



Les Documents de Travail de l'IREDU

Working Papers

Institut de Recherche sur l'Education

Sociologie et Economie de l'Education

Institute for Research in the Sociology and Economics of Education

**La loterie des notes au bac
Un réexamen de l'arbitraire de la notation des élèves**

Bruno Suchaut

Mars 2008
DT 2008/3

La loterie des notes au bac

Un réexamen de l'arbitraire de la notation des élèves

Bruno Suchaut
Irédu-CNRS et Université de Bourgogne
Mars 2008

La question de la notation à l'école n'est pas récente, les nombreuses études réalisées sur ce thème mettent toutes en évidence l'incertitude de la mesure des performances scolaires des élèves par les notes et les divers biais associés à cette pratique évaluative (Merle, 1996, 1998, 2007). Ce sont les psychologues qui ont le plus contribué à l'étude de la notation dans une perspective docimologique (Noizet, 1961 ; Pierron, 1963 ; Noizet, Caverni, 1985). Pour la majorité d'entre eux, ces travaux ont consisté à montrer de façon expérimentale le manque de validité et de fidélité des épreuves d'examen. Ce constat est à présent bien établi et ce texte n'a pas l'ambition d'exposer des analyses novatrices en ce domaine, il s'agit simplement de témoigner à nouveau du caractère aléatoire de la notation sur la base d'une expérimentation mobilisant des données récentes. A l'heure où l'on s'interroge en France sur les procédures d'évaluation et de certification des élèves, notamment au niveau du baccalauréat, une analyse réalisée sur des données actuelles ne peut qu'alimenter la réflexion sur la pertinence des évaluations certificatives et plus spécifiquement des examens qui balisent la scolarité des élèves.

L'objet d'étude n'est pas nouveau dans la mesure où, dès les années 1930, la notation des épreuves du baccalauréat a fait l'objet d'une célèbre recherche (Laugier, Weinberg, 1936). A cette époque, la commission française pour l'enquête Carnégie a réalisé des analyses statistiques sur les notes en soumettant à 5 correcteurs, 100 copies dans 6 disciplines. Les résultats bien connus de cette recherche ont mis en évidence de forts écarts de notation entre les différents correcteurs dans les disciplines littéraires, mais aussi, de façon moins marquée, dans les disciplines scientifiques. Cette recherche a également permis de souligner l'incertitude du jugement professoral qui pèse sur les résultats à l'examen du baccalauréat car environ 30% des candidats peuvent être admis ou ajournés selon la composition du jury. Il a fallu attendre les années 1990 pour que le baccalauréat fasse l'objet de nouvelles études, confirmant les tendances passées en matière d'écarts de notation entre correcteurs (Merle, 1996).

Notre démarche est limitée puisqu'elle se centre essentiellement sur l'examen et l'analyse de la variété de la notation à des copies du baccalauréat. L'objectif est de confirmer, sur la base de données actuelles, l'incertitude de la notation à l'examen. Après un bref rappel des conclusions des travaux sur la notation des élèves, ce texte rendra compte d'une

expérimentation menée dans 2 académies sur des copies de baccalauréat en sciences économiques et sociales (S.E.S.) soumises à la correction d'un large échantillon de professeurs.

I Ce que l'on sait sur la notation des élèves

La notation est une pratique persistante et généralisée dans le système éducatif français. Depuis le début de l'école élémentaire jusqu'au terme des études supérieures, les productions des élèves sont fréquemment et régulièrement notées. Il existe bien sûr des situations où des enseignants et des établissements se basent sur d'autres outils pour évaluer les résultats des élèves, mais ces usages ne sont pas majoritaires. La simplicité de la notation en fait en partie son succès, le chiffrage des performances des élèves est en effet lisible aisément par tous les acteurs : les enseignants, les élèves et leurs parents. Pour l'enseignant, la notation peut également être utilisée plus ou moins indirectement comme un instrument de pouvoir sur les élèves traduisant une gratification ou une sanction. La persistance de cette pratique peut aussi se lire comme un frein à un changement plus global des pratiques pédagogiques. L'abandon des classements et des notes conduirait ainsi à une autre conception de l'évaluation et peut-être même à une autre école (Perrenoud, 1991).

Les nombreuses critiques associées à la notation concernent plusieurs dimensions. Une première a trait à la qualité de la mesure des connaissances et des compétences des élèves, la note ne rendant compte qu'imparfaitement des performances. Ainsi, les corrélations relevées entre les scores obtenus à des tests de compétences scolaires et des notes d'épreuves du bac sont faibles (Jouvanceau, 1989). Par ailleurs, la relation entre les notes obtenues en cours d'année et les notes au bac n'est pas non plus parfaite (Chatel, 1994). C'est aussi le cas au niveau du collège où l'examen des corrélations entre notes au brevet, contrôle continu et épreuve commune amènent à s'interroger sur la validité des différentes mesures (Murat, 1998). Ces relations imparfaites tiennent au fait que la notation est toujours contextualisée : elle est attribuée par un enseignant, dans une classe, dans un établissement et à un élève particulier.

Les caractéristiques individuelles de l'élève et celles de son environnement scolaire vont donc engendrer des biais récurrents tout au long de la carrière scolaire des élèves (Duru-Bellat, Mingat 1993). Ces biais ont été étudiés dans de nombreuses recherches et concernent les caractéristiques personnelles des individus : attributs physiques (Nilson, Nias, 1977), sociaux (Weiss, 1969 ; Pourtois, 1978) et sexués (Spear, 1989). Le niveau (Caverni, Fabre, Noizet, 1975) et le statut scolaire (Bonniol, Caverni, Noizet, 1972) de l'élève conditionnent également le jugement de l'enseignement et se traduisent aussi par des biais de notation. L'objet n'est pas ici de recenser les résultats de ces travaux, on mentionnera simplement qu'à niveau de compétences identique, les élèves vont être évalués différemment par leurs

enseignants : les élèves de milieux sociaux défavorisés, les élèves au passé scolaire difficile et en retard scolaire sont ceux qui sont les plus pénalisés.

Le contexte de scolarisation n'est pas non plus indépendant des pratiques en matière de notation. Selon l'établissement fréquenté, les performances des élèves sont évaluées différemment. Les collèges, recevant un public d'origine populaire et de faible niveau scolaire, ayant tendance à être plus indulgents, alors que ceux scolarisant un public plus favorisé et de niveau scolaire élevé font preuve de plus de sévérité (Duru-Bellat, Mingat, 1993). Les pratiques de notation reflètent alors un aspect de politiques d'établissement différenciées qui doivent être lues en complémentarité d'autres mécanismes comme l'orientation et la sélection des élèves.

Les conséquences des aléas et des biais de la notation sur les élèves ne sont pas négligeables. Tout d'abord au niveau psychologique, la réussite scolaire conditionnant encore plus qu'auparavant la vie professionnelle et sociale, la validation des acquis prend une importance toute particulière dans la vie de l'élève ou de l'étudiant. Les examens sont donc source d'angoisse et de stress non négligeables. En outre, tout au long du cursus scolaire, la notation va témoigner de la réussite ou de l'échec de l'élève. Or, on connaît les conséquences de l'échec sur les comportements des élèves : la réussite a tendance à engendrer la réussite et l'échec à engendrer l'échec. Ce phénomène, qui s'exprime comme une prophétie qui se réalise d'elle-même (Self-Prophecy), a été étudié dans les célèbres travaux de Rosenthal et Jacobson (1968). Ces conséquences sont d'autant plus regrettables qu'elles découlent d'une mesure chiffrée imprécise de la performance scolaire. L'influence de la notation concerne aussi les parcours des élèves. De nombreux travaux, réalisés dans des contextes scolaires très variés, ont bien mis en évidence les inégalités produites par la notation dans les décisions de redoublement et d'orientation, la notation étant très marquée par le contexte de la classe ou de l'établissement (Duru-Bellat, Mingat, 1993 ; Jarousse, Leroy-Audouin, 1997 ; Jarousse, Suchaut, 2002 ; Mingat, Rakotomalala, Suchaut, 1999).

II L'expérimentation sur les copies du bac

L'expérimentation a été réalisée dans 2 académies au cours de 2 années scolaires consécutives (2006 et 2007). Dans chacune des académies, 3 copies d'élèves ayant passé le bac (en juin 2006 et en juin 2007) ont été soumises à la correction d'une trentaine de professeurs de sciences économiques et sociales¹. Les enseignants devaient attribuer une note à chacune des 3 copies et formuler des commentaires écrits justifiant la note. Dans chaque académie, les copies ont été choisies de sorte à avoir une production d'un « bon » niveau (note obtenue au bac : 15) et deux copies jugées moyennes (une copie ayant obtenu 9 et une copie ayant obtenu

¹ Cette expérimentation a été réalisée dans les académies de Dijon et de Besançon et a servi de support à une formation à l'évaluation à destination des professeurs de sciences économiques et sociales. Les enseignants n'ont pas eu connaissance des notes effectivement obtenues par les auteurs des copies au baccalauréat.

11 au bac). L'échantillon total rassemble 98 notes (3 notes pour chacun des 34 correcteurs de l'académie 1 et 3 notes pour chacun 32 correcteurs de l'académie 2).

Des notes très variées pour une même copie

Le tableau 1 présente l'ensemble des données collectées dans le cadre de cette expérimentation.

Tableau 1 : Notes attribuées aux 6 copies de S.E.S. par les 66 correcteurs

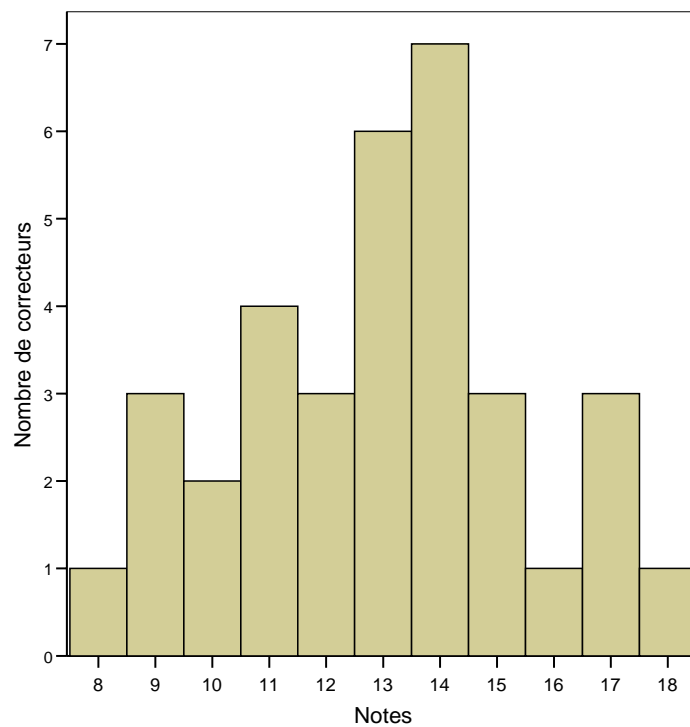
Correcteurs	Académie 1			Correcteurs	Académie 2		
	Notes				Notes		
	Copie 1	Copie 2	Copie 3		Copie 4	Copie 5	Copie 6
1	6	9	10	35	7	15	10
2	7	6	10	36	10	11	8
3	12	7	14	37	10	13	8
4	7	5	14	38	9	11	8
5	11	15	13	39	6	9	4
6	8	5	17	40	8	13	9
7	7	9	15	41	9	12	7
8	11	7	14	42	9	14	8
9	7	12	8	43	9	13	7
10	8	11	11	44	6	13	9
11	8	13	12	45	7	13	9
12	7	12	15	46	9	14	7
13	15	8	17	47	8	10	8
14	11	9	13	48	13	15	10
15	11	14	9	49	11	11	8
16	8	6	15	50	12	10	7
17	6	16	11	51	8	12	6
18	14	8	12	52	7	15	9
19	11	7	14	53	12	16	7
20	13	7	9	54	8	17	10
21	5	10	14	55	11	13	14
22	9	12	13	56	9	12	9
23	8	9	14	57	13	15	8
24	9	6	17	58	8	11	8
25	6	8	11	59	9	15	5
26	10	9	13	60	10	15	8
27	8	8	11	61	3	16	6
28	7	9	14	62	7	8	6
29	9	6	13	63	9	13	10
30	7	10	16	64	10	12	7
31	7	9	18	65	9	14	8
32	9	8	12	66	7	12	9
33	10	8	13				
34	6	8	9				

Sur l'ensemble des 198 notes, la moyenne est de 10 avec un écart-type de 3, les notes variant de 3 à 18. L'examen du tableau par colonne montre qu'il existe, pour chaque copie, des variations très fortes d'un correcteur à l'autre dans les 2 académies. Une première étape est de décrire la répartition des notes attribuées à une même copie par les différents correcteurs de l'échantillon. Le tableau 1 présente les principaux paramètres statistiques des distributions des notes des 6 copies.

Tableau 2 : Paramètres des distributions des notes attribuées aux 6 copies

	Minimum	Maximum	Moyenne	Ecart-type	Médiane	Mode
Académie 1 (N = 34)						
Copie 1	5	15	8,8	2,4	8,0	7
Copie 2	5	16	9,0	2,7	8,5	8
Copie 3	8	18	13,0	2,5	13,0	14
Académie 2 (N = 32)						
Copie 4	3	13	8,8	2,1	9,0	9
Copie 5	8	17	12,9	2,1	13,0	13
Copie 6	4	14	8,0	1,8	8,0	8

Les répartitions reflètent tout d'abord un phénomène statistique classique puisque les distributions se rapprochent de la loi gaussienne avec une concentration des notes autour de la moyenne et de la médiane. A titre d'illustration, le graphique 1 permet d'avoir une image concrète de la distribution des notes relative à la copie n°3.



Graphique 1 : Distribution des notes attribuées à la copie n°3

Un deuxième constat concerne la dispersion associée aux distributions avec des valeurs extrêmes très éloignées de la moyenne. Pour 4 copies l'écart maximal est de 10 points et il est de 9 et 11 points pour les deux autres copies, ce qui est considérable. Les écart-types, qui fournissent une mesure statistique de la dispersion moyenne des notes, présentent également des valeurs élevées (de 1,8 à 2,7 points). Ces données récentes confirment donc bien l'incertitude de la notation des élèves, les notes d'une même copie pouvant fortement varier d'un correcteur à l'autre. A titre d'illustration, l'élève auteur de la copie n°2 (qui a réellement obtenu 9 à l'examen du bac), selon que sa copie soit corrigée par l'un ou l'autre des 34 correcteurs, a une chance sur 6 d'obtenir une note supérieure à 12, mais aussi... un risque sur 6 d'obtenir une note inférieure à 6 !

L'origine des différences de notation

Il s'agit ici de chercher à expliquer statistiquement l'origine des différences entre les notes en fonction des variables disponibles dans l'expérimentation. Un premier facteur explicatif peut concerner les échantillons de correcteurs, ceux-ci se différenciant par leur académie d'appartenance et le sujet de l'examen, lui-même différent d'une année sur l'autre. Il n'existe pas de différences statistiquement significatives entre les moyennes des notes relevées sur chacun des échantillons de correcteurs (10,2 pour l'académie 1 et 9,9 pour l'académie 2). Les notes ne se différencient donc pas selon l'appartenance à une académie ou une autre ou, si l'on examine cela sous un autre angle, d'un sujet d'examen à un autre.

Un second facteur est l'élève auteur de la copie puisque nous disposons sur l'ensemble de l'échantillon des copies provenant de 6 élèves différents (3 dans chaque académie). Rappelons que les copies ont été choisies de sorte à avoir deux copies moyennes (notes de 9 et de 11) et une bonne copie (note de 15). Le tableau suivant présente les résultats des comparaisons de moyennes (t de Student pour échantillons appariés) réalisées sur chaque série de 3 copies dans chaque académie. Ces résultats nous indiquent que dans les 2 académies, et en moyenne, les correcteurs distinguent dans leur notation la « bonne » copie (copie 3 et copie 5) des deux autres. Les deux copies moyennes (respectivement la 1 et la 2 et la 4 et la 6 pour les académies 1 et 2) présentent des différences plus faibles (0,2 et 0,8 points) que l'écart réel initial de 2 points (entre les notes 9 et 11). Il y a donc bien, en moyenne, sur les 2 échantillons, une capacité des correcteurs à discriminer les copies, même si, comme nous l'avons déjà évoqué, les écarts intercorrecteurs pour une même copie sont considérables.

Tableau 3 : Comparaison des moyennes entre les copies

	Différence moyennes	t de Student
Académie 1 (N = 34)		
Copie 1 et copie 2	-0,2	0,3 n.s.
Copie 1 et copie 3	-4,2	7,3 ***
Copie 2 et copie 3	-4,0	5,4 ***
Académie 2 (N = 32)		
Copie 4 et copie 5	-4,1	7,9 ***
Copie 4 et copie 6	+0,8	1,9 **
Copie 5 et copie 6	+4,8	11,6 ***

n.s. : non significatif, ** : significatif au seuil de 5%,

*** : significatif au seuil de 1%

Le troisième facteur est central dans notre démonstration et concerne le correcteur lui-même. Les informations précédentes (tableau 1) invitent à mesurer l'ampleur de la variété des notes entre les correcteurs. En même temps, il est possible de répondre à une question complémentaire qui est celle de la régularité ou de la cohérence des jugements de chacun des correcteurs lors de la correction des 63 copies.

Instabilité de la notation d'un même correcteur

On pourrait penser qu'un même correcteur est fidèle à sa pratique en matière de notation. Or, on sait depuis longtemps (Hartog, Rhodes, 1935) qu'il existe des phénomènes de variations importants relevés dans la constance des corrections d'un même examinateur dus à différents facteurs liés au correcteur lui-même (fatigue, distraction etc...), mais aussi au contexte de correction, dont l'ordre de correction des copies (Bonniol, 1965 ; Bonniol, Piolat, 1971). Ces facteurs ne peuvent jouer qu'à la marge dans notre expérimentation, le nombre de copies étant beaucoup trop faible. Il est possible en revanche de montrer que les enseignants n'ont pas la même constance dans leur évaluation en termes d'indulgence ou de sévérité, et c'est donc davantage la modification des critères de corrections, d'une copie à l'autre qui peuvent être évoqués. Compte tenu de la structure des données, notamment du rapport entre le nombre de copies et le nombre de correcteurs², deux analyses statistiques ont été réalisées pour mettre en évidence ce phénomène. La première consiste à réaliser une analyse de variance selon un plan à mesures répétées (ANOVA pour échantillons appariés) dans laquelle on considère que chaque sujet (en l'occurrence ici un correcteur) est associé à plusieurs mesures (les 3 notes de la série de copies). Trois sources de variation sont identifiées dans cette analyse : entre les copies, entre les correcteurs et une source de variation due à l'erreur expérimentale.

Les tableaux 4 et 5 présentent les résultats des analyses de variance réalisées sur les 2 échantillons de correcteurs.

² Les conditions de l'expérimentation n'ont pas permis d'envisager que chaque correcteur puisse corriger un nombre important de copies, ce qui limite évidemment les possibilités d'analyses statistiques.

Tableau 4 : Analyse de variance des notes : plan à mesures répétées (académie 1)

Source de variation	Somme des carrés	Degrés de liberté	Carré moyen	F
Intercopies	379,8	2	189,9	25,5 ***
Intercorrecteurs	165,5	33	5,0	0,67 n.s.
Erreur	491,5	66	7,4	
Totale	1036,9	101		

*** : significatif au seuil de 1% , n.s. : non significatif

Tableau 5 : Analyse de variance des notes : plan à mesures répétées (académie 2)

Source de variation	Somme des carrés	Degrés de liberté	Carré moyen	F
Intercopies	436,6	2	218,3	65,2 ***
Intercorrecteurs	174,5	31	5,6	1,7 n.s.
Erreur	207,4	62	3,3	
Totale	818,5	95		

*** : significatif au seuil de 1% , n.s. : non significatif

Les résultats des ANOVA sont assez comparables sur les deux échantillons puisque seules les variations intercopies sont significatives³. Le fait que les variations entre les correcteurs (intercorrecteurs) soient faibles et non significatives montre qu'il n'existe pas d'effets de régularité en matière de sévérité ou d'indulgence sur cet échantillon⁴. Enfin, la part de variance due à l'erreur d'échantillonnage est importante (elle est plus forte dans la première académie que dans la deuxième) et témoigne de l'aspect aléatoire de la notation. Autrement dit, d'autres échantillons de correcteurs auraient donné lieu à des distributions de notes différentes.

Une seconde possibilité pour tester cette instabilité dans les pratiques de correction, est de calculer des corrélations entre les notes attribuées aux différentes copies. Un coefficient de corrélation affichant une valeur élevée signifierait que les correcteurs se classeraient de la même manière en terme de sévérité ou d'indulgence pour les 3 copies corrigées. Le tableau suivant présente, pour chaque série de copies, les coefficients de corrélations entre les notes. Les résultats sont nets car aucune corrélation positive et significative n'est relevée. Cela signifie que les correcteurs les plus sévères (ou les plus généreux) pour une copie particulière ne sont pas ceux qui font preuve de plus de sévérité (ou d'indulgence) pour les 2 autres copies. C'est même la tendance inverse qui se manifeste quand on compare la notation des copies 2 et 3 : les correcteurs sévères pour la copie 2, on tendance à évaluer avec plus d'indulgence la copie 3 et réciproquement.

³ Les variations entre les copies représentent 37% des variations totales des notes pour l'académie 1 et 53% pour l'académie 2.

⁴ Les termes « sévérité » et « indulgence » ont un sens très relatif puisque l'on ne dispose pas de réelle référence pour juger dans l'absolu du degré d'indulgence ou de sévérité des correcteurs.

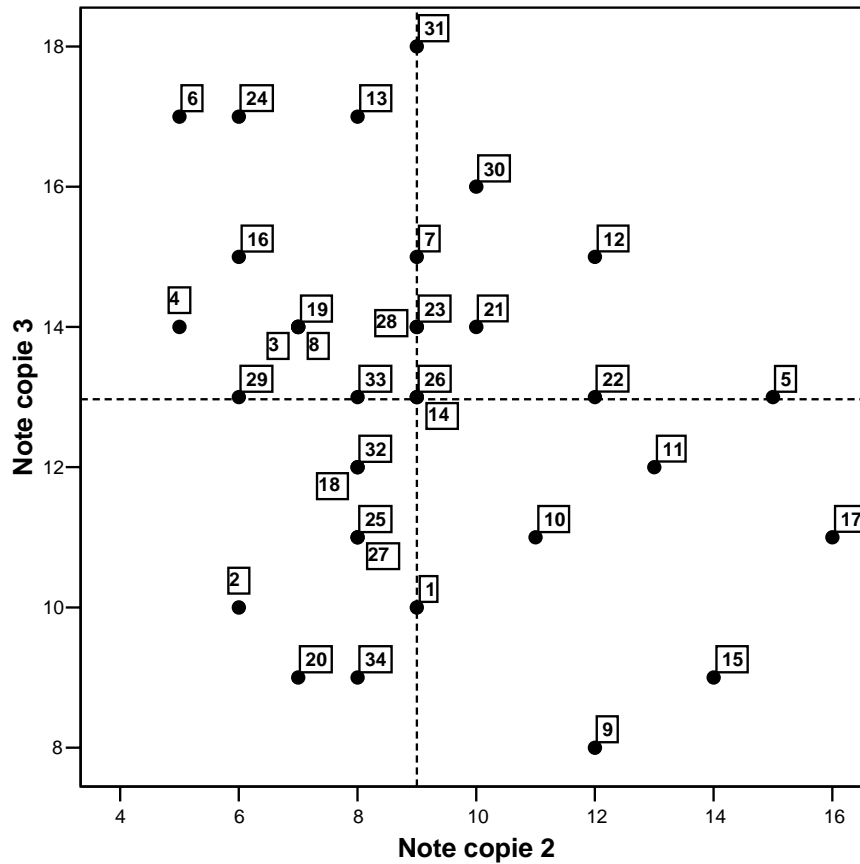
Tableau 6 : Corrélations entre les notes attribuées aux copies dans les deux académies

Académie 1 (N = 34)		
	Copie 1	Copie 2
Copie 2	-0,13 n.s.	
Copie 3	+0,08 n.s.	-0,30 n.s.
Académie 2 (N = 32)		
	Copie 4	Copie 5
Copie 5	+ 0,07 n.s.	
Copie 6	+0,23 n.s.	+0,28 n.s.

n.s. : non significatif

Le graphique suivant permet d'examiner plus en détail la logique de notation des différents correcteurs en ce qui concerne les copies 2 et 3⁵ ; les axes matérialisent les notes des copies et les points représentent les 34 correcteurs de l'académie 1. Si les correcteurs étaient aussi généreux (ou indulgents) pour la notation des deux copies, les points seraient sensiblement alignés sur une droite avec une pente positive. Autrement dit, le classement des correcteurs sur une échelle d'indulgence (ou de sévérité) serait identique pour les deux copies, ce qui est loin d'être le cas ici. Seuls quelques correcteurs semblent être réguliers. C'est par exemple le cas pour les correcteurs 26 et 14 qui attribuent aux deux copies la note moyenne relevée sur l'échantillon. De même, les correcteurs 2, 20 et 34 jugent les deux copies très défavorablement en leur accordant des notes très faibles. Les correcteurs ne semblent donc pas être constants dans leur pratique de notation. Ainsi, le correcteur 15 a attribué la note de 9 à la copie n°3 (soit l'une des plus faibles notes associées à cette copie) alors qu'il a attribué la note de 14 à la copie n°2 (soit l'une des notes les plus élevées de l'échantillon). De même, le correcteur 6 a noté sévèrement la copie n°2 (note de 5 sur 20) alors qu'il est beaucoup plus généreux pour la copie n°3 (note de 17 sur 20).

⁵ Les lignes pointillées représentent la moyenne des notes sur l'ensemble de l'échantillon des 34 correcteurs.



Graphique 2 : Relation entre les notes attribuées aux copies 2 et 3 par les 34 correcteurs

Les analyses tendent donc bien à montrer que les correcteurs ne se distinguent pas dans leur notation selon ce critère d'indulgence (ou de sévérité). Une question est à présent de savoir comment les notes sont justifiées par les correcteurs.

Comment les notes sont-elles justifiées ?

L'expérimentation a également permis de mobiliser des éléments plus qualitatifs avec le recueil des commentaires figurant sur les copies et justifiant les notes attribuées aux élèves. Ce travail d'analyse des commentaires a déjà été réalisé sur le même type de copies (épreuve de dissertation en S.E.S.) dans les années 95 (Simler, 1995 ; Merle, 1996), nos résultats vont dans le même sens avec une forte diversité des commentaires. Ceux-ci sont effectivement variés, à la fois dans leur contenu et dans leur volume. Les critères avancés par les correcteurs concernent le fond et la forme des productions. Au niveau du fond, les éléments suivants peuvent être distingués : plan, traitement du sujet, argumentation. Au niveau de la forme, ce sont la qualité de l'expression, l'orthographe et la grammaire qui dominent. Une tendance générale à la lecture des commentaires est que les commentaires sont beaucoup plus fournis pour les notes faibles que pour les notes élevées. On relève ainsi, sur l'ensemble des copies, une corrélation négative et significative entre le nombre de mots composant les commentaires

et les notes⁶. C'est comme si le correcteur se sentait d'autant plus contraint, même dans un cadre expérimental, de justifier sa pratique quand son jugement sur la production de l'élève est négatif. Il ne s'agit pas ici de présenter l'ensemble des propos rédigés par les correcteurs, mais uniquement de montrer comment peuvent être justifiées des notes très différentes attribuées à une même copie. Le tableau 7 expose des extraits des commentaires de 6 correcteurs pour 3 copies ; pour chacune de ces copies, ont été associées les remarques de deux correcteurs, l'un indulgent, l'autre plutôt sévère dans sa notation.

Comme cela vient d'être évoqué, les commentaires des copies ayant obtenu de faibles notes sont beaucoup plus fournis (les commentaires de ces copies n'ont d'ailleurs pas été reproduits dans leur totalité). On constate que les jugements professoraux diffèrent tout autant que les notes. La copie n°3 est évaluée par le correcteur 55 (note : 14) comme un bon devoir, les connaissances étant considérées comme maîtrisées ce qui est loin d'être l'avis du correcteur 39 : pour cet enseignant, l'élève ne maîtrise pas toutes les notions (la note attribuée est 4). En ce qui concerne la copie n°4 (notes : 3 et 13), les divergences entre les correcteurs ne sont pas négligeables. Pour l'un (correcteur 57), le sujet est très peu traité, mal analysé et les documents mal mobilisés ; pour le second (correcteur 61), le sujet est bien cerné et l'argumentaire est convaincant. La copie n°5 (notes : 9 et 16) ne réunit pas non plus des avis concordants de la part des deux correcteurs concernés (correcteurs 39 et 53). L'un juge globalement la production de l'élève comme un devoir correct, bien écrit avec des notions bien mobilisées. L'autre correcteur a un avis beaucoup plus réservé sur la copie : problématique non approfondie, notions mal mobilisées, documents mal utilisés...

⁶ $r = -0,40$, significatif au seuil de 1%.

Tableau 7 : Extraits de commentaires de correcteurs pour 3 copies

Copie n° 6	
Correcteur n° 39 - note attribuée : 4	Correcteur n° 55 - note attribuée : 14
Commentaires	Commentaires
<i>Pas vraiment de plan global et encore moins de sous-plan....Un style lourd et parfois emphatique. Des approximations, des clichés, tout au long du devoir, avec une accumulation sur la fin. Pas d'utilisation des documents. Dans l'introduction, des généralités en guise d'accroche et de définitions. Une problématique peu fouillée. Une annonce de plan ratée, qui ne dit rien. ..Il y a des approximations dans les mécanismes évoqués. Non seulement on ne voit pas où est le plan, mais en outre toutes les notions ne sont pas maîtrisées....</i>	<i>Bon devoir, dommage que l'aspect cumulatif ne soit évoqué qu'en conclusion, cela aurait mérité un réel développement, une introduction un peu confuse, des notions parfois implicites; des connaissances maîtrisées, attention à l'expression écrite. Quelques longueurs parfois</i>
Copie n° 4	
Correcteur n° 57 - note attribuée : 3	Correcteur n° 61 - note attribuée : 13
Commentaires	Commentaires
<i>Une introduction qui n'est pas satisfaisante. Un effort pour proposer une accroche factuelle mais le propos n'est pas très clair. Faute d'analyse du sujet, ce dernier n'est pas questionné, ni reformulé. Le plan annoncé est trop peu précisé, et la troisième partie est hors sujet... De nombreux propos page 6 et début 7 où les conséquences en terme d'exit de l'exclusion sont décrites mais le candidat déclare mobiliser le doc 5, or il semble que ce soit le 3. Un propos absolument non étayé ni par les connaissances ni par l'exploitation des documents (qui sont parfois évoqués mais non exploités). La sensation que le candidat fait du remplissage.</i>	<i>Le sujet est assez bien cerné. Des notions clés sont utilisées (flexibilité, précarité, chômage, pauvreté, stigmatisation, anomie,..). Les arguments sont développés et permettent de mettre en évidence les différents mécanismes de l'exclusion (micro-macro social). Cependant, le devoir manque parfois de nuance, de rigueur (dans le tri et la structuration des idées, expression, orthographe..) et les documents ne sont pas toujours bien exploités qui interrogent tout au long du développement... Appel à des concepts non définis... Un moment plus pertinent à la fin de la copie.</i>
Copie n° 5	
Correcteur n° 39 - note attribuée : 9	Correcteur n° 53 - note attribuée : 16
Commentaires	Commentaires
<i>Introduction correcte, peu approfondie mais claire. Les étapes sont respectées, le sujet est cadré, l'exclusion est saisie dans ces deux aspects. Le plan est pertinent mais pas assez articulé. La problématique n'est pas approfondie. La notion de changement social est intéressante, mais elle mériterait d'être précisée. En somme, l'élève semble aller dans la bonne direction, les termes du débat sont posés, mais de façon presque trop synthétique. L'idée de commencer sur le contexte de la crise du travail et ses effets escomptés quant à la résorption du chômage est intelligente. L'ensemble est maladroit, avec des phrases qui se contredisent. Première partie : La crise du travail est évoquée, mais pas le chômage de masse (ou à peine... ou trop tardivement)..... aucun lien entre pauvreté et exclusion. On nous dit que la crise du travail est un facteur de désintégration sociale. Mais l'explication est courte. Il y a volonté de faire la transition avec la deuxième partie mais celle-ci est ratée ! D'autres éléments sont évoqués à mauvais escient on reste sur notre faim... En somme, même si on retrouve certaines idées attendues, l'ensemble n'est ni expliqué, ni structuré. Pas vraiment de sous-parties. Les documents sont mal utilisés (mauvaise lecture des chiffres).</i>	<i>Sujet cerné, définit correctement l'exclusion. le plan annoncé aurait pu être explicité : lien des thèmes avec l'exclusion. Conclusion : assez bien elle montre bien que l'exclusion est d'une part un processus, d'autre part multiforme. Expression : aisée dans l'ensemble, agréable à lire Documents : bonne capacité d'analyse et d'exploitation. Première partie : plan cohérent, transitions entre les idées mécanismes à l'œuvre bien comprise. Bonne utilisation des notions. Quelques arguments à compléter Deuxième partie : Prend en compte les différents aspects de l'intégration sociale. Aurait dû soulever les problèmes relatifs à la protection sociale. Dommage qu'il y ait absence de référence à des auteurs mais le candidat a parfaitement compris le problème à traiter et démontre de bonnes capacités dans l'analyse.</i>

III En conclusion : quels enseignements pour le baccalauréat ?

Cette expérimentation ne fait que confirmer les conclusions bien connues des recherches antérieures sur la question des aléas de la notation. Ces aléas persistants interrogent forcément sur leurs conséquences sur la carrière des élèves et sur l'obtention des diplômes, et en l'occurrence ici, du baccalauréat. Les écarts de notation d'un correcteur à l'autre sont très importants (jusqu'à 10 points) et peuvent avoir, pour certains élèves, une réelle influence sur la réussite à l'examen. Ainsi, pour l'épreuve de sciences économiques et sociales qui est affectée d'un coefficient 7, un écart de 5 points en matière de notation entraîne au total des différences de 35 points, ce qui est loin d'être négligeable... Evidemment, on peut penser que les aléas de notation ont tendance à se compenser avec le nombre d'épreuves mais les coefficients variables affectés à ces épreuves ne donnent pas à toutes les erreurs de mesure le même poids. La réflexion suscitée par l'expérimentation présentée dans ce texte peut être menée à plusieurs niveaux.

Le premier niveau concerne la notation des élèves lors des examens et la question de la réduction de l'incertitude de la mesure est difficile à traiter. Une voie possible est de fournir des outils permettant de renforcer la cohérence des pratiques en la matière. Les grilles de correction avec un barème précis peuvent contribuer à garantir des aléas moins importants. Toutefois, même dans des disciplines scientifiques, l'utilisation d'un barème précis détaillé en de nombreuses sous-questions est loin d'être un outil parfait et les biais ne sont pas non plus négligeables (Aymes, 1979). Une autre possibilité est de multiplier le nombre de correcteurs, comme cela peut se faire dans certains concours avec notamment le principe de double correction. Pour des raisons évidentes de nature économique, cela n'est pas possible et la recherche de la « vraie note » (au sens statistique du terme), celle qui réduirait au maximum les aléas de correction n'est concrètement pas envisageable (Leclercq, Nicaise, Demeuse, 2004).

Au niveau des examens encore, les procédures d'harmonisation des notes ne limitent que très imparfaitement les biais puisqu'elles ne portent que sur l'échelle globale de notation (moyenne et dispersion de la distribution) et n'agissent aucunement sur les écarts entre les correcteurs. Une solution pourrait être, quand la nature de l'épreuve le permet, de recourir à des formes de type Q.C.M. pour lesquelles les marges d'interprétation des réponses sont nulles. Le coût en temps est plus élevé au niveau de la construction des épreuves que pour une forme traditionnelle, mais les gains en termes d'équité de la notation sont significatifs, de même que ceux relatifs au temps alloué à la correction. Bien sûr certaines disciplines se prêtent moins que d'autres à cette forme d'évaluation, sachant par ailleurs que les épreuves orales n'échappent pas non plus aux biais d'évaluation.

Un deuxième niveau de réflexion concerne le baccalauréat lui-même. Au début du XXI^{ème} siècle, 200 ans après sa création, la question de la pertinence de cet examen emblématique se pose et des interrogations sur sa valeur, son coût et son organisation sont légitimes (Solaux,

2001). Il est évident, qu'au fil des décennies, le bac a perdu de son intérêt en terme de sélectivité et même si les différentes séries du bac ne se valent pas toutes pour la poursuite d'études ultérieures, sa dévalorisation est néanmoins évidente (Duru-Bellat, 2006). Un autre aspect est son coût élevé⁷ à la fois monétaire et en temps pour les élèves avec la mobilisation des locaux scolaires pour les centres d'examen. Il serait aussi possible d'envisager des épreuves plus légères possédant le même caractère prédictif de la réussite, mais moins coûteuses en matière d'organisation (Oget, 1999), mais le problème des aléas de la notation ne serait pas davantage résolu. C'est donc clairement la question de la suppression de l'examen qui peut se poser sachant que son remplacement par le contrôle continu n'est pas la panacée, celui-ci étant très marqué par le contexte local au niveau de l'établissement.

Enfin, un dernier niveau de réflexion concerne plus généralement la pratique de la notation au quotidien en soulignant que cette pratique évaluative n'existe pas dans un certain nombre de pays. L'expérimentation présentée et analysée dans ce texte, ne fait que confirmer avec des données actuelles, ce que l'on sait depuis longtemps déjà sur l'incertitude de la notation. On ne peut que regretter que cette pratique, si répandue à tous les échelons de notre système éducatif, puisse avoir des conséquences en termes d'inégalités dans la certification des acquis scolaires. Faire le deuil de la notation, renvoie aussi à changer plus globalement la vision de la finalité de l'acte d'enseignement. Un changement en la matière obligerait à revoir totalement les mécanismes de sélection, d'orientation et de certification des élèves, mais aussi, au quotidien, le regard que porte l'enseignant sur l'élève.

⁷ Il est difficile d'avoir une estimation précise du coût du bac. Les chiffres disponibles varient selon les éléments intégrés dans les calculs. Pour l'année 2002 le coût moyen par élève présent au baccalauréat général s'élève à 44,6 euros. Ce coût étant établi sur la base des dépenses d'indemnités de jury et de remboursement des frais de déplacement (hors frais d'organisation matérielle dont certaines dépenses sont affectées à un ensemble d'examens). Données du SISEC.

Bibliographie

Aymes J. (1979). Une expérience de multicorrection. *Bulletin de l'Association des professeurs de mathématiques de l'enseignement public*, n° 321.

Bonniol J.J. (1965). Les divergences de notation tenant aux effets d'ordre de la correction ». *Cahiers de psychologie*, N°8, pp.181-188.

Bonniol J.J., Piolat M. (1971). Comparaison des effets d'ancrage obtenus dans une tâche d'évaluation. Expérience de multi-correction en mathématique et en anglais, in *Actes du XVIIe Congrès international de psychologie appliquée*, 8, pp. 1179-1189

Bonniol, J-J., Caverni, J-P., Noizet, G. (1972). Le statut scolaire des élèves comme déterminant de l'évaluation des devoirs qu'ils produisent. *Cahiers de psychologie*, N°15, pp.83-92

Caverni, J.-P., Fabre, J.-M., Noizet, G. (1975). Dépendance des évaluations scolaires par rapport à des évaluations antérieures: études en situation simulée. *Le Travail Humain*, 38, pp. 213-222.

Chatel E. (1994). Qu'est-ce qu'une note : recherche sur la pluralité des modes d'éducation et d'évaluation. *Les dossiers d'Education et Formations*, 47, 1994, pp. 183-200.

Duru-Bellat M. (2006). *L'inflation scolaire : les désillusions de la méritocratie*, La république des idées, Seuil, Paris. 106 p.

Duru-Bellat M., Mingat A. (1993). *Pour une approche analytique du fonctionnement du système éducatif*. PUF, 192 p.

Hartog P., Rhodes E.C. (1935). *An examination of examinations*. (International Institute Examinations Inquiry.) London and New York, Macmillan, 1935. 81 p.

Murat F. (1998). Les différentes façons d'évaluer le niveau des élèves en fin de collège, *Education et Formation*, 53, pp.35-49.

Jarousse J.P., Leroy-Audouin C. (1997). *Analyse du fonctionnement du premier cycle de l'enseignement fondamental au Maroc*. Rapport pour l'Unesco. Irédu 1997. 260p.

Jarousse J.P., Suchaut B. (2002). *Notation, orientation et abandons dans l'enseignement fondamental en Mauritanie : le cas de la 2ème année de l'enseignement fondamental*. Rapport

pour le Ministère de l'Education nationale mauritanien, projet Education-Formation, mai 2002. 42 p.

Jouvanceau P. (1989), “ L'épreuve anticipée du baccalauréat et l'évaluation ”, *Education et Formation*, 18, pp.13-21.

Laugier H., Weinberg D. (1936). *Commission française pour l'enquête Carnegie sur les examens et concours. La correction écrites des épreuves du baccalauréat*. Paris, Maison du livre.

Leclercq D., Nicaise J., Demeuse M. (2004). Docimologie critique : des difficultés de noter des copies et d'attribuer des notes aux élèves. In *Introduction aux théories et aux méthodes de la mesure en sciences psychologiques et en sciences de l'éducation*. Liège : Les Editions de l'Université de Liège, pp. 273-292.

Merle P. (1996). *L'évaluation des élèves. Enquête sur le jugement professoral*, PUF, 323 p.

Merle P. (1998). *Sociologie de l'évaluation scolaire*, PUF. 128 p.

Mingat A., Rakotomalala R., Suchaut B. (1999). *Une analyse du fonctionnement de l'école primaire au Gabon : acquisitions et carrières scolaires des élèves*. Pour le ministère français des Affaires étrangères et pour le ministère de l'Education nationale du Gabon. Décembre 1999, 95 p.

Murat F. (1998). Les différentes façons d'évaluer le niveau des élèves en fin de collège. *Education et Formation*, 53, pp.35-49.

Nilson G., Nias L. (1977). *Le charme a ses raisons*. Paris. Tchou.

Noizet G. (1961). Etude docimologique sur la correction de l'écrit du baccalauréat. *Bulletin de l'Institut national d'étude du travail et d'orientation professionnelle*. N°4 pp. 257-267.

Noizet G., Caverni J.P. (1978). *Psychologie de l'évaluation scolaire*, PUF, 230 p.

Noizet G., Caverni J.P. (1985). La notation en classe terminale et au baccalauréat : données monographiques et effets d'intervention expérimentale. *Les sciences de l'éducation*. Octobre-décembre.

Oget D. (1999). *Efficacité et coût du baccalauréat général et technologique : quelle alternative à l'organisation des épreuves ?* Thèse de doctorat, université de Bourgogne, 1999.

Perrenoud P. (1991), Pour une approche pragmatique de l'évaluation formative, *Mesure et évaluation en éducation*, 4 (13), pp. 49-81.

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

Pourtois J.P. (1978). Le niveau d'expectation de l'examineur est-il influencé par l'appartenance sociale de l'enfant ? *Revue française de pédagogie*. 44, pp 34-37.

Rosenthal R. A., Jacobson L. (1968). *Pygmalion in the classroom : teachers expectations and pupil's intellectual development*. New York : Holt, Rinehart and Winston

Simler B. (1985). *Comportement de notation au baccalauréat. Une étude comparée dans trois académies : Lille, Nancy, Poitiers*. Etude ronéotypée.

Solaux G. (2001). *Le Bac*. Le cavalier bleu éditions. Collection idées reçues économie et société.

Spear M.G. (1989). Sex Bias in Science Teachers' Rating of Work and Pupils Characteristics. *Euroean Journal of Science Education*. 6, 4. pp. 369-377.

Weiss R. (1969). *The reliability of the number marking system : An Austrian study*. In (J. Lauwerys & D. Scalon eds.). *Examinations*. London : Evans Brothers. pp. 101-107.