récréation, dans la rue. (Sujet n° **16**, moyenne : 8,6)*
– *En approchant d'un buisson, Jean voit un oiseau qui s'envole, il découvre un nid et… Racontez le reste et dites vos réflexions.* (Sujet n° **3**, moyenne : 8,7)
– *Racontez comment vous avez passé vos vacances de Pâques (jeux, voyages, devoirs, aides à vos parents, etc.) Avez-vous été heureux(se) ou ennuyé(e) de rentrer en classe ?* (Sujet n° **11**, moyenne : 8,7)
– *Décrivez votre habitation. Quels sentiments vous inspire-t-elle ?* (Sujet n° **12**, moyenne : 8,7)
– *Un parent (oncle, tante ou parrain) vous a fait un cadeau à l'occasion de votre anniversaire ou pour vos étrennes. Vous écrivez pour le remercier.* (Sujet n° **8**, moyenne : 8,9)

Voici le libellé des cinq sujets les mieux réussis (plus de 10 en moyenne) :
– *Par un mercredi pluvieux, vous observez à travers les vitres ce qui se passe dans la rue. Racontez ce que vous avez vu, dites quelles impressions vous a laissées le spectacle et les réflexions qu'il vous a inspirées.* (Sujet n° **20**, moyenne : 11,5)
– *Vous avez appris un certain nombre de fables. Quelle est celle qui vous a le plus intéressé(e). Racontez-la à votre manière.* (Sujet n° **10**, moyenne : 10,5)
– *Racontez un événement gai ou triste dont vous avez été le témoin ou dont on vous a fait le récit.* (Sujet n° **4**, moyenne : 10,2)
– *Vous raconterez d'une manière bien personnelle un événement fâcheux causé involontairement, ou bien par votre faute, maladresse, distraction, oubli…* (Sujet n° **7**, moyenne : 10,1)
– *Un homme mort il y a cent ans revient dans les lieux où il a vécu. Dépeignez son étonnement et indiquez les choses nouvelles qui vont le frapper particulièrement.* (Sujet n° **19**, moyenne : 10,1)

Il semble bien que la piste du scénario à construire soit plus pertinente que celle du type de texte à produire. En effet, tous les sujets bien traités permettent de reconstruire ou inventer aisément un scénario, qu'il s'agisse d'un récit dont le cadre est préconstruit (sujets **20** et **19**), centré sur un événement dont le sens est donné (sujets **4** et **7**) ou d'un récit dont le scénario n'est pas à construire mais à récupérer en mémoire (sujet **10** sur la fable). Parmi les moins bien traités, on trouve quatre sujets très typiques des années 1920 (sur l'acte de courage, la maraude des oisillons seuls dans le nid, les portraits du bon et du mauvais camarades et la lettre). Ils ne font plus partie des lectures ordinaires de l'école (ni, pour certains, de l'expérience des enfants). En effet, les lectures faites en classe constituaient une source privilégiée pour savoir ce qui était attendu dans ce genre de devoir. Les deux sujets sur les vacances demandent de construire un écrit en sélectionnant des informations pertinentes dans une masse de données d'expérience, de façon à construire un texte cohérent et original, sans qu'on puisse s'appuyer sur le libellé du sujet pour y parvenir. Rappelons que dans les années 20, certains instituteurs faisaient apprendre par cœur à leurs élèves des canevas-types sur les sujets dont ils savaient qu'ils étaient régulièrement donnés à l'examen.

Si l'on prend la meilleure moitié des élèves(graphique 6), les écarts de note moyenne entre les différents sujets de rédaction ne sont pas les mêmes. L'écart maximum est de 2,9 points (de 10 à 12,9) alors qu'il était de 3,9 points (de 7,6 à 11,5) sur la totalité des copies . Le sujet le moins bien réussi est la lettre de remerciement, le mieux réussi reste le même que pour la population entière (récit d'un mercredi). Tout se passe donc, comme en dictée, comme s'il y avait des seuils de difficultés. Ce qui fait statistiquement problème pour l'ensemble de la population ne le fait pas toujours pour la meilleure moitié. Les sujets pour lesquels les écarts sont les plus forts sont les sujets **1**, **11**, **12** et **16**. On peut expliquer cette différence par un « effet-classe ». En effet, ces élèves

GRAPHIQUE 6 – Notes moyennes à l'épreuve de la meilleure moitié des élèves et de l'ensemble des élèves de l'échantillon 1995 selon le sujet

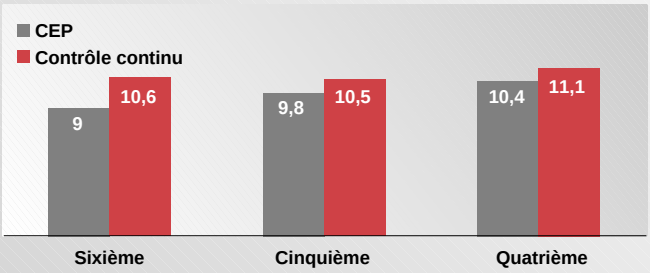


appartiennent aux classes pour lesquelles les moyennes au contrôle continu en mathématiques, orthographe, rédaction, sont les meilleures.

Influence de la classe et de l'âge sur les résultats en rédaction

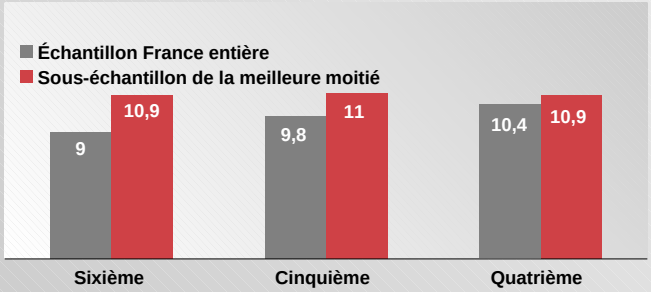
Comme pour les autres notes de contrôle continu, la note moyenne varie peu selon la classe (elle est toujours légèrement supérieure à 10. En quatrième, elle atteint 11,1 mais sur un effectif réduit d'élèves). En revanche, la note de rédaction augmente de classe en classe, même si c'est de façon moindre qu'en dictée, puisque l'écart de départ est moins important (graphique 7).

D'autre part, les différences entre population totale et meilleure moitié ne sont pas significatives classe par classe (graphique 8). Les notes moyennes des élèves sont de 9 en sixième, 9,8 en cinquième et 10,4 en quatrième. Les notes moyennes des élèves « à l'heure » sont de 9,6, 10,7 et 10,9. Celles des élèves en avance, les meilleures du lot, progressent de 10,8 à 11,3 puis 11,9. Ainsi, les élèves en avance ont une moyenne d'un point supérieur à celle des élèves « à l'heure ». Enfin, les élèves en retard progressent de 7,7 à 8,1 et 9,2. Là encore, ce sont les élèves les plus faibles qui progressent le plus entre cinquième et



GRAPHIQUE 7 – Notes moyennes de rédaction au contrôle continu et au CEP selon la classe

LECTURE – En 1995, les élèves de sixième ont obtenu une moyenne de 10,6/20 au contrôle continu en expression écrite, et de 9/20 à l'épreuve de rédaction du CEP.



GRAPHIQUE 8 – Notes moyennes de rédaction à l'épreuve du CEP de la meilleure moitié des élèves de 1995 et de l'ensemble des élèves de l'échantillon de 1995 selon la classe

LECTURE – Les élèves de sixième de l'échantillon France entière ont obtenu une moyenne de 9/20 à l'épreuve de rédaction du CEP.
Les élèves de sixième du sous-échantillon de la meilleure moitié des élèves ont obtenu à la même épreuve une moyenne de 10,9/20.

quatrième (de 1,1 points de moyenne) au moment où les autres ne progressent pratiquement pas (0,2 pour les élèves « à l'heure » et 0,6 pour les élèves en avance). En l'absence d'analyses supplémentaires plus précises, on ne peut guère expliquer d'où provient ce phénomène de saturation des progrès déjà observé en dictée. Les critères de notation sur la rédaction ne permettent pas aux élèves des plus grandes classes de manifester leur supériorité sur ces libellés de sujets aussi nettement qu'en orthographe où les écarts de départ sont très forts.

Comparaison des résultats des élèves des années 20 et des meilleurs élèves de 1995

Lorsque l'on compare la distribution des notes données par le correcteur de 1920 aux notes données à une population comparable de 1995 (la meilleure moitié des élèves), on obtient deux courbes exactement superposables (graphique 9). La note modale est la même (5,5). Cela signifie que les performances des élèves d'aujourd'hui, jugés par un évaluateur d'aujourd'hui, sont tout à fait semblables à celles des élèves des années 20 jugés par les jurys d'instituteurs, alors que les sujets proposés, qui étaient familiers à ceux qui préparaient le certificat d'études, le sont sans doute moins aujourd'hui. Les standards en la matière ont donc dû beaucoup moins évoluer qu'en orthographe.

On dispose pour chaque élève d'aujourd'hui de deux notes pour le même devoir, celle donnée par le professeur de l'élève et celle donnée par l'évaluateur (graphique 10). Là encore, les distributions des notes se recouvrent, même si les notes données par le professeur sont légèrement plus resserrées autour de la moyenne (moins de notes très basses ou très hautes).

On peut donc dire que les critères de notation, tels que la distribution des notes les traduit, sont restés parfaitement stables dans le corps enseignant entre des copies d'hier et d'aujourd'hui et entre deux correcteurs devant un même lot de copies actuelles. Ceci peut paraître contradictoire avec le fait souligné plus haut (absence de fidélité du jugement; subjectivité des critères, variations dans la notation d'une même copie, etc.) qui expliquait la faible corrélation entre contrôle continu et rédaction.

En fait, il ne faut pas confondre deux phénomènes : la façon dont un correcteur applique à une copie un jugement qui synthétise un ensemble complexe de critères (ce qui fait que la même copie peut être diversement appréciée) et la façon dont un correcteur distribue ses notations et donc ses jugements sur un ensemble de productions d'élèves. C'est du deuxième phénomène qu'il s'agit ici. Ce qui ressort de la superposition parfaite des différentes courbes montre que les critères de notation sont restés stables. La rédaction est en effet un exercice complexe, dans lequel la note synthétise un ensemble de jugements de valeur, portant à la fois sur le contenu et la forme, le sujet traité (choix des éléments racontés ou décrits, vécus ou imaginés) et la langue employée pour le dire (longueur et cohérence du texte, correction syntaxique, justesse du vocabulaire, etc.). Si au fil du temps, le poids respectif des différents éléments en cause avait changé, on aboutirait à un autre profil de notation et à une autre distribution de l'ensemble des notes. Par exemple, dans les copies d'hier jugées par des instituteurs de 1920, les qualités retenues (contenu, forme, exigences) ne pèseraient pas du même poids que pour les évaluateurs d'aujourd'hui (des professeurs des écoles stagiaires), et les copies corrigées aujourd'hui par des évaluateurs du primaire débutants ne seraient pas jugées de la même façon par des professeurs du secondaire connaissant leurs élèves. Sur tous ces points, le jugement des évaluateurs paraît (statistiquement) très stable (ce qui ne signifie pas, évidemment, qu'une copie obtient la même note aux deux évaluations). On peut donc conclure que l'exercice de la rédaction, même transformée en expression écrite, fait toujours partie de la réalité scolaire contemporaine et que les élèves y sont évalués comme autrefois et y obtiennent des résultats semblables, contrairement à ce qui se passe pour la dictée. Le point de différence essentiel – la longueur du texte produit – joue en faveur des élèves d'aujourd'hui (80 % de copies de plus de 20 lignes en 1995, 56 % en 1920), tandis que les autres facteurs sont soit identiques soit en balance (la maîtrise dans l'emploi des temps est un peu meilleure en 1920 mais la ponctuation et la présentation graphique sont meilleures aujourd'hui).

À l'inverse de la dictée, des résultats en parfaite continuité en rédaction

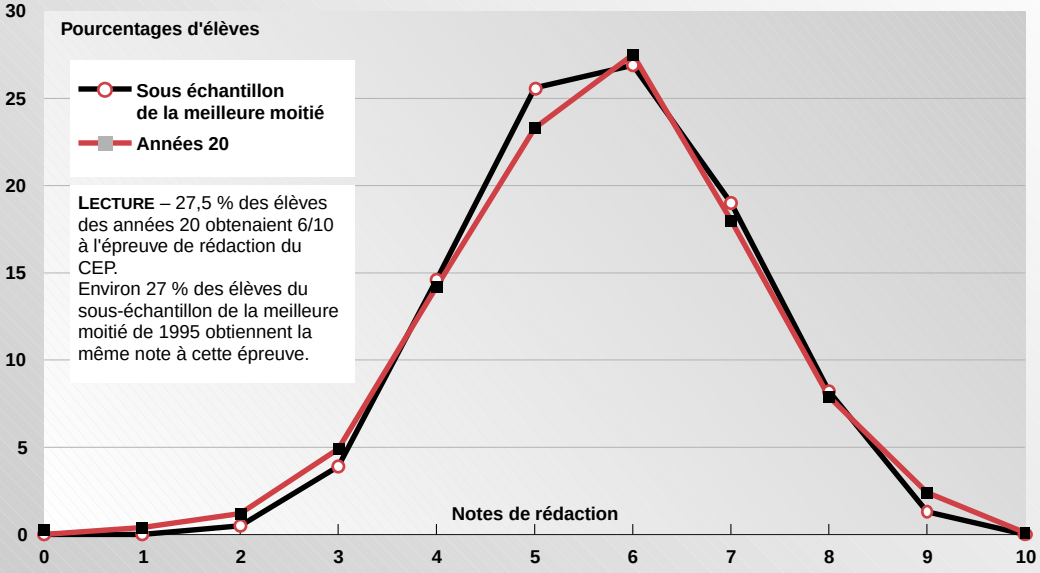
Comme pour la dictée, on peut tirer cinq conclusions de ces comparaisons de résultats – mais des conclusions parfois différentes :

– les distributions des résultats globaux au contrôle continu et à l'épreuve de rédaction sont proches ;

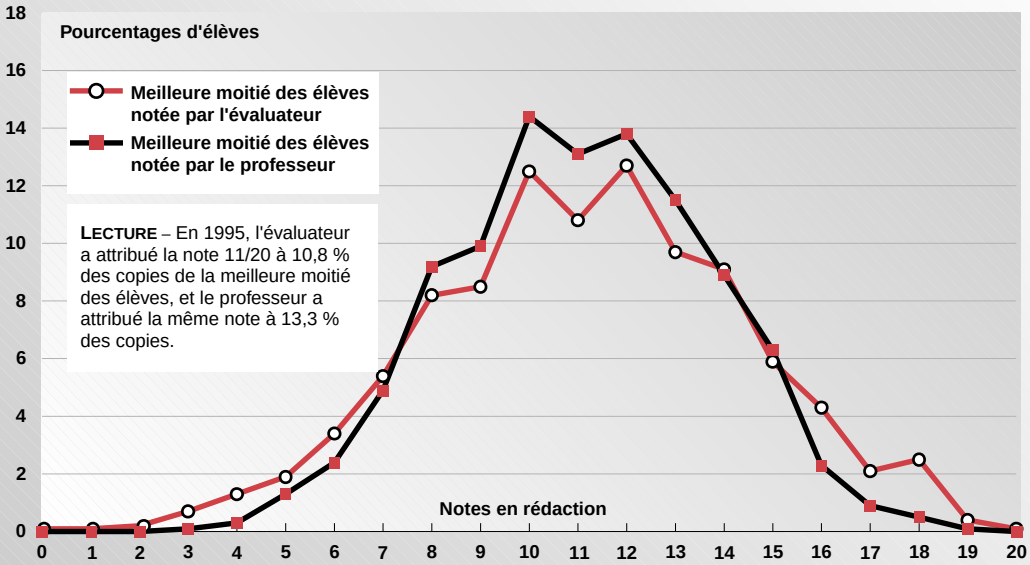
– la corrélation entre les résultats à l'une et l'autre épreuve est faible car les notes sont relativement dispersées autour d'une note moyenne, même si, statistiquement, les résultats augmentent en parallèle dans les deux épreuves ;

– les sujets de rédaction les mieux réussis sont ceux qui permettent de retrouver ou construire le plus facilement un scénario pour le texte à produire ;

GRAPHIQUE 9 – Épreuve de rédaction du CEP : résultats des élèves des années 20 et des meilleurs élèves de 1995



GRAPHIQUE 10 – Liaison entre note donnée par le professeur et note donnée par l'évaluateur en 1995 (Sous-échantillon de la meilleure moitié des élèves)



– les élèves qui ont les meilleurs résultats sont les élèves en avance, ce qui prouve que la rédaction est un bon indice de la réussite globale, mais les élèves qui progressent le plus sont les élèves les moins bons, sans qu'on puisse expliquer pourquoi les résultats des meilleurs semblent plafonner en quatrième ;

– enfin, les critères de jugement des enseignants sont remarquablement stables, lorsqu'il s'agit de noter les copies d'hier ou d'aujourd'hui, ce qui semble prouver que l'ancien exercice de rédaction s'est perpétué dans l'expression écrite. Les résultats des élèves des années 1920 et de 1995 sont, de ce point de vue, en parfaite continuité.

En introduction, nous proposions deux hypothèses pour expliquer les résultats des élèves d'aujourd'hui en dictée et en rédaction. La première mettait en cause le niveau général des élèves (qui se serait effondré en orthographe, et se serait renforcé en production de texte), la seconde la pertinence de l'exercice pour tester le niveau dans ces deux domaines.

S'agissant de la rédaction, on peut dire qu'elle demeure une forme d'exercice inchangé, malgré le changement d'appellation. Si l'on juge les critères de jugement des enseignants du primaire et du secondaire d'hier et d'aujourd'hui à travers leur notation, on pourrait conclure qu'ils se recouvrent parfaitement. L'entraînement reçu aujourd'hui par les élèves les conduit à une réussite supérieure à celle des années 1920, alors que les référents immédiats qui constituaient des aides pour les candidats au CEP (lectures modèles, familiarité des sujets, etc.) ont disparu. De la rédaction à l'expression écrite, il y aurait changement d'appellation plutôt que de forme, si bien que les élèves sont capables de s'adapter facilement lorsqu'on leur propose de se couler dans le moule de l'ancienne rédaction.

Il n'en est pas de même pour la dictée. Cet exercice typique de l'école primaire, pratiqué intensivement dans la préparation au CEP, produisait des résultats qu'on est loin de retrouver aujourd'hui, même chez les bons élèves. La dictée n'est donc plus une « forme scolaire » sur laquelle on compte pour produire de bons apprentissages, même si elle continue d'être utilisée ponctuellement au collège comme outil d'évaluation. Cependant, de l'effondrement des résultats en dictée, on ne peut conclure trop vite à l'effondrement des compétences orthographiques des élèves, pour deux raisons. D'une part, les résultats orthographiques des élèves de 1920 hors de la dictée sont nettement moins bons que celle-ci pourrait le laisser croire. La compétence en dictée n'est pas un indicateur suffisamment fiable du niveau général en orthographe, pas plus hier qu'aujourd'hui. D'autre part, le niveau en orthographe des élèves d'aujourd'hui, nettement inférieur à celui des années 1920, progresse régulièrement jusqu'en quatrième et ce sont les mauvais élèves qui progressent le plus. Quels sont les facteurs qui produisent ces progrès ? D'après nos analyses, la vigilance attentionnelle, la capacité à s'autocorriger joueraient un rôle essentiel dans les résultats, ce qui expliquerait que le niveau en orthographe soit un indicateur du niveau de réussite globale des élèves. Cela permettrait d'imaginer de nouvelles voies pour aider les élèves à progresser dans ce domaine.

À LIRE

C. PONS, « Connaissances en français et en calcul des élèves des années 20 et d'aujourd'hui », *Note d'Information,* 96.19, MEN-Direction de l'évaluation et de la prospective, avril 1996.
Connaissances en français et en calcul des élèves des années 20 et d'aujourd'hui - Comparaison à partir des épreuves du certificat d'études primaires, Les dossiers d'Éducation et Formations, n° 62, MEN-Direction de l'évaluation et de la prospective, février 1996.

Les différentes façons d'évaluer le niveau des élèves en fin de collège

Évaluation et notation des élèves (problèmes de mesures)

Comment évaluer correctement le niveau des élèves ? Le problème n'est pas simple. On dispose d'au moins trois types de mesures : les notes de contrôle continu, les résultats aux examens, et les scores aux évaluations menées par la Direction de la programmation et du développement. Or, si ces trois méthodes aboutissent dans l'ensemble, pour les élèves en fin de collège, à des conclusions assez convergentes, leur adéquation est loin d'être parfaite. En fait, il semble difficile d'élaborer un indicateur idéal du niveau des élèves, et ce, pour différentes raisons, telles que la variation de leurs performances, ou la subjectivité des professeurs.
Cet article propose diverses pistes méthodologiques pour tenir compte de la complexité des phénomènes étudiés et leurs interactions.

Fabrice MURAT
Bureau de l'évaluation des élèves
Direction de la programmation et du développement

L'évaluation des connaissances des élèves arrivés en fin de troisième – en l'occurrence, celle réalisée en juin 1995 – permet de déterminer pour chacun d'entre eux un score mesurant son niveau dans chaque discipline. Idéalement, ce score devrait donner des résultats sinon identiques, du moins équivalents à ceux obtenus par ces deux autres méthodes de mesure des connaissances que sont les moyennes de contrôle continu (calculées ici sur les deux dernières années scolaires, et fondées sur la notation par les professeurs) et les notes obtenues au brevet. Cependant, une analyse comparée de ces trois méthodes pour une même cohorte d'élèves montre des différences de performances qui peuvent être importantes, et qui amènent à s'interroger sur les outils de mesure utilisés et sur leur validité.

On posera en préalable que les comparaisons entre les différentes moyennes obtenues par les élèves (tableau 1) sont parfois hasardeuses. Elles sont fondamentalement relatives. On ne peut en particulier en tirer de conclusions sur le « niveau » global des élèves selon les disciplines, les épreuves ainsi que les exigences des professeurs étant trop différentes[1].

L'intérêt du tableau 1 est en fait de permettre de voir si, pour une même discipline, les différents types d'évaluation ciblent le même public d'élèves, et si la corrélation entre eux est simple (c'est-à-dire

NOTE

1. Il suffit pour s'en convaincre d'observer que les élèves de LV2, en anglais comme en allemand, ont des notes de contrôle continu supérieures à celles des élèves de LV1, ce qui ne correspond absolument pas à une supériorité réelle, comme le montre l'étude des scores, fondés sur des épreuves identiques en LV1 et LV2.

TABLEAU 1 – Moyennes et écarts-types des scores et des notes (contrôle continu et examens) pour les disciplines évaluées

	Scores		Notes de contrôle continu		Notes à l'examen	
	Moyenne	Écart-type	Moyenne	Écart-type	Moyenne	Écart-type
Allemand LV1	13,2	3,6	12,1	3,3		
Allemand LV2	10,3	3,7	12,8	3,1		
Anglais LV1	11,7	3,5	11,1	3,2		
Anglais LV2	11,8	3,2	13,0	3,2		
Espagnol LV2	7,9	3,2	11,6	2,9		
Français	14,2	2,5	10,9	2,5	9,8	3,3
Histoire-géographie	12,8	2,7			10,8	3,2
Mathématiques	12,3	3,5	10,9	3,4	10,5	4,6
Sciences physiques	9,5	3,1	11,4	2,8		
Sciences de la vie et de la Terre	11,3	2,4	11,6	2,7		
Technologie	15,1	2,5	13,0	2,2		

Remarque : les scores et les notes sont calculés sur 20.

linéaire). Deux exemples peuvent illustrer ce phénomène : les liens entre notes de contrôle continu et scores de juin 1995 pour les mathématiques et pour le français (graphiques 1.2 et 2.2). Pour les mathématiques, la dépendance entre note et score apparaît assez clairement linéaire ; et on remarque dans le tableau 1 que les moyennes sont effectivement assez proches. En revanche, pour le français (où les protocoles d'évaluation à l'origine du score sont nettement plus faciles que le contrôle continu), la tendance est plutôt « courbe ». Une épreuve facile permet de distinguer les élèves ayant quelques problèmes de ceux qui sont en grande difficulté, car ils échouent même aux questions les plus simples. Ainsi les élèves ayant en français une note comprise entre 8 et 10, ont un score compris entre 6 et 13 : on a en quelque sorte procédé à un *zoom* sur ces élèves pour isoler les plus faibles (en termes mathématiques, on pourrait dire qu'on a changé de métrique en élargissant les écarts pour les notes les plus basses). C'est ce qui explique la déformation de la courbe.

UNE PREMIÈRE APPROCHE DES VARIATIONS INDIVIDUELLES

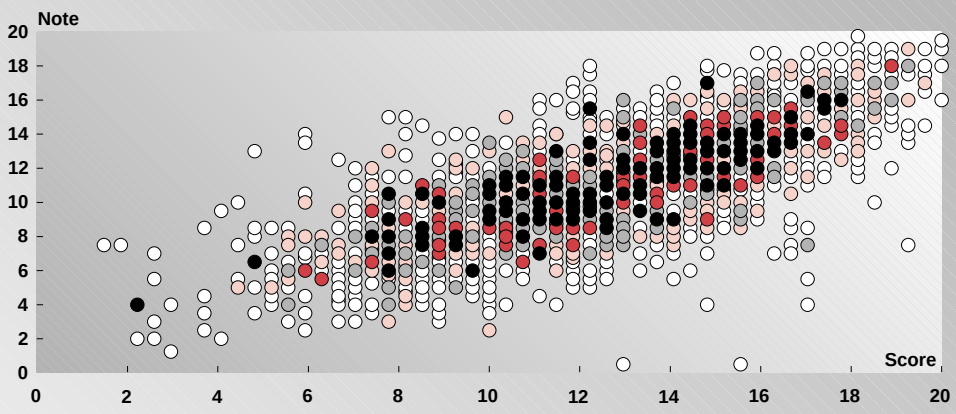
Les graphiques 1.1 et 2.1 permettent de se faire une première idée de l'ampleur de la dépendance entre score de l'enquête de juin 1995 et note de contrôle continu ; ils présentent le croisement de ces deux types d'évaluations, en mathématiques et en français, pour tous les élèves concernés (environ 1 500). On voit que si la dépendance entre les variables est nette (les coins sud-est et nord-ouest sont à peu près vides), les individus sont toutefois assez dispersés autour de la tendance. Ces variations individuelles n'ont rien d'étonnant et plusieurs facteurs peuvent les expliquer – indépendamment bien sûr d'éventuelles erreurs de traitements informatiques.

Les différentes évaluations n'ont pas le même objectif[2]. Elles ne sont pas non plus de même nature. La note de contrôle continu dépend fortement du professeur, tant à cause de la correction que du choix des sujets de devoir. En d'autres termes, il existe des professeurs sévères et d'autres plus indulgents ; il s'agira d'évaluer cet « effet-classe » (le fait que la note est calculée sur deux ans complique un peu la recherche puisque, si l'élève a changé de professeur, il y a deux « effets-classe » qui se combinent, voire s'annulent). À un moindre degré, les notes du brevet des collèges souffrent de la variation des sujets (différents selon les académies) et des correcteurs. Seuls

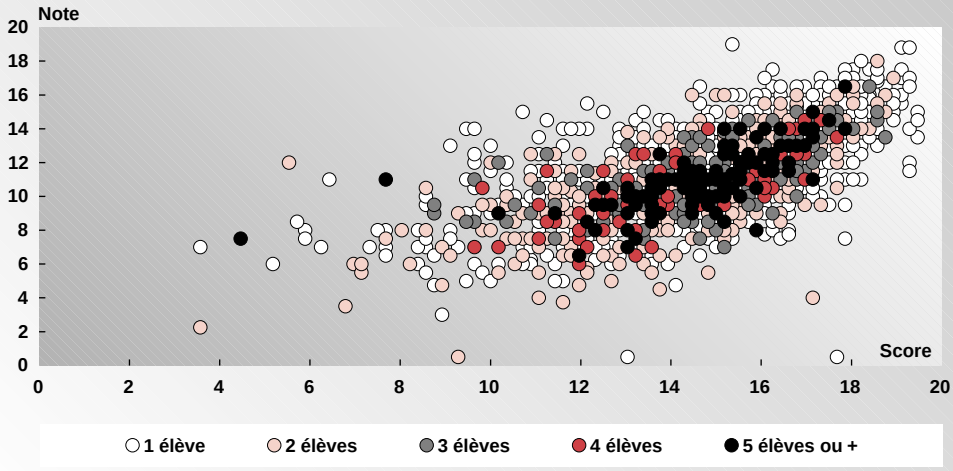
NOTE

2. Les protocoles d'évaluation de juin 1995 visaient à déterminer ce que savent les élèves et ce qu'ils ne savent pas, au niveau de l'ensemble de la cohorte en fin de collège ; la notation en classe (contrôle continu) a un but diagnostique et se situe au niveau individuel ; les notes du brevet des collèges sanctionnent un examen.

GRAPHIQUE 1.1 – Note de contrôle continu en fonction du score, niveau Élève (mathématiques)



GRAPHIQUE 2.1 – Note de contrôle continu en fonction du score, niveau Élève (français)



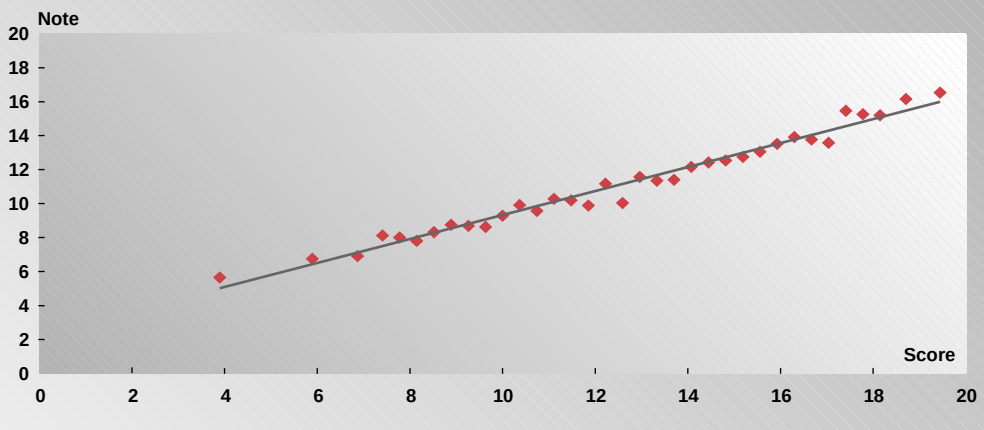
les scores de juin 1995 peuvent être considérés comme « objectifs » puisque les élèves ont tous passé le même protocole et que la correction était standardisée.

Il faut tenir compte aussi de la période à laquelle, et pendant laquelle, est faite l'évaluation. La note de contrôle continu est calculée sur les deux dernières années scolaires ; les notes à l'examen et les scores sont obtenus, eux, en fin de troisième, à un moment donné. On risque donc de constater des divergences dues au fait que l'élève a nettement progressé. Plus généralement, les conditions de passation peuvent provoquer des incohérences : il est par exemple difficile d'évaluer les compétences en technologie par une épreuve papier-crayon ; il est aussi très délicat d'évaluer objectivement l'expression orale en langue vivante et la production de texte en français.

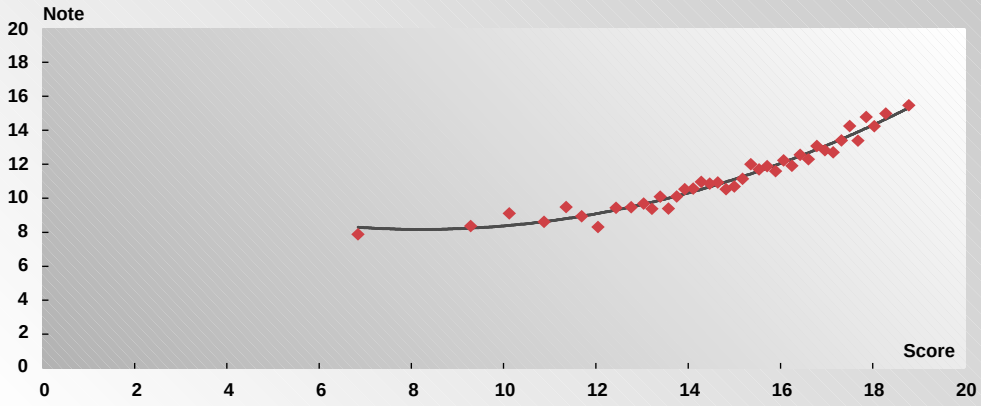
Enfin, une grande part des divergences sont d'origine individuelle : un élève n'a pas un niveau constant ; il a des points forts et des points faibles. De plus, ses caractéristiques (âge, sexe…) ont peut-être une influence sur la façon dont il est noté.

On voit que de nombreux phénomènes peuvent expliquer des variations individuelles. C'est pourquoi les corrélations qui seront mises en évidence ne sont pas parfaites. Il est toutefois possible, en raisonnant à un niveau agrégé, de montrer que la dépendance entre les différents systèmes d'évaluation est nette. Pour ce faire, on a regroupé les élèves ayant le même score et calculé la note moyenne correspondante – en veillant à ce que les groupes soient suffisamment grands, c'est-à-dire en travaillant par intervalles de scores au besoin. Ainsi ont été construits les graphiques 1.2 et 2.2, où la dépendance apparaît clairement (en tenant compte de la « déformation » déjà évoquée pour ce qui concerne le français). Ceci indique que, même si les évaluations diffèrent par leur nature et leurs objectifs, elles ont un fort lien de parenté et

GRAPHIQUE 1.2 – Note de contrôle continu en fonction du score, niveau Agrégé (mathématiques)



GRAPHIQUE 2.2 – Note de contrôle continu en fonction du score, niveau Agrégé (français)



mesurent effectivement à peu près la même chose : les élèves ayant un score faible ont *en moyenne* aussi une note faible (la relation au niveau agrégé étant presque parfaite). À un niveau individuel, en revanche, on observe de fortes variations : deux élèves ayant le même score peuvent avoir des notes différentes, et inversement, deux élèves ayant la même note peuvent avoir des scores différents ; même si leurs résultats à l'une des évaluations se trouvent assez près de ce que prédit l'autre d'après la courbe de tendance. La principale source de perturbations semble donc l'instabilité des performances et le fait qu'aucune évaluation ne peut prétendre mesurer à la perfection le niveau d'un élève dans une discipline donnée.

LES CORRÉLATIONS ENTRE LES ÉVALUATIONS

Les analyses qui précèdent demeurent très qualitatives. Il est nécessaire, pour mesurer avec une certaine précision l'ampleur de la dépendance entre les différentes mesures et pour la comparer d'une discipline à l'autre, d'utiliser un indicateur synthétique. Pour ce faire, celui qui semble le mieux adapté est le R^2, qui est le carré de ce qu'on appelle le coefficient de corrélation linéaire. Cet indicateur, compris entre 0 et 100 %, représente la part des variations de la variable expliquée (par exemple la note d'allemand LV1) que l'on peut prédire à l'aide de la variable explicative (le score). Il indique s'il existe une dépendance parfaite (de type linéaire) entre les variables (R^2 proche de 100 %) ou si les variations individuelles mises en évidence l'emportent (R^2 proche de 0 %).

La comparaison portant sur l'ensemble des élèves

➔ *L'indicateur varie selon la discipline*

La première partie du tableau 2 (colonnes 2 et 3) présente les résultats obtenus quand on travaille sur l'ensemble des élèves concernés. On voit que l'on se trouve dans une situation intermédiaire : l'indicateur varie de 30 % à 50 % environ suivant la discipline (si

TABLEAU 2 – R^2 global, « interclasses » et « intraclasse » pour chaque discipline

	Nombre d'élèves	R^2 (%)	Nombre de classes	R^2 (%)	R^2 moyen (%)	R^2 médian (%)	Nombre de classes au R^2 <10 %
Allemand LV1 (note de c.c./score)	1 233	39,9	85	16,1	58,5	62,3	5
Allemand LV2 (note de c.c./score)	1 060	36,1	81	12,7	52,0	57,5	7
Anglais LV1 (note de c.c./score)	3 497	49,2	166	33,9	61,7	66,8	5
Anglais LV2 (note de c.c./score)	1 126	43,5	83	31,8	54,0	62,4	9
Espagnol LV2 (note de c.c./score)	1 538	32,3	79	20,2	44,6	44,3	8
Français (note de c.c./score)	1 765	39,5	84	38,7	45,3	49,8	9
Français (note à l'examen/score)	1 623	32,3	78	40,9	31,4	33,3	19
Français (note de c.c./note à l'examen)	1 552	40,4	75	29,7	52,9	56,3	4
Histoire-géographie (note à l'examen/score)	1 666	28,9	78	26,6	33,2	34,5	14
Mathématiques (note de c.c./score)	1 772	52,4	83	48,5	58,1	63,7	5
Mathématiques (note à l'examen/score)	1 599	48,8	77	53,3	48,5	54,1	5
Mathématiques (note de c.c./ note à l'examen)	1 532	51,0	74	27,5	59,2	61,0	2
Sciences physiques (note de c.c./score)	1 700	47,4	80	47,7	52,0	54,0	4
Sciences de la vie et de la Terre (note de c.c./score)	1 735	29,5	81	23,0	34,4	34,1	7
Technologie (note de c.c./score)	1 947	10,8	89	2,0	21,6	17,5	29

LECTURE – c.c. : contrôle continu.
– pour chaque discipline, on a calculé les corrélations entre les différentes évaluations disponibles : d'une part, sur l'ensemble des élèves (deuxième et troisième colonnes, donnant l'effectif de l'échantillon et l'intensité de la relation) ; d'autre part, sur l'ensemble des classes (quatrième et cinquième colonnes, donnant le nombre de classes et l'ampleur de la relation à ce niveau) ; enfin, entre élèves au sein de chaque classe (les dernières colonnes donnent la moyenne et la médiane de l'ensemble de ces corrélations par classe, ainsi que le nombre de classes où « les choses se passent vraiment mal ».

l'on excepte la technologie, créditée de seulement 10,8 %). Les résultats sont de qualité variable suivant la discipline : le R^2 tourne autour de 50 % en mathématiques, en sciences physiques et en anglais LV1 ; il est proche de 40 % en anglais LV2, en français (moins nettement pour la relation entre score et note à l'examen dans cette discipline) et en allemand ; il dépasse peu ou pas 30 % en espagnol, en histoire-géographie et en sciences de la vie et de la Terre.

➔ *Des résultats équivalents par tableaux croisés*

Il est possible de se faire une idée plus concrète de la situation en comparant les classements des élèves selon les différentes évaluations dans trois exemples : relation entre score et note de contrôle continu en mathématiques, en anglais LV2, et en sciences de la vie et de la Terre (les R^2 étant respectivement de 52,4 %, 43,5 % et 29,5 %).

Pour chacune de ces trois disciplines, on procède à un croisement des quartiles d'élèves selon les scores et selon les notes. Les quartiles correspondent à quatre catégories d'élèves, définis comme les faibles, les moyens-faibles, les moyens-forts, les forts (tableau 3). Les effectifs des quartiles ne sont pas nécessairement exactement égaux car, de nombreux élèves ayant les mêmes résultats, il peut être impossible de les partitionner en quatre ensembles d'effectifs identiques.

Il s'agit ensuite, pour une discipline donnée, de comparer le classement selon le score et celui selon la note afin de savoir si les élèves classés comme faibles pour une évaluation le sont aussi pour l'autre.

L'étude des tableaux croisés amène à des conclusions identiques à celles obtenues avec les R^2 : on trouve une meilleure cohérence en mathématiques qu'en anglais LV2 et surtout qu'en sciences de la vie et de la Terre.

Ainsi, on voit que pour les mathématiques, 27,3 % de l'ensemble des élèves ont un score faible. Parmi ceux-ci, 17,4 % (de l'ensemble des élèves) ont une note faible ; si on rapporte ce pourcentage à la proportion d'élèves de score faible, on établit donc qu'il y a (17,4/27,3)100 = 63,7 % de chances qu'un élève de score faible ait aussi une note faible.

Un calcul équivalent donne 60,7 % pour l'anglais LV2 et 56,7 % en sciences de la vie et de la Terre.

TABLEAU 3 – Croisement des quartiles selon les scores et les notes de contrôle continu pour trois disciplines

Mathématiques

Score/note	Faibles	Moyens-faibles	Moyens-forts	Forts	Total
Faibles	17,4	7,5	2,0	0,5	27,3
Moyens-faibles	6,5	11,3	5,9	1,6	25,3
Moyens-forts	2,2	6,3	11,1	5,9	25,6
Forts	0,7	1,4	6,1	13,5	21,7
Total	**26,8**	**26,5**	**25,2**	**21,5**	**100,0**

Anglais (LV2)

Score/note	Faibles	Moyens-faibles	Moyens-forts	Forts	Total
Faibles	15,6	7,1	2,7	0,2	25,7
Moyens-faibles	7,8	7,6	6,3	2,6	24,3
Moyens-forts	3,0	6,0	8,5	7,8	25,3
Forts	1,0	2,5	6,0	15,2	24,7
Total	**27,5**	**23,3**	**23,5**	**25,8**	**100,0**

Sciences de la vie et de la Terre

Score/note	Faibles	Moyens-faibles	Moyens-forts	Forts	Total
Faibles	14,8	5,6	4,7	1,0	26,1
Moyens-faibles	9,1	8,5	6,0	3,2	26,8
Moyens-forts	3,8	5,8	7,9	5,6	23,2
Forts	2,2	3,3	6,0	12,4	23,9
Total	**29,9**	**23,3**	**24,6**	**22,2**	**100,0**

LECTURE – En colonnes, répartition des élèves selon la note ; en lignes, selon le score.

Si on considère maintenant, pour les mathématiques, les élèves se situant dans la meilleure moitié selon le score (c'est-à-dire en réunissant les forts et les moyens-forts), on calcule de la même façon qu'ils ont 77,4 % de chances d'être dans la meilleure moitié selon la note. Les résultats équivalents sont, pour l'anglais LV2, 75,0 % ; et pour les sciences de la vie et de la Terre, 67,7 %.

Une remarque générale s'impose en étudiant ces classements : la cohérence est plus grande entre les classes extrêmes qu'entre les classes intermédiaires. Ainsi, en sciences de la vie et de la Terre, 14,8 % des élèves se trouvent dans la case « faible (note)/faible (score) », mais seulement 8,5 % dans la case « moyen-faible (note)/moyen faible (score) ».

➔ *L'étude de la cohérence interne des scores*

Grâce aux R^2 ou aux tableaux croisés, on peut comparer les disciplines entre elles, déterminer celle où la corrélation entre les différentes évaluations est la plus forte. Il est plus difficile d'élaborer un jugement absolu : peut-on considérer que la dépendance en mathématiques est forte, qu'elle est faible en sciences de la vie et de la Terre ? On manque de références. En particulier, comment isoler ce qui est dû à des différences au niveau de l'instrument de mesure (les évaluations ne mesurent pas le niveau de l'élève de la même façon) de ce qui est dû à l'« objet » de la mesure (l'élève lui-même, dont les performances sont très variables) ?

Pour tenter de mesurer les variations de performance indépendamment de la méthode d'évaluation, on ne s'intéresse qu'aux protocoles de juin 1995. On divise aléatoirement, pour chaque discipline, les items en deux groupes (les items de numéro pair et les items de numéro impair), ce qui permet de construire deux nouveaux scores. On étudie ensuite la relation linéaire entre ces deux scores, comme précédemment entre score et note. Le R^2 est alors de l'ordre de 85 % en histoire-géographie, 80 % en anglais (LV1 et LV2) et en sciences physiques, 75 % en français et en mathématiques, 65 % en allemand (LV1 et LV2), en espagnol et en sciences de la vie et de la Terre, 40 % en technologie.

Le nombre d'items utilisés dans chaque score (la moitié du nombre total) intervient de façon sensible dans leur stabilité : ce sont dans les disciplines aux items les plus nombreux que les scores sont les plus stables (histoire-géographie, anglais). On remarque toutefois que l'adéquation n'est pas parfaite, même dans les meilleurs cas, alors qu'aucun des facteurs évoqués (différences d'objectifs, effet-classe, conditions de passation…) n'entrent en ligne de compte, même si rien n'assure que les deux sous-scores soient représentatifs du niveau de l'élève, l'épreuve n'ayant pas été conçue pour qu'ils le soient. Cette méthode met donc en évidence l'influence des variations de performance de l'élève, d'une compétence ou d'un domaine sur l'autre (on ne tient pas compte des variations dans le temps) ; mais elle montre aussi que d'autres facteurs doivent intervenir dans la relation entre score et note, puisque les R^2 que l'on trouve ici sont tout de même nettement plus élevés que ceux observés en comparant des évaluations de natures différentes.

On peut mettre ce phénomène en évidence en revenant au niveau élémentaire de l'item. Prenons un exemple tiré de l'épreuve de français. Le premier exercice proposait un texte contenant une vingtaine

de fautes d'orthographe. Les élèves devaient récrire le texte en corrigeant les fautes. En particulier, le texte comportait l'expression « tu m'a » qui devait être corrigée en « tu m'as », et l'expression « j'en met » qui devait être corrigée en « j'en mets ». La première correction est effectuée par 61 % des élèves ; la seconde par seulement 46 %. Bien que ces deux questions ne soient donc pas de difficultés identiques, elles mesurent a priori des compétences proches. On s'attendrait à ce que ceux qui corrigent l'une des fautes, corrigent assez souvent l'autre. C'est le cas, mais d'une façon qui est loin d'être massive. En effet, parmi ceux qui repondent bien à la première question, 52 % répondent bien à la seconde (à comparer au 46 % sur l'ensemble des élèves) ; le rapport est de 68 % contre 61 % quand on inverse l'ordre des questions dans la comparaison. Des tests peuvent montrer qu'il existe une corrélation significative mais elle n'est pas parfaite. On comprend mieux les incohérences qui demeurent quand on regroupe les items pour calculer un score (ce regroupement est d'ailleurs profitable car il amène à des indicateurs plus robustes).

Il existe une autre façon de juger de la qualité d'une relation : elle consiste à comparer le score dans une discipline, non seulement avec la note qui lui correspond, mais avec toutes les notes disponibles[3]. Les corrélations sont presque toutes significatives, ce qui signifie simplement qu'un élève bon dans une matière le sera souvent dans une autre. On remarque toutefois que les scores sont sensiblement mieux corrélés avec la ou les notes qui leur correspondent, ce qui montre que ces deux mesures s'attachent bien au même objet, d'une façon spécifique.

La comparaison portant sur les classes, et la mise en évidence d'un effet-classe

Parmi les nombreux facteurs qui peuvent expliquer pourquoi les corrélations entre évaluations ne sont pas parfaites, il en est un dont il est particulièrement important de connaître l'influence : c'est l'effet-classe, c'est-à-dire le fait que chaque professeur est responsable des notes qu'il donne et qu'il peut le faire comme il l'entend. Il est vraisemblable de supposer qu'il existe des professeurs plus sévères que d'autres. On attend tout au moins qu'au sein d'une même classe, la notation du professeur permette de distinguer les bons élèves des mauvais et qu'à peu de choses près, la hiérarchie qu'elle induit corresponde à celle qu'on obtient à partir des scores. Il est plus délicat de demander au professeur d'harmoniser les notes qu'il donne avec celles qui sont données dans une autre classe, dans un autre collège, aux caractéristiques sans doute très différentes. Peut-on noter les élèves d'un collège parisien renommé comme des élèves en ZEP ? Est-il souhaitable dans une classe très faible de ne pas mettre de note au-dessus de 8 ? Il faut cependant tenir compte du fait que les notes de contrôle continu sont utilisées pour l'obtention du brevet, ce qui suppose qu'elles soient équitables. Peut-on vraiment les considérer comme telles ? Quelques précautions sont indispensables dans l'utilisation des résultats. On a vu en effet que les protocoles d'évaluation et le contrôle continu n'ont pas tout à fait les mêmes objectifs ni les mêmes méthodes, ce qui peut expliquer des divergences. De plus, les notes de contrôle continu dont on dispose ont été calculées sur deux ans ; les résultats qui en découlent sont certainement atténués par rapport à la réalité à cause de l'influence « parasite » des notes de quatrième (ou de la précédente troisième pour les redoublants), dans une classe et avec un professeur aux caractéristiques peut-être très différentes. En pratique, cette atténuation est d'ailleurs un phénomène plutôt positif puisqu'elle rend les notes moins subjectives et plus exploitables dans l'optique du diplôme. Enfin, il faut tenir compte de la procédure d'échantillonnage : on a tiré une seule classe par établissement, dont on a interrogé tous les élèves. Cela signifie que l'on compare des classes d'établissements différents ; les différences entre les classes peuvent être dues en partie à des différences entre collèges.

On peut étudier la relation entre la note moyenne d'une classe et son score moyen, comme a été étudiée celle entre la note d'un élève et son score. En comparant la qualité de cette relation entre classes avec celle que l'on trouve entre élèves, on aura une première idée de l'influence de ce niveau.

Les indicateurs rendant compte de la relation entre les différentes formes d'évaluation au niveau Classe se trouvent dans la deuxième partie du tableau 2 (colonnes 4 et 5). Ils sont parfois sensiblement différents de ceux du niveau Élève. En effet, si les R^2 sont relativement proches de ceux de la première partie du tableau pour les mathématiques, les sciences physiques, le français et la technologie, il n'en va pas de même pour les sciences de la vie et de la Terre (23,0 % au lieu de 29,5 %), la technologie (2,0 % au lieu de 10,8 %), et surtout les langues vivantes (en espagnol, par exemple, le R^2 passe de 32,3 % au niveau Élève à 20,0 % au niveau Classe) et la relation entre note de contrôle continu et note à

NOTE

3. Dans *Évaluation pédagogique en fin de troisième générale et technologique,* (les dossiers d'Éducation et formations, n° 86, mai 1997, MEN-DEP), est proposé (p. 181) un tableau présentant la corrélation entre toutes les notes et tous les scores.

l'examen en français (de 40,4 % à 29,7 %) et en mathématiques (de 51,0 % à 27,5 %).

Puisque la note et le score ne correspondent pas parfaitement sur l'ensemble des classes, c'est qu'il existe un effet-classe. Certains professeurs attribuent à l'ensemble de leur classe une note moyenne qui est supérieure ou inférieure à ce que le score moyen laisserait attendre. En d'autres termes, une classe ayant une bonne note moyenne n'aura pas toujours un score moyen élevé.

Un tel résultat était prévisible, compte tenu de ce qui a été établi au niveau Élève. Les incohérences entre classes sont en grande partie le reflet des variations de performances individuelles. Sachant que l'on observe une corrélation moyenne ou mauvaise entre score et note sur l'ensemble des élèves, il était peu probable d'observer une amélioration en regroupant les élèves par classe. Aussi, la situation normale est celle où l'on observe des R^2 comparables aux deux niveaux. On ne peut affirmer l'existence d'un effet-classe que dans le cas où une dégradation de la corrélation est observée pour ce niveau, c'est-à-dire quand le R^2 entre classes est bien plus mauvais que celui entre élèves. Pour l'allemand LV1 par exemple, la corrélation au niveau Classe, si on compare le score et la note de contrôle continu, dépasse à peine 20 %, alors qu'elle approche de 40 % au niveau Élève. Cela signifie que l'on observe un lien assez fort au niveau Élève (un élève brillant pour une évaluation le sera pour l'autre), qui est bien moins net au niveau Classe (une classe ayant une moyenne élevée n'aura pas forcément un score moyen élevé : les professeurs d'allemand attribuent un niveau de note à leur classe plus en fonction de leur propre « sévérité » qu'en fonction du niveau réel de la classe, lequel est assez bien mesuré par les scores « objectifs »). On peut donc penser que dans cette discipline, les choses se passent mieux au sein de chaque classe, mais que la situation se dégrade du fait que les professeurs n'harmonisent pas leurs notes entre eux.

Les exemples du français et des mathématiques permettent l'étude de la relation entre score et note à l'examen (et non plus note de contrôle continu). Les relations au niveau Classe sont créditées de R^2 meilleurs que celles au niveau Élève (40,9 % au lieu de 32,3 % en français ; 53,3 % au lieu de 48,8 % en mathématiques). Puisque les deux évaluations peuvent être considérées comme objectives, et que dans les deux cas la marge de manœuvre du correcteur est minimale, il est normal de ne pas observer d'influence de la classe (c'est-à-dire du professeur qui corrige) ni par conséquent de détérioration du R^2. Si l'on observe une amélioration, cela tient sans doute au fait que l'on travaille sur une population composée de moins d'individus (il y a moins de classes que d'élèves). En effet, on attribue d'autant plus de « significativité » à un R^2 donné que la population est nombreuse. En termes pratiques, trouver une tendance linéaire sur cinq individus est rassurant, mais il est bien plus intéressant d'en observer une sur mille individus. Il ne faut donc pas s'étonner si l'on trouve des R^2 supérieurs au niveau Classe en cas d'absence d'incohérence ; le fait d'en trouver de moins bons est par contre d'autant plus significatif.

La comparaison portant sur les élèves d'une classe

Les résultats qui précèdent amènent à se demander si le lien entre les évaluations n'est pas meilleur quand on travaille au sein d'une classe donnée. Pour cela, on calcule le R^2 non plus sur l'ensemble de la population des élèves mais pour les élèves de chaque classe. Il n'est bien sûr pas question de présenter ici le R^2 correspondant à chacune d'entre elles. On trouvera simplement, dans la troisième partie du tableau 2 (colonnes 6 et 7), la moyenne et la médiane de la distribution de ces R^2 suivant la classe.

Les indicateurs sont en général plus élevés que ceux que l'on obtient au niveau global. Tout en restant prudent (pour une raison déjà évoquée : comme il y a bien moins d'individus dans une classe que dans la population totale, il est normal que les R^2 soient plus élevés), on peut énoncer que la situation est meilleure quand on se place au sein d'une classe donnée. L'écart est particulièrement net en langues vivantes : en allemand LV1, le R^2 entre le score et la note de contrôle continu dépasse 62,3 % pour la moitié des classes. Or il est à peine supérieur à 20 % quand on utilise la classe comme unité statistique. Il semble bien que les professeurs d'allemand classent convenablement leurs élèves les uns par rapport aux autres au sein de leur classe, mais qu'ils ne donnent pas à la classe dans son ensemble la moyenne qui lui correspond. Les comparaisons entre élèves de classes différentes se trouvent par conséquent un peu faussées, d'où la qualité moyenne du R^2 au niveau global.

Il existe un certain nombre de classes pour lesquelles il ne semble y avoir aucun rapport entre score et note : les R^2 sont inférieurs à 10 %. Plutôt que de supposer un comportement particulier du professeur, il est plus plausible d'invoquer dans nombre de cas des erreurs informatiques dans la liaison entre le score et les notes. La preuve en est que l'on a très peu de classes de ce type quand on compare notes de contrôle continu et notes à l'examen, qui étaient collectées ensemble, indépendamment des exercices du protocole de juin 1995.

Une étude par discipline

Les résultats par discipline se lisent sur le tableau 2 en sens inverse (de droite à gauche), ce qui correspond mieux à la réalité : on part d'un certain niveau de corrélation au sein de chaque classe, qui peut être détérioré ou non par une adéquation plus ou moins bonne entre les classes ; on obtient enfin la corrélation au niveau global.

En allemand, la corrélation au sein de chaque classe est assez bonne (le R^2 au sein d'une classe donnée est en moyenne de 58,5 % en LV1, de 52,0 % en LV2 ; les R^2 médians sont encore meilleurs). Elle est en revanche plutôt mauvaise entre classes (autour de 20 %). Cela explique pourquoi la corrélation au niveau global ne dépasse pas 40 %.

En anglais, les conclusions sont identiques, à ceci près que les modèles s'ajustent mieux, particulièrement en LV1 (on atteint 49 % de corrélation au niveau global). On constate en effet que la cohérence entre classes est meilleure qu'en allemand (33,9 % en LV1, 31,8 % en LV2), même si elle reste inférieure à ce qu'on observe au sein des classes (le R^2 moyen dépasse 60 % en LV1).

Cette tendance s'observe aussi en espagnol mais les corrélations sont un peu moins bonnes qu'en allemand (R^2 de 32 % sur l'ensemble des élèves).

En français, on a déjà remarqué qu'il n'existe pas d'effet-classe significatif quand on compare score et note à l'examen (le R^2 entre classes est même le plus élevé de tous les indicateurs). La comparaison des R^2 « interclasses » et « intraclasse » pour les deux autres croisements semble indiquer en revanche une influence du niveau Classe quand on compare le score ou la note à l'examen avec la note de contrôle continu ; ceci est plus net quand il s'agit de comparer note à l'examen et note de contrôle continu : on passe de 52,9 % pour le R^2 intraclasse moyen à 29,7 % pour le R^2 interclasses.

En histoire-géographie, on ne constate pas non plus d'effet-classe marqué. La corrélation n'est d'ailleurs pas très élevée – autour de 30 % selon les modèles. Rappelons que l'on comparait le score avec la note à l'examen.

En mathématiques, les modèles s'ajustent bien, avec des R^2 d'environ 50 %. Les conclusions en ce qui concerne d'éventuels effets-classes sont les mêmes qu'en français : ils n'existent pas quand on compare score et note à l'examen, mais apparaissent si l'on compare l'un des deux à la note de contrôle continu. Là encore, c'est particulièrement net lorsque l'on confronte note à l'examen et note en contrôle continu (le R^2 intraclasse moyen est de 61,0 % ; le R^2 interclasses n'est que de 27,5 %).

En sciences physiques, l'adéquation linéaire est bonne (50 % de la variance est expliquée) et le niveau Classe ne semble pas avoir d'influence très nette.

En sciences de la vie et de la Terre, l'adéquation tourne autour de 30 %. L'influence du niveau Classe semble assez légère.

Enfin, les corrélations concernant la technologie sont assez faibles, ce qui pourrait indiquer que les deux évaluations pour cette discipline ne visent pas exactement les mêmes compétences.

En définitive, on peut résumer comme suit les résultats par discipline :

– en langues vivantes, la cohérence est bonne au sein des classes, mais comme elle est moins bonne entre les classes, on observe des corrélations assez moyennes au niveau global. On peut aussi placer dans ce groupe les confrontations entre note à l'examen et note de contrôle continu en français et en mathématiques, qui ont les mêmes caractéristiques ;

– en ce qui concerne la confrontation entre score et note à l'examen, elle est bonne en mathématiques, moins nette en français et en histoire-géographie. Il n'y a pas d'effet-classe (ou alors très léger, en histoire-géographie) ;

– la confrontation entre score et note de contrôle continu en mathématiques, français et sciences physiques, fait apparaître une corrélation assez bonne et une légère influence du niveau Classe. Cette influence est à peu près du même ordre en sciences de la vie et de la Terre et en technologie mais l'adéquation linéaire est nettement moins bonne.

❒ Influence de quelques variables sur la façon de noter

On n'a fait jusqu'à présent que mettre en évidence un effet-classe. Il convient de chercher à l'expliquer, en essayant même de répondre, par la même occasion, à une question plus générale : pourquoi un élève n'est-il pas noté comme il « devrait » l'être ?

Une méthode de mesure

On peut repérer les classes « anormales » de deux façons : soit les élèves au sein d'une classe donnée ne sont pas classés de la même manière suivant la note et suivant le score (instabilité interne) ; soit ces élèves sont systématiquement surnotés ou sous-notés par rapport à ceux des autres classes (instabilité externe).

Ces deux problèmes sont à peu près indépendants. L'étude du premier est assez délicate puisqu'il s'agit de s'intéresser aux professeurs qui, semble-t-il, ont

une conception de la notation très différente de celle utilisée dans les protocoles d'évaluation. Pour expliquer ces différences, il faudrait sans doute avoir plus d'informations sur les opinions et pratiques pédagogiques du professeur dans sa discipline. On s'intéressera plutôt à l'influence des caractéristiques de l'élève sur la façon dont le professeur le note, pour déterminer si elles peuvent être sources de biais et d'incohérences entre scores et notes.

On rejoint alors la problématique associée au second cas. Il s'agit de déterminer si un individu (que ce soit un élève ou une classe) a été surévalué ou sous-évalué. Plusieurs méthodes sont possibles pour décider que tel individu a été « mal » noté. Celle adoptée est relativement simple, même si elle nécessite un travail préliminaire sur les variables : on doit, en effet, faire en sorte que les évaluations soient de difficultés comparables. Cela consiste à caler toutes les notes et tous les scores sur une moyenne égale à 10. Par exemple, la note moyenne de contrôle continu en mathématiques est 10,9 (tableau 1) ; on enlève alors 0,9 point à chaque note. Le score moyen en français est 14,2 ; on retire 4,2 points à tous les scores.

Ceci fait, on calcule pour chaque élève la différence entre son score et sa note (« corrigés ») ; pour chaque classe, la différence entre le score moyen et la note moyenne.

Il suffit alors de comparer, pour chaque élève et pour chaque classe, cette différence avec celle observée sur la population, qui est nulle, puisqu'on a centré les variables sur 10. Ainsi, quand on observe pour un élève que la différence (score - note) est négative, cela signifie que le score est inférieur à la note, tandis que par construction ces deux variables sont égales sur l'ensemble de la population. L'élève est donc surnoté, il a mieux réussi le contrôle continu que les protocoles. Si la différence pour l'élève est positive (score supérieur à la note), il est sous-noté.

En d'autres termes, quand on veut savoir, pour une population donnée (par exemple, les élèves en retard scolaire), si elle est ou non sous-notée, on calcule la différence moyenne entre score et note et on en regarde le signe.

Résultats pour diverses populations

Le tableau 4 présente ces différences pour quelques populations intéressantes tout d'abord pour des variables relatives à la classe (niveau de la classe, appartenance à une ZEP), puis pour des variables relatives à l'élève (âge, redoublement).

Les résultats concernant l'influence du niveau de la classe confirment et éclairent ce qui a été dit sur l'effet-classe. Dans l'ensemble, on remarque que les élèves des classes faibles sont surnotés, tandis que ceux des classes fortes sont sous-notés. Ce phénomène assez naturel (les professeurs n'« osent » pas

TABLEAU 4 – Différences entre score et note corrigés pour diverses populations d'élèves

	Niveau de la classe				ZEP		Âge de l'élève		Redoublement	
	Faible	Moyen-faible	Moyen-fort	Fort	Hors ZEP	En ZEP	En retard	« A l'heure »	Oui	Non
Allemand LV1 (note c.c./score)	- 1,6	- 0,8	0,9	2,0	0,3	- 2,3	0,4	- 0,1	0,5	- 0,2
Allemand LV2 (note c.c./score)	- 2,1	- 0,9	0,3	2,6	0,0	- 0,4	0,1	- 0,1	1,1	- 0,1
Anglais LV1 (note c.c./score)	- 1,6	- 0,2	0,7	1,6	0,1	- 0,4	-,0,0	0,0	0,7	- 0,1
Anglais LV2 (note c.c./score)	- 1,2	- 0,4	0,1	1,5	- 0,0	- 0,2	- 0,2	0,1	1,0	- 0,0
Espagnol LV2 (note c.c./score)	- 1,9	- 0,3	0,7	2,0	0,0	0,2	0,2	- 0,2	1,0	- 0,1
Français (note c.c./score)	- 0,8	0,4	0,3	0,4	0,2	- 1,3	0,4	- 0,2	0,6	- 0,1
Français (note examen/score)	- 0,3	0,0	0,3	0,0	- 0,0	0,0	0,4	- 0,2	0,4	- 0,0
Français (note c.c./note examen)	- 1,2	- 0,3	0,5	1,9	0,1	- 1,4	0,0	- 0,1	0,3	- 0,1
Histoire-géographie (note examen/score)	- 1,1	0,5	0,1	0,6	-0,1	0,7	0,5	- 0,3	0,3	0,0
Mathématiques (note c.c./score)	- 1,2	- 0,2	0,5	1,3	0,1	- 0,9	- 0,1	0,0	0,9	- 0,1
Mathématiques (note examen/score)	- 0,0	0,2	- 0,1	- 0,1	- 0,1	0,4	0,4	- 0,3	0,2	- 0,0
Mathématiques (note c.c./note examen)	- 2,0	- 0,9	1,5	2,4	0,2	- 1,2	- 0,6	0,3	0,4	- 0,1
Sciences physiques (note c.c./score)	- 1,0	- 0,4	0,2	1,5	0,0	- 0,2	0,3	- 0,2	0,8	- 0,1
Sciences de la vie et de la Terre (note c.c./score)	- 0,9	0,1	0,3	0,7	- 0,1	0,5	0,5	- 0,3	0,8	- 0,1
Technologie (note c.c./score)	- 1,3	0,0	0,4	0,9	0,1	- 0,4	0,4	- 0,2	0,4	- 0,1

LECTURE – c.c. : contrôle continu.
Remarque : il faut comparer ces données avec la différence que l'on obtient quand on travaille sur l'ensemble de la population de référence, c'est-à-dire 0.
Si la différence est négative, cela signifie que le score est plus grand que la note ; la population est donc sous-notée.

mettre de trop mauvaises notes ou de trop bonnes) explique sans doute une grande part de l'effet-classe puisqu'il lui est en quelque sorte parallèle : il n'existe pas dans les croisements où l'on n'observait pas d'influence de la classe (confrontation score/note à l'examen) ; il est de faible ampleur pour les disciplines où le niveau Classe se manifestait faiblement (les élèves de classes faibles gagnant alors de 1 à 1,5 points par rapport à leur score, les élèves de classes fortes étant pénalisés dans les mêmes proportions) ; il est très net pour les autres croisements (on approche ou on dépasse 2 points d'écart entre les deux évaluations).

Le fait d'être en ZEP a aussi une influence sur la notation. Dans la plupart des disciplines, l'écart entre score et note de contrôle continu est négatif pour les élèves en ZEP, ce qui signifie, on l'a vu, qu'ils sont surnotés. Ce phénomène ne s'observe pas quand on confronte scores et notes à l'examen. Il semble même s'inverser en mathématiques et en histoire-géographie.

Le retard scolaire n'a pas d'influence très marquée même si les élèves âgés semblent légèrement sous-notés dans certaines disciplines.

En revanche, le redoublement est déterminant. Les redoublants sont assez systématiquement sous-notés (c'est beaucoup moins net, une fois encore, pour la comparaison score/note à l'examen). Il faut cependant rappeler que les notes de contrôle continu ont été calculées sur deux ans, ce qui inclut l'année à la fin de laquelle l'élève a redoublé ; on peut légitimement penser que les notes obtenues durant cette année étaient particulièrement mauvaises, ce qui explique pourquoi la note de contrôle continu est basse. L'élève peut néanmoins avoir suffisamment progressé durant la seconde année pour obtenir des résultats corrects à l'examen et à l'évaluation en fin de troisième ; la sous-notation observée serait alors plutôt le signe d'un effet bénéfique du redoublement.

L'impossible perfection

La méthode employée pour tenter d'expliquer l'effet-classe est assez simple mais elle a l'inconvénient d'établir un lien arbitraire entre score et note. Elle postule en effet qu'un point de score « vaut » un point de note et qu'il est donc légitime de faire la différence de l'un par l'autre. Une autre méthode est possible, celle du « résidu » (*cf.* « La méthode du résidu »). Elle n'est malheureusement pas satisfaisante, pour une raison qui illustre un aspect intéressant et paradoxal des problèmes de mesure de niveau des élèves, que l'on va évoquer rapidement.

La confrontation entre score et note de contrôle continu en mathématiques peut servir d'exemple. On va étudier l'influence de l'âge à score constant. On a vu qu'il n'y a aucune surnotation selon le retard dans ce cas. Si on ne garde que les élèves ayant un score entre 10 et 12 (351 dans l'échantillon), on constate que parmi eux, 145 ont du retard, 206 n'en ont pas. Leurs scores moyens sont respectivement 11,0 et 10,9 ; leurs notes, respectivement 10,2 et 9,5. On est tenté de dire que les élèves « à l'heure » sont surnotés en contrôle continu par rapport aux élèves en retard, puisqu'ils obtiennent une note supérieure alors que les scores sont quasiment égaux. On inverse maintenant la perspective (en fixant donc la note de contrôle continu) : on ne garde que les élèves ayant une note comprise entre 10 et 12 (374 dans l'échantillon) ; 141 ont du retard, 233 n'en ont pas. Leurs notes sont respectivement 10,9 et 10,6 ; leurs scores sont respectivement 12,4 et 11,4. Le score des élèves « à l'heure » semble surévalué par les protocoles de 1995 par rapport aux élèves en retard. Il apparaît donc que les élèves « à l'heure » sont surévalués à la fois par le contrôle continu et par les protocoles d'évaluation de 1995 ! On aboutirait aux mêmes conclusions en changeant de discipline ou en choisissant un autre intervalle de scores (tableau 5).

THÈME

TABLEAU 5 – Différences entre élèves en retard et élèves « à l'heure » pour une évaluation quand on fixe l'autre (exemple des mathématiques)

Score fixé					Note fixée				
Tranches de scores	Score des élèves :		Note des élèves :		Tranches de notes	Score des élèves :		Note des élèves :	
	« à l'heure »	en retard	« à l'heure »	en retard		« à l'heure »	en retard	« à l'heure »	en retard
06-08	7,4	7,3	8,1	7,5	06-08	10,0	8,8	7,0	6,8
08-10	9,0	8,9	9,1	8,1	08-10	11,2	10,4	8,8	8,8
10-12	11,0	10,9	10,2	9,5	10-12	12,4	11,4	10,9	10,6
12-14	13,0	12,8	11,5	10,4	12-14	14,1	13,0	12,7	12,7
14-16	15,0	14,9	12,9	12,2	14-16	15,3	14,5	14,6	14,8
16-18	17,0	16,8	14,6	11,3	16-18	16,6	15,1	16,5	16,7

Ce constat est *a priori* contradictoire : il signifie que deux élèves ayant le même score ou la même note…n'ont pas forcément le même niveau ! Cela tient à l'imperfection des évaluations et des liaisons entre elles. Tout d'abord, il ne faut pas s'étonner si en fixant une des évaluations, on observe des écarts sur l'autre : puisque le lien entre elles n'est pas parfait, il est normal que la deuxième évaluation reste assez libre. Ce qui est moins normal, c'est que l'écart soit systématiquement en faveur des élèves « à l'heure ». Cela tient au lien que l'on a souvent mis en évidence entre réussite scolaire et retard scolaire. Là où la première évaluation n'a pas suffi à discriminer les deux élèves, cette liaison apporte une information supplémentaire. On observe que certains élèves peuvent faire illusion sur une évaluation mais pas sur deux. En d'autres termes, un élève âgé qui réussit une évaluation a moins de chance d'en réussir une autre qu'un élève « à l'heure » qui a aussi bien réussi que lui la première évaluation.

Si cette explication, liée à l'instabilité des performances des élèves, est cohérente, le paradoxe évoqué ci-dessus montre, s'il en était besoin, que la mesure des connaissances est un problème très délicat. Il semble suggérer que seule une évaluation parfaitement objective, s'effectuant sur un temps assez long et comportant un nombre important de questions, serait capable de donner une image absolument fiable des capacités d'un élève.

Présentation de l'enquête

Le dossier d'Éducation et formations n° 86 (ministère de l'Éducation nationale, Direction de l'évaluation et de la prospective, mai 1997), présente un descriptif complet du dispositif TROISIÈME 95. Rappelons ici quelques éléments essentiels. Il faut savoir que les 6 963 élèves présents dans l'échantillon de troisième générale (les troisièmes technologiques ont été par commodité écartées du champ de cet article) ne passaient pas tous les protocoles disciplinaires (allemand LV1, allemand LV2, anglais LV1, anglais LV2, espagnol LV2, français, histoire-géographie, mathématiques, sciences physiques, sciences de la vie et de la Terre, technologie), mais seulement deux protocoles de langues (leur LV1 et leur LV2) et deux parmi les autres disciplines. Ainsi pour chaque discipline, on ne peut calculer un score que pour le tiers de l'échantillon, soit environ 1 500 élèves (exception faite de l'anglais LV1 qui concernait les deux tiers de l'échantillon global).

Les notes de contrôle continu et à l'examen devaient théoriquement être renvoyées pour tous les élèves ; mais certains secrétariats ont omis de le faire. Il y a donc quelques non-réponses. Pour mener à bien l'analyse, il a fallu se restreindre aux élèves pour lesquels on connaissait à la fois la note et le score. On trouve dans la deuxième colonne du tableau 2 les effectifs correspondants pour chaque discipline.

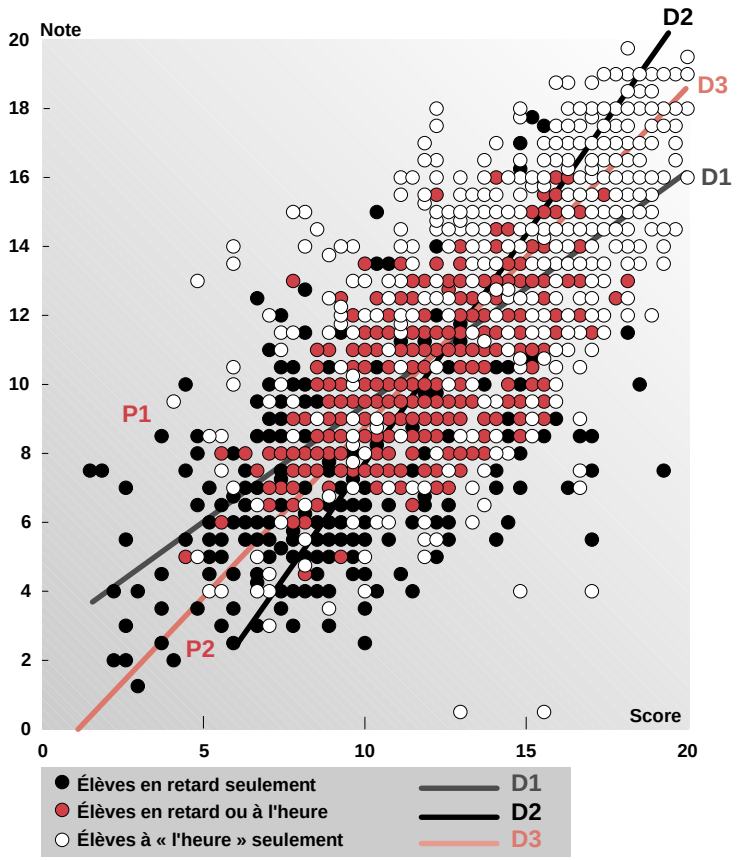
Opinions des élèves sur la notation

Grâce au questionnaire « Vie scolaire », on connaît l'opinion des élèves de troisième générale sur l'évaluation des résultats (notes, bulletins). On leur demandait en effet, à quoi elle sert essentiellement. Parmi les réponses proposées (qui ne s'excluaient pas les unes des autres), c'est l'information aux familles qui est le plus souvent avancée (84 % des élèves ont coché cette réponse). Les élèves sont aussi assez nombreux à considérer la notation comme une aide pour mesurer leur progrès (quatre élèves sur cinq). En revanche, seulement trois sur cinq environ pensent que cela leur permet d'avoir une meilleure vision de leurs atouts. La proportion de ceux à qui elle sert à se situer par rapport aux autres élèves, est la même. Moins de 55 % attribuent aux notes un intérêt pour l'information du professeur et à peine deux sur cinq disent qu'elles sont censées leur donner envie de travailler. Quant à leur objectif à long terme, ils sont plus nombreux à croire qu'elle va leur permettre de poursuivre leurs études (68 %) qu'à leur donner un sens pour un choix professionnel (57 % tout de même).

La méthode du résidu

On a mis en évidence surévaluation et sous-évaluation en faisant la différence entre score et note. Mais, si une note est comprise entre 0 et 20, un score est habituellement estimé en %. C'est pourquoi les scores ont été divisés par 5. De plus, on a été obligé de caler les moyennes sur 10. Postuler qu'un point de note vaut un point de score est donc peut-être arbitraire. D'autre part, on s'aperçoit vite que, par construction, un élève ayant un score très élevé (proche de 20) sera presque à coup sûr donné comme sous-noté ; en tous cas, il ne peut pas être surnoté. Inversement les élèves dont les résultats aux protocoles d'évaluation de 1995 sont faibles, sont donnés comme surnotés : il pourrait donc être plus approprié de raisonner à score constant ou à note constante.

GRAPHIQUE 3 – Note de contrôle continu en fonction du score, niveau Élève (mathématiques) et selon le retard scolaire



Le graphique 3 propose l'exemple du croisement entre score et note de contrôle continu en mathématiques ; on a isolé les élèves « à l'heure » et les élèves en retard. La méthode de régression linéaire permet de déterminer la droite D1 passant globalement au plus près de tous les points représentant les élèves, quand on minimise les écarts verticaux entre les points et la droite. Cette droite représente la meilleure approximation linéaire de la note en fonction du score. Son équation dans l'exemple choisi est
$N = 0,69S + 2,40$ (avec N = note, S = score).

Soit un élève i dont la note est N_i et le score S_i. D'après le modèle, sa note « prédite » est donc $N_{pi} = 0,69S_i + 2,40$. Si cette note théorique est supérieure à sa note observée N_i, il est sous-noté. En cas contraire, il est surnoté. Graphiquement, les élèves au-dessus de D1 sont surnotés ; en dessous, ils sont sous-notés. La différence $N_i - N_{pi}$, appelée « résidu » du modèle, mesure la surnotation ou la sous-notation. Si le résidu moyen d'une population est positif, cette population est surnotée. Par exemple, les élèves en retard sont sous-notés (résidu = - 0,48), ceux « à l'heure » surnotés (résidu = 0,31).

La méthode du résidu (suite)

Cette méthode a le défaut de n'être pas symétrique. La solution obtenue est différente si on exprime S en fonction de N et non N en fonction de S : il s'agit en effet de minimiser les écarts horizontaux (et non plus les écarts verticaux) entre les points et la droite idéale. Il suffit d'ailleurs de tourner le graphique d'un quart de tour vers la gauche pour se rendre compte que D1 n'est plus satisfaisante.

Il faut donc déterminer une droite D2 qui donne la meilleure approximation linéaire de S en fonction de N. Son équation est S = 0,76N + 4,04. On obtient en comparant score observé et score prédit par la note, un nouveau résidu.

On calcule que les élèves en retard ont un score inférieur à ce que leur note laisse attendre : ils sont « sous-scorés » (résidu = - 0,52) ; et que les élèves « à l'heure » sont « surscorés » (résidu = 0,31).

D'où ce paradoxe : la population des élèves en retard est sous-notée par rapport au score et sous-scorée par rapport à la note, c'est-à-dire sous-évaluée dans les deux types d'évaluation.

La comparaison entre l'élève P1 et l'élève P2 permet de mieux comprendre ce phénomène. L'élève P1 est au-dessus de D1 et à gauche de D2 : il est sous-noté et surscoré, il ne pose donc pas de problème. L'élève P2 est paradoxal : étant au-dessous de D1 et à gauche de D2, il est sous-noté et sous-scoré. Les conclusions que donnent les deux résidus sont contradictoires. On aboutirait au même problème si l'on raisonnait à score constant ou note constante en utilisant les techniques d'analyse de la covariance (GLM). On comprendra alors les résultats concernant l'influence du retard scolaire, sachant que 64 % des élèves ayant deux résidus négatifs sont en retard (et seulement 13 % parmi ceux qui ont deux résidus positifs).

Aussi est-il préférable d'utiliser la méthode par différence entre score et note, qui revient, graphiquement, à repérer la position des points par rapport à une nouvelle droite D3.

On rappelle qu'il s'agit de « corriger » le score et la note (en les calant sur un score moyen de 10 et une note moyenne de 10), d'en faire la différence, et de considérer le signe de cette différence : si la différence (score corrigé - note corrigée) est négative, l'élève est surnoté ; si elle est positive, il est sous-noté.

Autrement dit, dans l'exemple illustré par le graphique 3 :
– un élève est surnoté si (S -2,3) - (N - 0,9) < 0 ;
– un élève est sous-noté si (S -2,3) - (N - 0,9) > 0.
cela revient à dire que l'équation de la droite D3 est :
S - 2,3 = N - 0,9 soit S = N + 1,4.

Cette droite est en situation intermédiaire entre D1 et D2. Les individus « paradoxaux » proches de D3 pourront être considérés comme n'étant ni sous-évalués, ni surévalués pour les deux notations. Cette approche est donc plus satisfaisante puisqu'elle préserve la symétrie.

L'observation du graphique montre que les élèves en retard sont situés principalement dans le quart sud-ouest, mais qu'ils sont aussi surreprésentés dans le quart sud-est et le quart nord-ouest. Le tableau 6 confirme que les élèves âgés ont tendance à rater les deux évaluations (66,6 % d'entre eux se situent dans la mauvaise moitié selon le score et selon la note).

La méthode du résidu (suite)

TABLEAU 6 – Croisement selon le score et la note, et selon le retard scolaire

Élèves en retard (%)			
Score/note	Mauvais	Bons	**Total**
Mauvais	66,6	9,0	**75,6**
Bons	10,6	13,8	**24,4**
Total	**77,3**	**22,7**	**100,0**

Élèves « à l'heure » (%)			
Score/note	Mauvais	Bons	Total
Mauvais	27,7	10,8	**38,5**
Bons	10,7	50,8	**61,5**
Total	**38,4**	**61,6**	**100,0**

Ensemble (%)			
Score/note	Mauvais	Bons	**Total**
Mauvais	42,6	10,0	**52,7**
Bons	10,7	36,7	**47,3**
Total	53,3	46,7	**100,0**

LECTURE – en colonnes, répartition entre « mauvais » et « bons » élèves selon la note ; en lignes, selon le score.

Les proportions d'entre eux qui ne réussissent qu'une des deux évaluations sont les mêmes que pour les élèves « à l'heure » (environ 10 % dans tous les cas) mais ces proportions n'ont pas le même poids. Ainsi, un élève âgé qui a réussi le contrôle continu a 13,8/24,4 = 56,5 chances sur 100 de réussir les protocoles d'évaluation, tandis qu'un élève « à l'heure » a dans ce cas 50,8/61,5 = 82,6 chances. Ces données sont équivalentes si on considère la situation inverse (chances de réussite au contrôle continu pour les élèves ayant réussi les protocoles).

Réussir une des évaluations ne signifie pas, pour un élève âgé, qu'il va réussir l'autre ; en tous cas, le lien est moins fort que pour un élève « à l'heure ». De fait, les deux évaluations sont moins bien corrélées (R^2 de 40,2 % entre score et note pour les élèves en retard, contre 47,2 % pour ceux « à l'heure »).

Quel est le rapport entre ce phénomène et celui que nous avons déjà mis en évidence ? Supposons que la relation entre score et note soit meilleure pour les élèves âgés, que tous ceux qui aient réussi l'une des évaluations aient réussi l'autre. Dans ce cas, on aurait toujours la plus grande proportion de la population dans le quart sud-ouest, et les deux résidus moyens seraient sans doute encore négatifs. En revanche, les considérations qui ont été développées dans le texte sur les résultats à « score constant » seraient fortement modifiées. En effet, nous avons observé que si l'on fixe l'une des évaluations à une valeur donnée, on a systématiquement sur l'autre un écart en faveur des élèves « à l'heure » (voir tableau 5). Il est probable que, dans le cas fictif décrit plus haut, cet écart ne s'observerait plus que pour des valeurs basses de la notation fixée. Dans le cas des valeurs élevées, au contraire, on observerait sans doute un écart en sens inverse. On voit donc combien ces deux phénomènes s'entremêlent et jouent un rôle complémentaire l'un par rapport à l'autre.

La notation par les professeurs de lycée :

variations selon les disciplines et les situations

➔ *Les professeurs de lycée évaluent de manière différente les productions des élèves selon les circonstances de la correction. La spécificité des contenus de l'enseignement dispensé intervient dans leurs façons de faire, si bien que cette dimension, liée à la nature des connaissances qu'il s'agit d'évaluer, paraît plus intéressante que l'appartenance des professeurs à telle ou telle catégorie statistiquement repérable. Cette affirmation est étayée par deux enquêtes, dont l'une concerne un corpus de notes (traces quantifiées de l'évaluation) et l'autre un corpus d'appréciations et d'annotations inscrites sur des copies (traces textuelles de l'évaluation).*

Élisabeth CHATEL
IDHE-CNRS
ENS Cachan

On sait depuis longtemps que les notes attribuées par les professeurs aux travaux des élèves, bien qu'exprimées par des chiffres, n'ont pas les qualités attendues d'une mesure cardinale. Il convient alors de s'interroger sur la nature de l'acte de notation, qui, à l'encontre de son apparence de quantification, semble se rebeller à assurer une mesure exacte et fiable.

L'évaluation est-elle uniquement l'application d'une règle ou d'un barème, ou surtout l'exercice d'un jugement dans une situation ? Ceux qui la réalisent sont en effet des acteurs socialement insérés, et à ce titre ils ne sont pas seulement des agents pris dans des déterminations qu'ils reproduiraient à leur insu. Le professeur correcteur, s'il est l'acteur principal dans l'affaire, n'est pas seul. Il doit tenir compte des autres et avec eux des enjeux de la situation. En mettant l'accent sur la négociation entre les personnes par la notion d'arrangement, on privilégie cependant une des fonctions de l'évaluation – sa fonction de sélection et de classement – et on risque de laisser de côté ce sur quoi ce classement se fonde. Or les qualités nouvelles acquises par les élèves durant leur scolarité, leurs connaissances, leur savoir-faire, leurs comportements à l'égard des connaissances à acquérir, etc., constituent aussi l'enjeu de l'évaluation. La notation en effet n'est pas seulement une sanction *a posteriori* permettant sélection et classement, elle est également un élément régulateur de la vie des classes et de l'effort des élèves pour apprendre *« quelque chose qui vaille »*[1] en s'ajustant aux attentes manifestées par les professeurs.

NOTE

1. Comme le dit fort justement J.-C.Forquin [1] de l'activité enseignante et de l'École plus généralement.

TABLEAU 1 – Relation entre note au bac dans une matière et la moyenne obtenue en classe dans la même matière

Modèle : note au baccalauréat = f (note en classe)

	Ensemble	Mathématiques	Philosophie	Histoire-géographie	SES	Langue vivante 1
Effet sur la note au bac d'un point de plus en classe	0,95	0,88	0,5	0,77	0,68	0,79
Part de variance expliquée	51,30 %	42,70 %	11,00 %	24,60 %	17 %	41 %
Constante	- 0,207	- 0,2	3,67	1,75	2,43	2

LECTURE – La note en mathématiques obtenue en classe « explique » 42,7 % de la variance de la note en mathématiques au baccalauréat. Cette dernière gagne 0,88 point sur 20 pour un point de plus en classe.

NOTATION DES PROFESSEURS : QUE DISENT LES NOTES ?

Les notes, en tant qu'expression numérique d'un jugement, forment un matériau qui peut être traité par des techniques quantitatives. Cependant elles ne sont pas de « bonnes mesures » puisqu'on peut repérer différentes manières de noter selon les disciplines et au sein d'une même discipline.

La notation varie selon les disciplines

Les notes ne correspondent pas de la même manière à l'idée d'une mesure : nous insisterons ici sur cette principale manifestation d'une différence entre les disciplines. L'analyse montre l'existence de spécificités disciplinaires de l'évaluation, révélant l'impact des contenus de connaissance sur l'évaluation.

Cette affirmation s'appuie sur les résultats d'une enquête menée sur les notes obtenues par des élèves de terminale B (future ES) durant l'année scolaire 1991-1992 et au baccalauréat 1992 (*cf.* « Méthodologie de l'enquête »). Mais des résultats analogues sont obtenus par Noizet et Caverni [2] lors d'une enquête du même type, portant sur l'académie d'Aix-Marseille en 1974, bien qu'ils ne les interprètent pas de cette façon. Les éléments sur les notes au baccalauréat 1995 rapportés dans le rapport de l'Inspection générale de 1997 [3], convergent avec nos observations[2] sur les différences entre les disciplines.

Selon les disciplines, les notes ne correspondent pas de manière identique aux caractéristiques d'une mesure. En effet si les notes étaient de bonnes mesures des compétences disciplinaires des élèves, les notes obtenues en classe devraient prédire celles du baccalauréat. Les élèves se préparent dans l'année à l'examen ; les professeurs, de leur côté, cherchent en général à les tester sur des épreuves qui anticipent et préparent celles de l'examen. Or la liaison que nous obtenons entre notes données en classe et au baccalauréat n'est pas très bonne. Certes les élèves qui obtiennent la moyenne en classe sont reçus au baccalauréat pour 79,7 % d'entre eux au premier groupe d'épreuves et 95 % si on prend les deux groupes. Mais il y a aussi 26 % des élèves qui n'ont pas cette moyenne et qui sont néanmoins reçus au premier groupe d'épreuves, et ce pourcentage est porté à 53 % si on ajoute les résultats des deux groupes d'épreuves.

Si la liaison entre moyenne en classe dans les cinq matières retenues et moyenne au baccalauréat dans les même matières est assez forte, elle est beaucoup plus faible matière par matière. Les différences entre les matières attirent principalement notre attention ici (tableau 1). Elles nous conduisent à opposer deux sortes de matières selon qu'elles se prêtent plus ou moins bien à la mesure. On aurait d'un côté les mathématiques et la langue vivante. Les notes dans ces matières sont plus dispersées (en classe comme au baccalauréat), mais surtout leur fiabilité paraît plus grande : la liaison entre note obtenue en classe et note au baccalauréat est plus forte. Inversement, la philosophie et les sciences économiques et sociales sont des disciplines où les notes correspondent assez peu aux caractéristiques d'une mesure, l'histoire-géographie se situant entre ces deux pôles.

Un bref regard sur les caractéristiques de ces matières permet de comprendre ce phénomène. Les mathématiques et la langue vivante 1 se prêtent assez volontiers à des épreuves où les élèves sont jugés sur l'exactitude de leurs connaissances de façon très délimitée. On peut évaluer leur capacité à connaître, mobiliser et appliquer correctement une règle ou un algorithme. En conséquence, les évaluations sont construites comme une série de petites questions ou exercices successifs dont la correction appelle un diagnostic vrai-faux, exact-inexact et ne laisse pas une grande marge d'interprétation au correcteur. À l'inverse, en philosophie et en sciences économiques

NOTE

2. Cependant, nos données, portant sur le baccalauréat économique et social, n'informent pas sur les caractéristiques des notes dans les disciplines scientifiques, technologiques, artistiques ou sportives.

et sociales, on demande aux élèves un effort de problématisation autonome, et une rédaction de synthèse, qui s'appuient normalement sur les connaissances et démarches acquises en classe. Même lorsqu'on cherche, comme on le fait aujourd'hui en sciences économiques et sociales, à décomposer l'épreuve par des questions préalables préparant la synthèse ultérieure[3], l'activité intellectuelle requise reste difficile à délimiter. Réciproquement le correcteur détient de son côté une grande marge d'interprétation. Il juge une prestation de façon synthétique plus qu'il n'applique un barème.

Les sciences économiques et sociales, sur lesquelles nous avons plus systématiquement fait porter nos analyses, sont donc caractérisées par le fait que les notes n'y sont pas de bonnes mesures. Mais rien ne permet de mettre en cause les comportements de notation des professeurs de ces matières, indépendamment de la nature des connaissances qu'ils évaluent.

Les caractéristiques statistiquement repérables des professeurs éclairent bien peu leurs façons de noter

On peut faire l'hypothèse que les façons d'évaluer sont assez différentes selon les professeurs. Pierre Merle [4] distingue ainsi les professeurs qui croient à l'exactitude des notes et sont, de ce fait, peu enclins à les négocier, à l'inverse de ceux qui, en doutant, acceptent plus volontiers la souplesse, et se montrent moins sévères. Cependant cette distinction est délicate à préciser par des analyses quantitatives car les notes, par définition, traduisent tout à la fois la façon de noter, c'est-à-dire de juger et de quantifier le jugement de celui qui évalue, et le niveau de réussite de la prestation évaluée. Il est difficile de séparer ces deux composantes dans l'analyse. Pour essayer d'identifier néanmoins un éventuel « effet » du professeur dans la façon de noter, il faut séparer ce qui dans la variance des notes renvoie à des différences dans la qualité du travail évalué, et ce qui provient de différences dans les façons de noter. Nous avons tenté de le faire en contrôlant le niveau de prestation des élèves par leur moyenne générale. Cette moyenne fournit un jugement quantifié de l'ensemble de l'équipe pédagogique et constitue ainsi un indicateur du niveau scolaire de l'élève, certes critiquable puisqu'il confond les matières, mais effaçant quelque peu les particularités de chacun des professeurs dans la façon de noter[4].

En procédant de la sorte, nous voyons tout d'abord que la relation entre les moyennes des différentes matières et la moyenne générale est forte. Cette relation n'a rien de surprenant, puisque la moyenne générale contient la moyenne de chaque matière. Cependant elle plaide pour considérer qu'existe une certaine convergence à la fois de la réussite des élèves dans plusieurs matières et des jugements des professeurs de ces diverses disciplines. Se manifeste aussi un effet conjoint de la spécificité de la classe et de la personne du professeur sur les notes (tableau 2)[5]. Cet effet est assez important ; il est variable selon les disciplines, allant de 9 % de la variance des notes en mathématiques à 21 % en philosophie (et 11,5 % en

THÈME

NOTES

3. Ce sont les caractéristiques de la nouvelle épreuve mise en place pour le baccalauréat à partir de 1995.

4. On suppose donc que la relation est forte entre la réussite dans une matière et la réussite scolaire en général. En contrôlant le niveau des élèves par la moyenne générale, on recherche un « effet net » de l'appartenance à tel ou tel groupe classe. Le calcul des « effets bruts » sur les notes donne la même hiérarchisation des matières (sur la différence entre « effet brut et net », voir [10]).

5. On travaille donc sur la variance des notes dans une matière à moyenne générale donnée, et on prend pour variable contextuelle du modèle un identifiant du groupe classe-professeur. Cette variable est dite muette puisqu'on ne sait en qualifier les modalités autrement qu'en distinguant les uns des autres les résultats de chacune des 126 classes différentes de notre échantillon.

TABLEAU 2 – Effet du groupe classe-professeur sur la variabilité des notes

Modèle : note en classe par matière = f (moyenne générale, spécification professeur-classe, caractéristiques sociodémographiques des élèves)

	Mathématiques	Philosophie	Histoire-géographie	Langue vivante 1	SES
Part de variance dont le modèle rend compte	61 %	62,80 %	64,50 %	57 %	69 %
dont part attribuable à la moyenne générale	50,90 %	40,63 %	51,10 %	41,80 %	57,20 %
dont part attribuable à la spécificité classe-professeur	9 %	21 %	11,50 %	12,50 %	11,50 %
dont part des autres variables	1,10 %	1,20 %	0,40 %	2,70 %	0,50 %
Constante	6,19	1,93	2,5	0,73	0,99

sciences économiques et sociales). Il montre qu'il y a des différences significatives dans les notes accordées par différents professeurs, mais rappelons-le, ces différences tiennent autant aux façons de noter qu'aux éventuelles différences de réussite d'une classe dans une matière spécifique, la faisant dévier de la moyenne générale. Toujours est-il que l'effet important de la variable classe-professeur renvoie probablement à l'activité du professeur, en tant qu'enseignant et évaluateur, et qu'elle s'inscrit incontestablement dans le contexte d'une discipline d'enseignement.

Nous avons voulu ensuite mieux cerner les facteurs de la variabilité de la notation des professeurs pour les sciences économiques et sociales, matière pour laquelle nous pouvions relier les notes des élèves en classe aux caractéristiques sociodémographiques et professionnelles des professeurs.

Le modèle reproduit dans le tableau 3 nous permet de voir l'effet de certaines caractéristiques du professeur évaluateur sur la note moyenne annuelle obtenue en sciences économiques.

Si ce modèle rend compte d'une part importante de la variance des notes, cela tient à la forte corrélation entre la moyenne générale et la note en SES. 57,2 % de la variance des notes en SES est attribuable à l'effet de la moyenne générale. L'effet global des caractéristiques sociodémographiques et professionnelles des professeurs et des caractéristiques sociodémographiques des élèves sur les notes est, quant à lui, très faible (1,8 % de part de variance)[6]. Nous avons fait intervenir la moyenne générale dans le modèle pour contrôler l'effet éventuel du niveau des élèves sur leur note en SES. Ainsi les effets des caractéristiques sociodémographiques des élèves et professeurs que nous observons s'exercent pour une moyenne générale donnée des élèves.

L'effet des caractéristiques sociodémographiques des élèves sur leurs notes de sciences économiques et sociales est peu important ; la seule variable qui joue significativement est le sexe. Bien que faible, cet effet mérite d'être souligné car il va à l'encontre de ce que l'on sait de la scolarité des jeunes filles. Elles ont

TABLEAU 3 – Notation selon les caractéristiques des professeurs

Modèle : note de SES en classe = f (moyenne générale en classe, caractéristiques sociodémographiques des professeurs et des élèves)

Part de variance dont le modèle rend compte	59 %
dont part de variance attribuable à la moyenne générale	57,20 %
dont part attribuable aux autres variables	1,80 %
Constante	1,05
Effet d'un point de plus de moyenne générale sur la note moyenne en SES	0,9

Effet des caractéristiques sociodémographiques des professeurs sur la note en SES

Variables de référence	Variables actives	Coefficients
Grades : maître auxiliaire	Agrégé	**0,2**
	Adjoint d'enseignement	**0,35**
	Certifié	ns
Sexes : homme	Femme	**0,17**
Académies : Versailles	Lille	**0,35**
	Montpellier	ns
	Paris	ns
	Rouen	**0,17**
Origines sociales : professions intermédiaires	Agriculteurs, commerçants, artisans	ns
	Cadres supérieurs	ns
	Employés	**- 0,24**
	Ouvriers	**- 0,23**
	Autres	ns
Âges : plus de 48 ans	moins de 30 ans	*- 0,23*
	de 31 à 40 ans	**- 0,27**
	de 41 à 47 ans	*- 0,15*
Formations initiales : économistes	Non-économistes	**- 0,2**
Niveaux de diplôme les plus élevés : maîtrise	En dessous de la maîtrise	ns

Effet des caractéristiques sociodémographiques des élèves sur la note en SES
(Seul le sexe a un effet significatif)

Sexes : garçons	Filles	**- 0,27**

LECTURE – Seuls les résultats significatifs ont été reportés. En gras : significativité au seuil de 1 % ; en italiques : au seuil de 5 % ; en romain : au seuil de 10 % ; ns : non significatif. Les coefficients se lisent ainsi : un élève noté par un agrégé obtient 0,2 point de plus qu'un élève noté par un maître auxiliaire (catégorie de référence), toutes les autres variables étant supposées égales par ailleurs.

NOTE

6. Nous avons calculé l'effet des seules caractéristiques sociodémographiques des professeurs de SES sur les notes en SES par un modèle, dont les résultats ne sont pas reproduits ici. Ce modèle prenait comme seules variables explicatives les caractéristiques sociodémographiques et professionnelles des professeurs et aboutissait à ce que 3,4 % de la variance des notes en classe soit attribuable à ces variables, lorsqu'on ne tenait pas compte de l'ancrage académique, et 4 % lorsqu'on en tenait compte. De la même façon, les caractéristiques sociodémographiques des élèves, prises comme seules variables explicatives des notes, ne rendent compte que de 3,45 % de la variance de celles-ci. Prises ensemble dans le modèle du tableau 3, ces variables ne rendent plus compte que de 1,8 % de la variance des notes, une partie de leur effet étant absorbé dans celui de la moyenne générale.

globalement de meilleurs résultats scolaires que les garçons, mais en SES elles obtiennent 0,27 point de moins qu'eux, toutes les autres variables étant égales par ailleurs. or divers travaux ont montré qu'elles maîtrisent mieux que ces derniers leur *« métier d'élève »* (Felouzis, 1993) et bénéficient de ce fait du regard bienveillant des professeurs. On ne peut donc pas interpréter leurs moins bons résultats en sciences économiques et sociales comme étant l'effet d'une sévérité particulière des professeurs à leur égard. Ils sont alors à attribuer à la qualité moindre de leurs prestations dans cette matière. Le même phénomène se produit en histoire-géographie, matière parente des SES en tant que sciences sociales, posant des questions de portée politique. Ceci interroge sur le lien entre la réussite sexuée des apprentissages et la nature des connaissances à acquérir. Mais revenons à notre question qui est la notation des professeurs.

Bien que ne rendant compte que d'une très faible part de la variance des notes, apparaissent cependant des façons significativement différentes de noter selon les caractéristiques des professeurs.

L'effet du grade est confus, puisqu'on ne voit pas de différence significative entre certifiés, catégorie numériquement la plus importante, et maîtres auxiliaires, mais une légère différence entre ceux-ci et les agrégés comme les adjoints d'enseignement, qui notent un peu plus largement, toutes autres choses étant égales par ailleurs. Chacune des ces deux dernières catégories est au demeurant peu nombreuse dans l'échantillon[7]. Le fait d'avoir ou non une maîtrise ne semble pas exercer d'influence sur la notation, tandis que, dans une profession dominée par des enseignants formés initialement en économie, les non-économistes notent plus sévèrement.

L'âge joue un rôle plus manifeste et plus facile à comprendre sur les notes. Les professeurs les plus jeunes (moins de 40 ans) notent plus sévèrement que leurs aînés : on peut supposer qu'il y a là un effet d'une moindre expérience. Le sexe joue aussi un peu ; les femmes notent un peu plus large que leurs collègues masculins, mais le coefficient n'est pas très élevé.

Il y a aussi un effet de l'origine sociale des professeurs sur leurs façons de noter. Les professeurs d'origine défavorisée (enfants d'employés et d'ouvriers) notent plutôt plus sévèrement que leurs collègues.

Remarquons enfin qu'existent des différences académiques dans les façons de noter. On note moins sévèrement en sciences économiques et sociales à Lille et à Rouen que dans les autres académies de l'échantillon, mais il se trouve aussi que dans ces deux académies, les résultats au baccalauréat 1992 et les notes sont généralement plus faibles que dans les autres, et ce dans toutes les matières[8].

Ainsi les variables caractérisant les formations des professeurs ne donnent pas d'indications faciles à interpréter tandis que l'effet de l'âge, du sexe, de l'académie d'exercice et de l'origine sociale, sans être très importants, jouent d'une façon qui est cohérente avec l'intuition ou avec d'autres résultats. Les caractéristiques sociodémographiques et professionnelles des professeurs exercent un effet sur leurs façons de noter ; il n'en reste pas moins que cet effet est globalement assez faible. Certes il conforte l'idée selon laquelle la notation n'est pas strictement une mesure, mais cela ne nous permet pas de faire un pas vraiment important pour comprendre ce qu'elle est.

C'est pourquoi il est intéressant de rechercher comment prend forme le jugement évaluatif sur les élèves au moment de la correction de leurs travaux.

Les critères de correction de copies en classe et au baccalauréat

Comme les notes, les corrections de copies constituent des traces du jugement porté sur les élèves. De plus, les corrections, étant exprimées par des énoncés, donnent à voir la façon dont les professeurs expriment et étayent leur jugement sur les élèves au moment même où ils l'élaborent. C'est un matériau précieux pour voir émerger en actes les critères divers mobilisés dans l'évaluation. Ces analyses montrent l'importance accordée par les évaluateurs à la dimension spécifique des connaissances de leur domaine et à la méthode de l'exercice. Leurs critères de correction ont un aspect technique et disciplinaire : on juge un travail dans un domaine plus que des qualités supposées d'une personne (*cf.* encadré p. 56).

Des caractéristiques assez générales relatives à la correction en sciences économiques et sociales ressortent de ces analyses ainsi que des différences dans

NOTES

7. Respectivement 19 agrégés ou bi-admissibles et 14 adjoints d'enseignement sur 126 professeurs.

8. Ce résultat prend un certain relief car il est postérieur à une expérience de multicorrection de copies de baccalauréat menée en 1984 à l'initiative du recteur de Lille. Celui-ci faisait l'hypothèse d'une sévérité accrue à Lille par rapport aux autres académies. Le test opéré en sciences économiques et sociales [4] avait renforcé cette supposition. Peut-être faut-il expliquer nos résultats par ses effets ultérieurs, conduisant les professeurs de sciences économiques et sociales de Lille et Rouen, convaincus d'être trop sévères, à se montrer plus larges

Matériaux de l'enquête et analyse lexicale

Le matériau est constitué des trois corpus d'énoncés correctifs inscrits sur des dissertations de sciences économiques et sociales. Le plus important, 554 copies, provient du baccalauréat 1992 à Montpellier ; ces copies ont été recueillies lors de la collecte des notes, durant l'enquête précédente. Elles appartenaient à 25 jurys différents, repérés par leur numéro, mais les correcteurs sont inconnus. On connaît au contraire la note et le sexe des candidats. Un deuxième corpus est formé par les corrections portées par 34 correcteurs, dont on connaît les caractéristiques sociodémographiques, corrigeant deux copies identiques dans une expérience de multicorrection qui a eu lieu à Lille en juin 1993. Le dernier corpus est fait de 58 copies effectuées dans 5 classes différentes dans les académies de Lille, Paris et Créteil. Ces devoirs ont été effectués à notre demande sur les mêmes sujets qu'à Montpellier, durant l'année 1993. On connaît les caractéristiques des élèves et professeurs, mais cet échantillon n'est pas représentatif.

Les énoncés correctifs, appréciations en tête de copie et annotations en marge, dactylographiés ont été traités par le logiciel d'analyse lexicale ALCESTE conçu par Max Reinert [6] [7]. ALCESTE permet principalement de rapprocher des énoncés précédemment découpés en unités de même longueur (propositions de 14 à 16 mots, phrases, paragraphes). Les classes stables d'énoncés sont formées par la proximité de leur lexique. Par exemple, si l'on trouve l'appréciation suivante : « Plan contestable, documents peu utilisés, cependant des remarques pertinentes », la présence des mots « plan », « document », « utilisé ou utilisation » rapprochera cet énoncé de celui-ci : « Des connaissances, une assez bonne utilisation des documents mais le plan est mal articulé ». Les « classes lexicales » indiquent que les sujets énonciateurs se réfèrent au même univers de significations en employant le même lexique. L'interprétation consiste ensuite à délimiter ces univers grâce aux vocables et par les propositions les plus représentatives de leur utilisation que nous livre le logiciel. Il est ensuite possible d'examiner les relations statistiques entre ces classes lexicales et certaines variables caractéristiques des copies telles que des données sur le candidat, le correcteur, la copie ou la situation d'évaluation.

les façons de corriger. Ces caractéristiques se retrouvent dans nos trois corpus, mais elles sont particulièrement mises en évidence dans le corpus de correction des copies de baccalauréat sur lequel, pour cette raison, nous appuierons notre première démonstration.

Le vocabulaire de la correction

Les mots que l'on retrouve le plus fréquemment dans ces énoncés sont communs aux trois corpus. La négation « pas » est le vocable le plus fréquent : corriger consiste à redresser le propos de l'autre. Cependant le vocable « sujet » vient juste après par ordre de fréquence – il s'agit du sujet de la dissertation. Le correcteur attend donc du candidat qu'il traite le sujet ; l'évaluation n'est pas sans contenu. Si on ajoute « sujet » et « hors sujet », on obtient l'expression qui a le plus d'occurrences. Cependant, hormis la forme lexicale « sujet », les vocables les plus fréquents sont évaluatifs : « ne », « pas », « mal », « non », « bien », « oui », « A », « peu », « assez », « trop ». L'évaluation négative l'emporte cependant sur les vocables plus positifs ou moins tranchés, confirmant que la norme est celle de l'excellence. Le professeurs disent les raisons de ne pas mettre une bonne note, plus qu'ils ne justifient les points accordés.

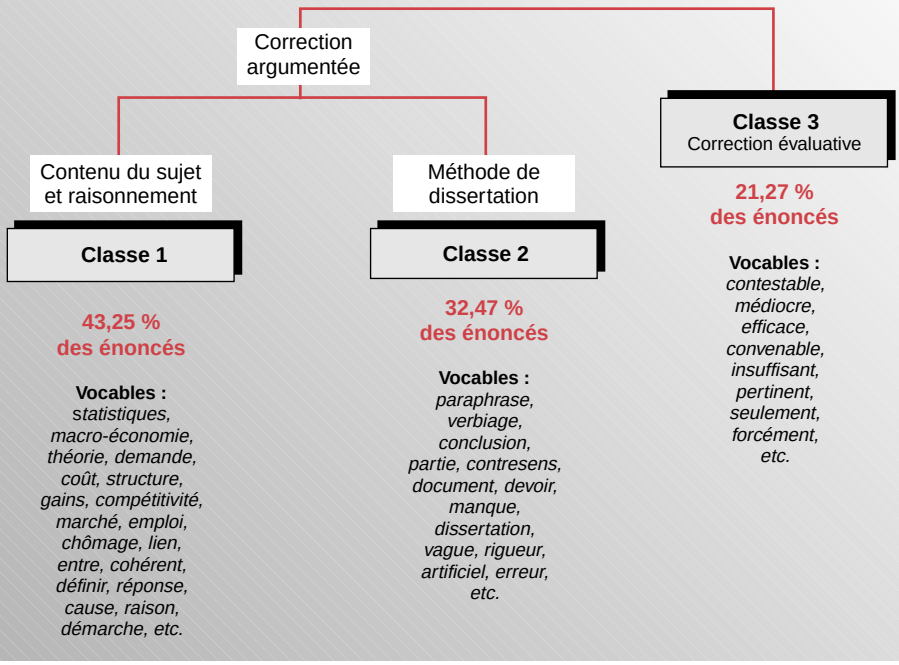
Les critères de correction

Les classes lexicales formées par l'analyse font apparaître le caractère très conventionnalisé des énoncés correctifs. Les classes formées par ALCESTE (*cf.* « Matériau de l'enquête et analyse lexicale ») sont très nettes pour les énoncés correctifs des copies de baccalauréat, plus encore pour les annotations que les appréciations. Sur les autres corpus, elles sont un peu moins nettes, cependant les interprétations convergent.

L'ensemble des énoncés correctifs des copies de baccalauréat découpés en propositions de 14 à 16 mots font émerger trois classes lexicales correspondant à trois sortes d'argumentations qu'on trouve sur les copies. Le graphique 1 en offre une représentation.

Une première catégorie regroupe des énoncés qui expriment des exigences relatives à la nécessité de traiter le sujet dans son contenu disciplinaire spéci-

GRAPHIQUE 1 – Arbre des registres sémantiques de la correction au baccalauréat (Montpellier, 1992)



fique, et de construire un raisonnement. Elle classe 43,25 % des énoncés classés ; or, 78 % des énoncés de ce corpus ayant été classés (ce qui est un bon résultat pour ALCESTE), cela veut dire que la majeure partie des énoncés correctifs évoquent les significations de cette première classe. On y trouve ainsi du vocabulaire appartenant au domaine économique et social, et des mots présents dans l'intitulé du sujet. Il y a aussi des vocables qui évoquent la nécessité de raisonner.

Voici un énoncé représentatif de cette classe : *« Confusion semble-t-il. Vous ne faites pas le lien entre compétitivité et transformations du marché du travail »*.

La deuxième classe lexicale ressortit, plus que la première, du registre de la méthode de la dissertation. Elle regroupe une proportion un peu moindre des énoncés classés (32,47 %). Elle exprime l'exigence d'une dissertation construite, utilisant les documents associés à l'épreuve et écrite correctement.

Cela donne par exemple un énoncé représentatif comme celui-ci : *« La construction formelle de la dissertation est respectée mais l'apport de connaissances et l'étude des documents sont insuffisants, le plan n'est pas respecté »*.

La troisième classe lexicale se distingue des deux autres en ce qu'elle renvoie quasi exclusivement au registre évaluatif. Elle ne regroupe que 21,27 % des énoncés classés.

Une appréciation-type de ce registre peut être la suivante : *« Ensemble insuffisant et maladroit. Banalités et lieux communs, des points insuffisamment développés »*.

Registre du contenu joint à celui du raisonnement ; registre plus spécifique à la méthode et à la forme dissertative ; registre évaluatif : avec des variantes et une intensité variable, on retrouve ces trois dimensions de la correction dans nos trois corpus. On peut les rapprocher des critères de correction mentionnés dans les comptes-rendus de commissions d'entente et d'harmonisation au baccalauréat [9] ; ils sont sensiblement les mêmes. Mais il faut garder à l'esprit qu'il s'agit de registres employés simultanément. Le correcteur étaye le plus souvent sa correction en combinant divers registres lexicaux. Cependant il les agence diversement selon les situations, comme nous allons le voir, ce qui apporte la preuve qu'en pratique la note est accordée en fonction d'une appréciation synthétique portant sur plusieurs dimensions, elle n'est pas l'addition de ces éléments.

NOTE

9. Ou dans d'autres travaux sur l'évaluation en sciences humaines et sociales, voir par exemple le travail sur l'évaluation de l'excellence scolaire par la Direction de l'évaluation et de la prospective, ministère de l'Éducation nationale, de la Recherche et de la Technologie, en 1996 [8].

Quelles différences entre les façons de corriger ?

L'étude de ces différences s'appuie sur des comparaisons que l'on peut mettre en évidence grâce à nos différents corpus de copies. Mais elles se dessinent déjà dans les distinctions entre les classes lexicales du corpus principal, celui de la correction au baccalauréat à Montpellier. Une première différence s'exprime par l'opposition entre les classes 1 et 2 (axe 1 du graphique 2) et la classe 3, distinguant ainsi une correction argumentée, soit au niveau de son contenu soit au niveau de la méthode de l'exercice, à une correction seulement évaluative et moins approfondie, dans laquelle on n'argumente pas sur ce que le candidat dit ou fait – on se contente de l'apprécier.

La deuxième différence (sur l'axe 2 du graphique) sépare une exigence relative à la construction du raisonnement portant sur le contenu du sujet à traiter (classe 1) de celles relatives à la méthode de l'exercice (faire un plan, utiliser les documents) (classe 2).

Ces divers registres entrent en proportions variables dans les jugements effectués selon les circonstances de l'évaluation, c'est-à-dire selon qu'il s'agit de corriger au baccalauréat ou en classe. Leur usage varie selon le niveau des copies. Leurs agencements dépendent aussi des correcteurs. En effet des styles de correction propres aux correcteurs émergent de ces analyses lexicales, cependant leur stabilité n'est pas sûre.

➔ *Différences entre correction en classe et au baccalauréat*

En classe et au baccalauréat, les correcteurs utilisent la même batterie de critères, ce qui confirme la permanence des exigences entre ces deux corrections. Cependant le contexte de correction intervient dans la façon de corriger. Le correcteur quand il corrige n'oublie pas les enjeux spécifiques de la situation dans laquelle il se trouve.

Ainsi les énoncés véritables de correction au baccalauréat sont différents de la multicorrection. Les correcteurs sont plus prolixes lors de la multicorrection[10] (à Lille en 1993). Dans ce dernier cas il s'agit d'une correction fictive visant à dégager les diverses interprétations possibles de la question posée et de voir, entre correcteurs, les problèmes que la correction peut soulever. Le correcteur prépare une discussion avec ses collègues et non avec le candidat.

Dans la correction en classe, la longueur des énoncés correctifs moyens est beaucoup plus importante que dans les deux autres circonstances[11]. Les occurrences de l'emploi des pronoms personnels « vous »[12] et « tu » plus que le pronom « il » ou l'expression « le candidat », montrent clairement qu'en corrigeant le professeur s'adresse à un élève connu de lui. Il s'adresse à lui dans l'intention de corriger son travail, donc de signaler les écarts à ses attentes, et pas seulement de justifier une note.

Les appréciations et annotations sont longues, la classe lexicale purement évaluative disparaît tandis qu'apparaissent deux classes lexicales correspondant au registre de la méthode (construction du devoir et expression, connaissances et utilisation des documents) et que se maintient une classe relative au contenu même du sujet, la plus importante (52,4 % des énoncés classés). Ceci conduit à penser que la correction en classe, du moins dans ce petit échantillon qui malheureusement n'est pas représentatif, a une dimension « formative » assez marquée. En classe le jugement n'a pas le même caractère d'irréversibilité qu'à l'examen, la correction peut servir à signaler ce qui ne va pas. On en trouve pour preuve supplémentaire le fait que, dans les copies faites en classe, il n'y a pas de relation statistique entre le niveau des copies et les registres lexicaux mobilisés pour corriger, alors qu'au baccalauréat elle est nette, signe que la préoccupation du classement ordinal des copies est plus patente au baccalauréat qu'en classe.

➔ *Correction et niveau de la copie*

L'usage des divers critères de corrections varie selon le niveau de la copie, très nettement en ce qui concerne les corrections de copies de baccalauréat (*cf.* graphique 1)[13].

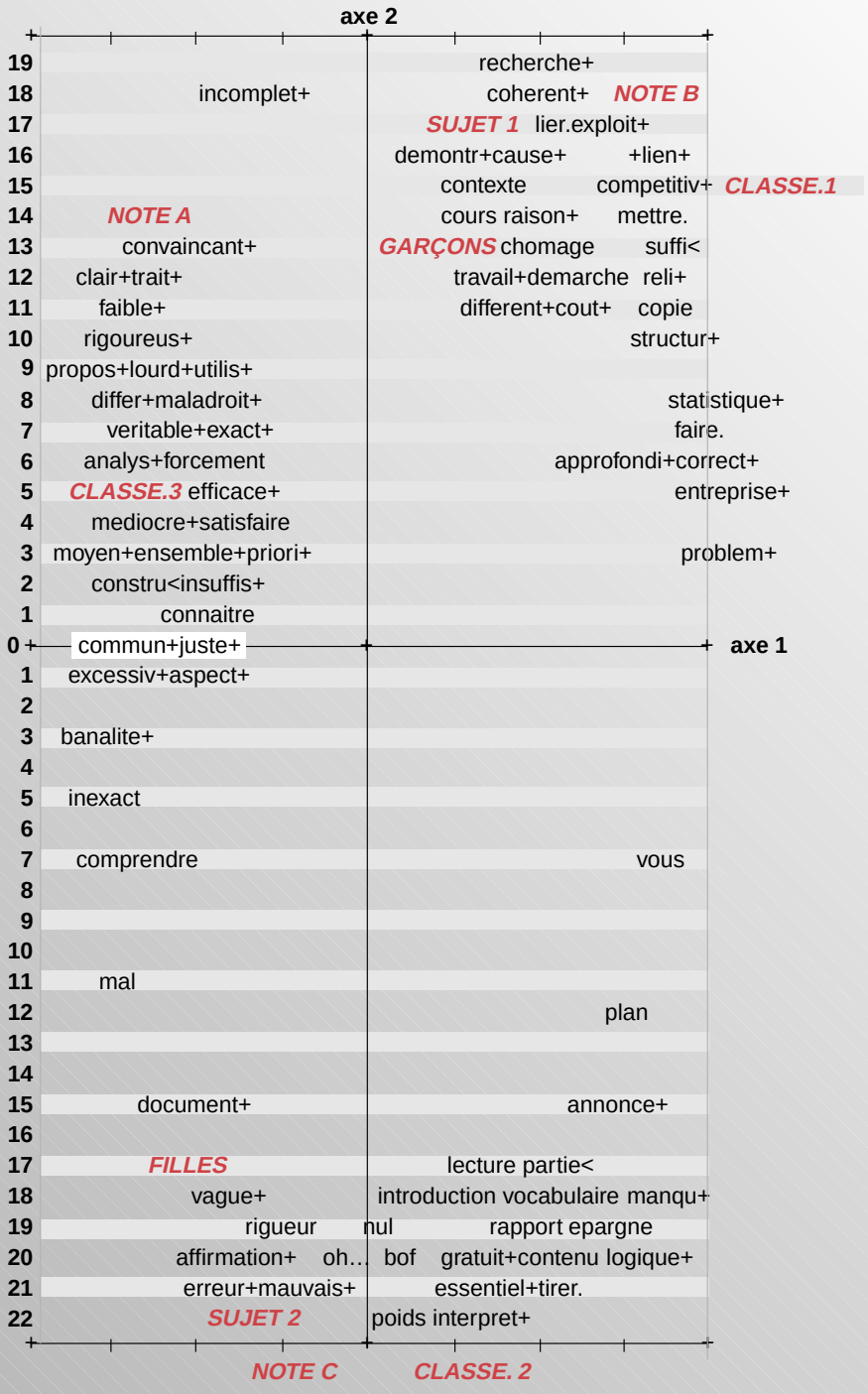
Alors que les copies les meilleures (12 et plus, note A sur le graphique) appellent des commentaires surtout évaluatifs, bien qu'assez vagues, et parfois rela-

NOTES

10. Leurs énoncés sont 3,18 fois plus longs en nombre de mots qu'au baccalauréat.
11. Les énoncés correctifs sont en moyenne 7,78 fois plus longs que dans la correction au baccalauréat, la longueur étant mesurée en nombre de mots.
12. 255 fois pour 58 copies en classe alors qu'il apparaît 153 fois dans les corrections de baccalauréat pour 554 copies.
13. Pour confirmer la liaison établie par l'analyse factorielle (projections des variables associées, inactives dans l'analyse, sur les classe lexicales), nous avons procédé aussi à l'inverse. Nous avons recherché les mots les plus significativement associés, en terme de X^2, à chaque niveau de notes. L'interprétation des vocables obtenus par ces tris est convergente avec celle des classes lexicales formées « spontanément » par ALCESTE (par rapprochement des énoncés ayant des vocables communs), au sens où ils évoquent les mêmes domaines de signification.

GRAPHIQUE 2 – Analyse factorielle du corpus des énoncés correctifs des copies du baccalauréat (Montpellier) – Sciences économiques et sociales

Projections des mots analysés sur le plan 1 2



Remarques – Les classes (1 ; 2 ; 3) sont des classes lexicales (voir texte) ;
– A, B, C sont les niveaux de note des élèves (> 12 ; de 8 à 11 ; < 8) ;
– la projection de la classe 3 sur le premier axe et de la classe 2 sur le deuxième sont particulièrement bonnes ;
– les vocables (dans leurs formes réduites) sont les variables actives de l'analyse ;
– le sexe est celui des candidats au baccalauréat.

tifs aux connaissances, les copies plus mauvaises (moins de 8, note C) sont jugées en exprimant des critères principalement méthodologiques et formels. Leurs insuffisances au niveau formel semblent interdire au correcteur d'argumenter au-delà. Dans les deux copies multicorrigées à Lille domine le registre de la méthode, ce qui est cohérent avec le fait qu'elles sont notées bas (4,74 et 5,74 sont les moyennes des notes qui leur ont été données par les 34 correcteurs). Les copies moyennes provoquent au contraire des énoncés relatifs au contenu du sujet et aux raisonnements à tenir pour le traiter. Elles sollicitent plus attentivement l'attention du correcteur sur leur contenu que les autres et leur imposent, semble-t-il, d'entrer dans le raisonnement de l'élève, d'évoquer l'objet de l'étude et la façon de l'aborder.

Pour les bonnes et les mauvaises copies, l'évaluation paraît relativement simple. Quand elles sont bonnes, on le dit, et on précise comment elles pourraient être meilleures. Quand elles ne le sont pas (moins de 8), on le dit aussi, le commentaire porte sur les insuffisances relatives à la méthode de l'exercice plus que sur son contenu en termes de connaissances. Cela indique, en creux, que les correcteurs exigent une dissertation dans une forme acceptable pour accorder au candidat la note autorisant un rattrapage possible à l'oral du deuxième groupe. On trouve plus souvent dans la correction de ces copies que dans les autres l'emploi du pronom personnel « vous », comme si elles provoquaient, par leurs insuffisances, une certaine implication relationnelle du correcteur.

Aux copies de niveau intermédiaire, les plus nombreuses (soit 50,9 % des copies de l'échantillon), est associé le vocabulaire du raisonnement et du contenu. À l'inverse des précédentes, on y emploie plus fréquemment qu'ailleurs le pronom « il », ce qui manifeste une sorte de distance du correcteur à l'égard du candidat qui s'exprime lorsque la correction est focalisée, plus spécifiquement, sur l'objet de connaissance dont il fallait traiter. L'évaluation de ces copies comporte des attendus plus élaborés, plus étayés en profondeur. Ce faisant, la correction des copies moyennes et mauvaises nous en dit plus que les meilleures sur les attentes des professeurs. La correction des plus mauvaises sur le registre de la méthode montre que celle-ci est conçue comme auxiliaire, elle est une condition nécessaire mais non suffisante de la réussite. Réussir suppose de traiter le sujet posé en construisant un raisonnement, attentes exprimées dans la correction des copies moyennes.

➔ *Styles de correction et correcteurs, une relation difficile à caractériser*

Peut-on établir une relation entre des « façons de corriger » repérables et des correcteurs identifiables ? Pour cela, il nous faut traiter les énoncés correctifs portés sur une même copie comme une entité unique et non les découper en propositions comme dans les analyses précédentes[14]. L'énoncé élémentaire est alors formé par l'appréciation complète d'un correcteur face au texte d'une copie. Les appréciations en tête de copie se prêtent mieux que les annotations en marge à ces analyses. Lorsque des classes lexicales sont obtenues à partir de ce mode de découpage des énoncés, nous les interprétons comme exprimant des « styles de correction » différents.

L'analyse ainsi effectuée des appréciation des copies de baccalauréat (Montpellier en 1992) fait apparaître 5 classes lexicales, mais trois d'entre elles classent si peu d'énoncés que nous n'en tiendrons pas compte. Deux autres catégories classent chacune environ la moitié des appréciations restantes[15]. On y retrouve présents les trois registres lexicaux que nous avons passés en revue précédemment. Ces deux classes sont donc toutes deux composites relativement à nos registres élémentaires, ce qui ne surprend pas. Elles se distinguent cependant l'une de l'autre par la place faite dans l'une, et non dans l'autre, à une référence précise et constante aux contenus même du sujet à traiter. Dans la première on trouve par exemple des appréciations comme celles-ci :

« Assez bon travail d'ensemble. Des connaissances et analyses parfois fines, utilisation assez satisfaisante des documents, l'ensemble demeure assez satisfaisant ».

« Ensemble peu convaincant. Le sujet n'est pas vraiment traité, connaissances insuffisantes ».

Dans la deuxième on trouve dans les énoncés des mots relatifs au contenu de la question à traiter (que nous soulignons) :

NOTES

14. Le logiciel ALCESTE n'est pas très performant pour cela. Il est construit pour faire émerger ce que Max Reinert [7], [8] nomme des « mondes lexicaux » ou « topoï », simultanément présents dans un texte. Cela s'obtient par un découpage du texte en propositions. La stabilité des classes obtenues par d'autres découpages du texte est moins assurée.

15. Respectivement 250 et 245 énoncés élémentaires.

*« Les connaissances en matière de **théorie économique** ne permettent pas de sauver ce travail trop bref. Le candidat oublie le "dans quelle mesure"[16], il propose une deuxième partie uniquement descriptive, on ne peut accepter la confusion contrats à durée **indéterminée-emplois atypiques ».***

*« Un travail très décevant. La problématique choisie est mal adaptée au traitement du sujet car elle n'établit pas de véritable lien entre **compétitivité et évolution du marché du travail.** Par ailleurs ce plan conduit à une importante disproportion entre les parties. Le candidat se contente le plus souvent de réciter son cours, sans se préoccuper de savoir si ces développements entrent dans le champ du sujet (ce qui est loin d'être le cas) ».*

Dans les énoncés de cette deuxième classe, les correcteurs ne se contentent pas de dire que la copie n'est pas satisfaisante, le plan déséquilibré ou la problématique inadaptée, ils ne signalent pas seulement des écarts à leurs attentes, ils cherchent à aller plus loin. On est assez étonné de ces approfondissements quand on sait que ces appréciations ne seront presque jamais lues par le candidat. Tout se passe comme si certains correcteurs, habitués à corriger en classe selon un certain style, ne s'en défaisaient pas au baccalauréat. Ils essaient de comprendre l'origine de la défaillance observée, ce qui les fait remonter à l'objet d'étude lui-même et entrer en détail dans le contenu de la copie, la démarche du candidat. Pour cette raison, on peut dire que la deuxième classe lexicale suggère un style de correction plus attentif au travail de chaque candidat, somme toute plus approfondi que la correction stylisée dans la première classe.

On peut donc rattacher la différence entre les deux classes d'énoncés à des « styles de correction » et attribuer ces styles à des correcteurs[17]. Mais, comme nous n'avons pas observé de liaison entre ces deux classes lexicales et un niveau de note des copies, on peut dire que ces façons différentes de corriger ne sont pas liées à la qualité de la prestation du candidat, ni à la sévérité ou la largesse dans la notation. Ces styles se distinguent par la façon d'exercer l'activité d'évaluation proprement dite, par le caractère plus ou moins approfondi de l'appréciation synthétique, en référence aux contenus à traiter.

Un traitement analogue des appréciations complètes portées, lors de la multicorrection du baccalauréat à Lille, par chacun des 34 correcteurs, ne nous permet malheureusement pas d'aller plus loin pour caractériser ces différences apparues entre les correcteurs. Ce corpus présente l'intérêt de nous donner des précisions sur les caractéristiques sociodémographiques des correcteurs. Mais le traitement des appréciations selon la même méthode que précédemment ne fournit pas des résultats interprétables, car les cinq classes lexicales obtenues ne peuvent pas être lues comme exprimant des styles de correction repérables. Peut-être cela tient-il au fait qu'à Montpellier, il s'agit d'une correction véritable, alors que dans l'autre cas, à Lille, la correction a un caractère expérimental. De plus, les deux copies multicorrigées étaient de niveau très faible, ce qui contribue, nous l'avons vu, à cantonner les corrections dans le registre de la méthode. Enfin on peut penser que le phénomène de multicorrection, bridant l'expression et lissant les façons de corriger par le contrôle mutuel des correcteurs, atténue d'éventuelles différences de style entre eux. Ces résultats ne permettent pas en tout cas d'étayer l'hypothèse d'existence de styles de correction assez stables attribuables à des correcteurs en toutes circonstances.

Fondements cognitifs et disciplinaires de la correction

La robustesse des critères de corrections qui émergent de l'analyse lexicale, quel que soit le corpus analysé, permet d'affirmer que les professeurs d'une même matière, les sciences économiques et sociales, recourent pour corriger à des critères qui sont similaires. S'il existe des différences dans leurs façons de corriger, elles sont liées au caractère plus ou moins approfondi de la correction.

Les différences les mieux mises en évidence par l'analyse lexicale sont liées aux circonstances et donc aux enjeux de la correction, comme c'est le cas selon que la correction est celle d'une copie de baccalauréat, d'une copie faite en classe, ou d'une copie-test. Les corrections dépendent aussi, à l'évidence, du

THÈME

NOTES

16. Expression qui se trouve dans l'intitulé du sujet posé.
17. Dans la majorité des cas (17 jurys sur 22), les énoncés portés sur des copies d'un même jury sont classés dans la même classe lexicale. Nous identifions les correcteurs et les distinguons ainsi les uns des autres par leur numéro de jury (un même correcteur ne peut pas être dans plusieurs jurys en sciences économiques et sociales), mais nous ne savons rien de plus sur ces correcteurs du baccalauréat.

contenu de la copie, elles se différencient selon la qualité de celle-ci. Les correcteurs agissent donc par rapport à la situation qu'ils rencontrent et pas uniquement en fonction d'attitudes *a priori*. Ils accordent de l'importance à la maîtrise des connaissances disciplinaires, aux raisonnements nécessaires au traitement du sujet et aux savoir-faire de l'exercice constitutif de l'épreuve. Le jugement ainsi formé, bien qu'il soit inégalement étayé selon les cas, porte non sur la personne de l'élève mais sur la qualité de ce qu'il a produit.

Cette conclusion, liée à l'analyse d'un corpus limité aux sciences économiques et sociales, va à l'encontre de celle de Bourdieu et Saint-Martin [9], relative aux « catégories » intellectuelles utilisées par les professeurs lorsqu'ils jugent leurs élèves. Analysant les appréciations portées par un professeur de philosophie en khâgne sur ses élèves durant trois années, ces auteurs observent une relation entre les adjectifs utilisés et l'origine sociale des élèves ; ils concluent à l'usage intériorisé par les professeurs de catégories de jugement fondées sur des proximités culturelles et socialement distinctives. Le vocabulaire qu'ils recensent[18] désigne en effet le plus souvent des qualités ou dispositions des personnes et non des aptitudes techniques à se conformer à des exigences rigoureusement définies. Le vocabulaire de la correction des dissertations de sciences économiques et sociales en 1992, dans l'échantillon représentatif que nous en avons, ne correspond pas à ces spécifications. Il est souvent assez technique, rarement emphatique, et désigne des exigences bien spécifiques. On ne peut nullement en déduire que les professeurs jugent *« des dispositions globales, au demeurant indéfinissables »*, mais au contraire qu'ils cherchent à juger une production d'élève dans un contexte.

Ainsi, il nous semble que les corrections effectives, bien qu'elles portent sur des travaux en classe de terminale et au baccalauréat, ne répondent pas à une pratique de pur classement des personnes, mais qu'au contraire ce classement est étayé sur des qualités attendues du travail des élèves. Il ne s'agit donc pas tant d'évaluer quelqu'un ou une pure « capacité », mais d'évaluer la manifestation de la capacité de quelqu'un à faire quelque chose. Certes, de tels jugements ouvrent à des interprétations multiples et à de nécessaires délibérations, mais ils ont pour fondement la prestation des élèves dans un type précis d'épreuve, inscrite dans une discipline, ce qui permet de délimiter les attentes.

Nous avons ainsi mis en évidence l'importance des critères disciplinaires, méthodologiques et cognitifs, dans la correction. De plus les correcteurs agençnt ces critères en les combinant différemment selon les situations de correction et le niveau de la prestation des élèves.

Nous avons pu repérer des styles de correction différents, mais leur stabilité n'est pas sûre et nous n'avons pas pu les relier à des caractéristiques repérables des correcteurs. Plus généralement, l'étude montre que les caractéristiques des correcteurs (leur appartenance à telle ou telle catégorie statistique) n'apportent qu'une faible lumière à la question tant de la notation que du style de correction. Il faudrait confirmer ces premières remarques par d'autres travaux prenant en compte diverses situations d'évaluation (en classe en particulier) et diverses disciplines. Nous avançons néanmoins l'hypothèse que les variables caractérisant *a priori* les correcteurs ne sont pas très pertinentes pour comprendre les pratiques effectives d'évaluation des professeurs. Il y a ici une analogie à établir avec différents travaux sur l'effet conjoint maître-classe. Cet effet, régulièrement mis en évidence par l'analyse statistique, reste cependant largement inexpliqué par les variables qu'on mobilise pour en rendre compte (Bressoux [10], Duru-Bellat et Leroy-Audouin [11], Serra et Thaurel-Richard[12]). ■

NOTE

18. Ils ne retiennent que les adjectifs.

Méthodologie de l'enquête

L'enquête a concerné 4 302 élèves de classe terminale dans les académies de Paris, Versailles, Lille, Rouen, Montpellier. Elle a été effectuée de la façon suivante. Un questionnaire a été envoyé par voie postale à 400 professeurs de sciences économiques et sociales de ces académies (cette matière a été choisie parce qu'elle est spécifique à cette section). 167 réponses nous ont été retournées. On a ensuite recueilli les notes obtenues en classe par les élèves de terminale de ces professeurs, ainsi que des données sur leurs caractéristiques sociodémographiques inscrites sur les fiches archivées dans leurs établissements. Les notes sont donc les moyennes trimestrielles des élèves sur 20. On a recherché ensuite dans les rectorats et au SIEC les notes obtenues à l'écrit du baccalauréat par ces mêmes élèves, dans les cinq matières de l'écrit du baccalauréat et en français (matière pour laquelle nous ne disposons pas des notes obtenues en classe). Le recoupement des deux séries de données a été obtenu pour 126 classes. L'échantillon est représentatif de la population d'élèves concernée. Il contient 63 % de jeunes filles, son âge moyen est de 18,8 ans. Quant aux professeurs de sciences économiques et sociales, notre échantillon a sensiblement la même structure par sexe (avec 41 % de femmes contre 44 %) que l'ensemble des professeurs de cette discipline, il contient plus de titulaires (81,5 % contre 79,7 %) puisque nous n'avons retenu que les professeurs exerçant en classe de terminale. Probablement aussi de ce fait il est un peu plus âgé que la population de référence avec une sous-représentation des plus jeunes (moins de 30 ans) et une surreprésentation des 40-49 ans. Ces matériaux sont traités ici par des modèles de régression multiple[19].

19. Une première analyse de ces matériaux a été publiée en 1994 [13]. Nous avons procédé ici à de nouvelles analyses, qui ont été effectuées par Danièle Trancart, maître de conférences à l'Université de Rouen. Nous la remercions vivement de son apport à cette recherche.

À LIRE

[1] **J.-C. Forquin**, *École et culture, le point de vue des sociologues britanniques,* Bruxelles, De Boeck-Éditions universitaires, 1989.

[2] **J.-P. Caverni** et **G.Noizet**, « La notation en classe terminale et au baccalauréat : données monographiques et effets d'intervention expérimentale », *Les sciences de l'éducation*, octobre-décembre 1985.

[3] *Rapport de l'Inspection générale 1997,* Ministère de l'Éducation nationale, de la Recherche et de la Technologie, La documentation française.

[4] **P. Merle**, *L'évaluation des élèves, enquête sur le jugement professoral,* Paris, PUF, 1996.

[5] **B. Simler** (dir.), *L'évaluation en sciences économiques et sociales*, CRDP de Lille, 1987.

[6] **M. Reinert**, « ALCESTE, une méthodologie d'analyse des données textuelles et une application : *Aurélia* de Gérard de Nerval », *Bulletin de méthodologie sociologique,* n° 26, mars 1990, pp. 24-54.

[7] **M. Reinert**, « Mondes lexicaux et Topoï dans l'approche ALCESTE », Intervention au *Congrès international des linguistes,* Paris, 21 et 25 juillet 1997, *mimeo.*

[8] *Évaluation des compétences scolaires des meilleurs élèves depuis quarante ans,* Les dossiers d'Éducation et formations, n° 69, MEN-Direction de l'évaluation et de la prospective, mai 1996.

[9] **P. Bourdieu** et **M. de Saint-Martin**, « Les catégories de l'entendement professoral », *Actes de la Recherche en sciences sociales,* n° 3, mai 1975, pp.68-93.

[10] **P. Bressoux**, *Les performances des écoles et des classes, le cas des acquisitions en lecture,* Les dossiers d'Éducation et formations, n° 30, MEN-Direction de l'évaluation et de la prospective, juin 1993.

[11] **C. Audouin-Leroy** et **M. Duru-Bellat**, « Pratiques pédagogiques et acquis des élèves au cours préparatoire », *revue Éducation et formations,* n° 26, MEN-Direction de l'évaluation et de la prospective, janvier-mars 1991, pp.3-13.

[12] **N. Serra** et **M. Thaurel-Richard**, « Les acquisitions des élèves au CE2 et les pratiques pédagogiques », *Revue française de pédagogie,* n° 107, avril-mai-juin 1994, pp.43-62.

[13] **É. Chatel**, *Qu'est-ce qu'une note ? Recherche sur la pluralité des modes d'éducation et d'évaluation,* Les dossiers d'Éducation et formations, n° 47, MEN-Direction de l'évaluation et de la prospective, décembre 1994, pp.183-203.