Une expérience docimologique

par Claire DUPUIS, François PLUVINAGE, IREM de Strasbourg.

Le présent article a son origine dans la reprise à Strasbourg de l'expérience de docimologie présentée dans le Bulletin A.P.M. n° 300 (septembre 75, p. 517). C'est un groupe de stagiaires IREM, animé par les auteurs de l'article, qui s'est chargé de ce travail.

1. Ecarts de correction et docimologie

Notre propos est de montrer dans un cas particulier que le *problème docimologique* est susceptible d'être un *faux problème*. Qu'est-ce à dire ? Nous allons donner une brève explication, avant de relater l'expérience docimologique qui s'avéra révélatrice.

Une épreuve d'examen est un moyen de repérer, à l'issue d'un apprentissage donné, soit des lacunes (cas de la dictée par exemple), soit des acquis (cas général des épreuves de mathématiques). La docimologie s'intéresse aux variations individuelles d'appréciation de ces lacunes ou de ces acquis sur des copies d'élèves données. Par exemple, on constate que les fautes d'orthographe "oubliées" par un instituteur dans sa classe se répartissent inégalement ; les oublis sont plus nombreux pour les "bons" élèves que pour les "mauvais".

Mais supposons qu'une épreuve laisse subsister l'incertitude la plus complète sur les lacunes ou les acquis de la plupart des élèves d'un niveau donné. Le correcteur, qui est *obligé* de noter toute copie, n'a d'autre solution que de procéder à une extrapolation hardie, reposant sur des indices très minces, de comportements dont il a eu (ou cru avoir) l'expérience en tant qu'enseignant. C'est-à-dire qu'il est conduit, en l'absence de véritable élément d'appréciation, à se rabattre sur un comportement qu'il estime moyen.

Et *paradoxalement*, l'*écart des notes* ainsi attribuées par plusieurs correcteurs (ou un même correcteur à des temps différents) peut alors être *inférieur à l'incertitude* que l'épreuve laisse subsister sur les lacunes ou les acquis des élèves interrogés. Or, il paraît clair que la docimologie doit précisément éliminer des cas de ce genre : des cas limites (par exemple distribution d'un énoncé incorrect) conduisent même à *ne pas corriger*, mais à faire recommencer l'épreuve.

2. Le déroulement de l'expérience

Le Bulletin n° 300 de l'A.P.M. rendait compte d'une expérience de docimologie faite par un groupe de l'I.R.E.M. de Toulouse. Nous n'y avions pas cru tout à fait. Peut-être nos collègues toulousains ou montalbanais n'avaient-ils pas suffisamment discuté du barème, des diverses solutions possibles, de la manière de tenir compte des erreurs prévisibles. Si l'on opérait exactement dans les conditions de la correction au B.E.P.C., l'écart serait-il moindre ? De plus, nous avions à Strasbourg des copies d'élèves de Seconde et Première à cause de la date de l'expérience (1er trimestre), ce qui risquait encore de réduire les écarts. Voici ce qui s'est passé.

Premier temps : l'épreuve de géométrie

Après discussion, le barème est subdivisé : ainsi l'attribution des 5 points de la deuxième question est partagée entre calculs de distances et résultat sur le triangle (O, B, C). Pour ce triangle, c'est surtout le fait qu'il soit rectangle qui est à prendre en compte. La troisième question admet plusieurs solutions qui sont examinées en détail. Et ainsi de suite...

De plus, les correcteurs auront à ajouter à la simple application du barème sur dix copies un travail d'appréciation de l'écart des copies deux à deux (voir annexe 1). Le fait de confronter ainsi systématiquement les copies entre elles risque d'améliorer l'appréciation . Par ailleurs, nous espérions, par application d'une méthode statistique (la méthode INDSCAL), faire des observations intéressantes à partir des différences entre correcteurs lors de ces confrontations des copies deux à deux.

Aux résultats ! Nous ne retrouvons pas à Strasbour le correcteur languedocien "P 2" (qui notait nettement en dessous de ses collègues), mais, pour le reste, *les résultats sont exactement les mêmes*. Ainsi, le plus grand écart relevé à Toulouse est de 11 points, il devient de 7 points si l'on "supprime" P 2 : notre écart maximum est de 7 points. Pour notre écart minimum, il est de 2 points (1 point pour Toulouse). Pas plus de possibilité d'équilibrer en ramenant tous les correcteurs à la même moyenne et au même écart type.

L'application de la méthode INDSCAL se révèle plutôt décevante : elle fait essentiellement apparaître que deux d'entre nous ont mal respecté la consigne de mettre entre deux copies, sur la

même question, une note d'écart au moins égale à la différence de leurs notes-barèmes. Ce manquement à la consigne, qui ne résulte pas que de questions d'arrondi (demi-point au point inférieur ou supérieur), est peut-être révélateur de sentiments sur la signification des notes. Il semble par ailleurs que, pour être vraiment interprétables, les écarts devraient être repérés non sur toute l'épreuve mais sur chaque question en particulier.

Deuxième temps : l'épreuve d'algèbre (voir annexe 2)

Nous n'allons pas recommencer comme pour la géométrie ! Nous allons faire confiance aux méridionaux. Il s'agit donc de trouver comment améliorer la fiabilité de la correction par une procédure qu'il serait facile de mettre en oeuvre pour les examens réels. Notre expérience des enquêtes nous a appris que les codages de réponses sont surs. Alors allons-y !

- Prélèvement d'un échantillon de copies (nous en prenons 8)
- Détermination pour chaque question d'un *codage disjonctif total* (1) (éventuellement en découpant en plusieurs rubriques)
- Détermination d'un barème associé automatiquement au codage.
- Correction du paquet de copies selon ce barème.

Simple !

Voici un résumé sommaire du déroulement de la procédure. A la fin d'une séance, les correcteurs prennent connaissance des huit copies de l'échantillon et en discutent par groupes, pour voir comment résumer les contenus des réponses. La séance suivante, le codage de la question 1 finit par être explicité comme suit, après environ 40 minutes.

(1) Chaque question ou rubrique donne lieu pour chaque élève à l'attribution d'un unique 1 et d'un certain nombre de zéros. C'est la place du 1 qui différencie les réponses. Ainsi, pour une question notée sur trois colonnes, il y a trois "notes" possibles : 001, 010 et 100. Dans la pratique, nous les désignons par 0, 1 et 2, la conversion en "0 et 1" étant effectuée en machine ; ceci pour des raisons de commodité.

Codage de la première question (trois rubriques)

Conformité à la consigne	Identité $(a+b)^2 = \ldots$	Coefficients et signes
0 Non-réponse 1 Non respect de la consigne (pas d'essai de développement) 2 Obtention d'un polynome développé, mais non réduit et ordonné. 3 Obtention d'un polynome réduit et ordonné	0 Non utilisée 1 Utilisation d'une identité incorrecte 2 Identités correctes	0 Pas de calcul 1 Erreurs de distribution de coefficient et de signe 2 Erreurs sur la distributivité seule 3 Erreurs de signe seules 4 Calcul sans erreur

Ainsi, celui qui n'a pas répondu à la question sera codé 000, l'élève qui a répondu correctement 324.

Joli !

Avant de poursuivre, nous décidons de vérifier la pertinence de notre codage, en demandant à chacun des membres du groupe de l'appliquer aux huit copies de l'échantillon et en confrontant les résultats. Les copies 1, 2, 3, 4, 6 et 7 sont codées de façon identique par tous, mais pas les copies 5 et 8. La difficulté sur la copie 5 provient de ce que l'élève a écrit

$$f(x) = 4x^2 - 4 - x^2 + 1$$

sans autre explication. A une erreur évidente sur l'identité, s'ajoute-t-il une erreur de signe pour expliquer le $-x^2 + 1$? La réponse est *probablement* affirmative. Mais ce n'est absolument pas certain car à la deuxième question, des élèves remplacent $(x-1)^2$ par $(x+1)(x-1)$ (pour pouvoir mettre en facteur) ; donc, on ne peut pas exclure complètement la possibilité : $(x+1)^2 = x^2 - 1$. Dans un tel cas, il est nécessaire de prendre pour le codage une décision d'autorité : on comptera "erreur de signe". A juste titre, deux membres du groupe sont réticents, mais n'est-on pas amené à faire de même pour fixer certaines notes ?

La copie 8 s'annonce plus mal. Au bout de dix minutes de discussion, qui amènent à préciser la rubrique "conformité à la consigne", une voix timide avance :

"Et s'il avait seulement répondu à la *deuxième* question ? "

Voici quel était le contenu de la copie sur cette question :

1) $f(x) = 4(x-1)^2 - (x+1)^2$

$f(x) = 4(x-1+x+1)(x-1-x-1)$

$f(x) = 4(2x)(-2) = (8x)(-2)$

L'élève a utilisé une mise en facteur. Par erreur, il aboutit à un polynome du premier degré. Il est difficile de dire s'il a cru ainsi répondre à la première question, ou si le "1)" n'est que de pure forme et si le travail commence sur la seconde question. Mais le fait est là : *19 correcteurs réunis peuvent ne pas arriver à savoir, sur cette épreuve, si un élève a répondu à telle question ou à telle autre* !

Enorme !

Après ce coup de $\left\{\begin{matrix}\text{maître}\\ \text{théâtre}\end{matrix}\right.$ (*), la question 2 se révéla elle aussi croustillante. L'auteur de l'énoncé voulait sans doute savoir si les élèves étaient capables de percevoir une différence de carrés dans l'expression

$$4(x-1)^2 - (x+1)^2$$

et d'appliquer l'identité $A^2 - B^2 = (A-B)(A+B)$. Un certain nombre d'élèves perçoivent qu'il y a "du $A^2 - B^2$ là-dessous". Mais, comme il intervient aussi l'identité $A^2B^2 = (AB)^2$, une difficulté supplémentaire s'introduit. De ce fait, il se produit un certain nombre de cas de "détournement de difficulté", dont le plus courant semble être de mettre 4 en facteur commun. Et le fait pour la plupart des élèves de prendre ensuite en compte le résultat qu'ils viennent d'obtenir conduit à un *échec complet* sur la suite du problème.

Conclusion de l'expérience

Autrement dit, sur cette épreuve, la plupart des élèves qui n'auront pas "vu" ici l'utilisation simultanée des deux identités concernant $A^2 - B^2$ et A^2B^2 n'auront livré aux correcteurs aucune information précise sur leurs acquis. Et que dire des élèves qui ont essayé de résoudre la deuxième question en partant du résultat trouvé à la première (l'énoncé n'avertissait pas de l'indépendance) ?

(*) Rayer l'un des deux mots au choix.

On comprend dans ces conditions que les correcteurs se trouvent devant une délicate situation de notation : la seule attitude "docimologiquement" valable consiste (puisqu'il n'y a pas d'information précise sur la plupart des cas autres que la réussite complète) à attribuer la note 0 à toute réponse entâchée d'erreur. Mais on imagine alors la catastrophe. Le problème docimologique n'apparaît donc que dans la mesure où l'on veut éviter cette catastrophe en appréciant des acquis "intermédiaires", qui sont en fait inappréciables sur cette épreuve.

3. Analyse de l'énoncé d'algèbre

a) *Liste des acquis isolés nécessaires à la résolution du problème.*

Connaissance, reconnaissance et utilisation des identités :

(1) $(a + b)^2 = a^2 + 2ab + b^2$

(2) $(a - b)^2 = a^2 - 2ab + b^2$

(3) $a^2 - b^2 = (a - b)(a + b)$

(4) $a^2 b^2 = (ab)^2$

(5) $\frac{ca}{cb} = \frac{a}{b}$ si $c \neq 0$

(6) Calcul polynomial : degré d'un monôme, réduction de monômes de même degré, écriture ordonnée d'un polynome.

(7) Pratique de la mise en facteur, d'un coefficient et d'un polynome. Suppression de parenthèses précédées du signe — .

(8) Fonctions : substitution d'une valeur à la variable.

(9) Résolution d'une équation du premier degré.

(10) Simplification de l'égalité de deux rapports (produits des extrêmes et des moyens).

(11) Calculs comportant des radicaux.

b) *Articulation du problème*

— la question 1 ne sert à rien pour la suite du problème,

— la première partie de la question 2 est susceptible d'être utilisée dans toute la suite du problème, alors que la deuxième partie de cette même question ne sert à rien dans la suite, sauf éventuellement au b) de la question 4,

— la question 3 nécessite le premier résultat de la question 2,

— la question 4 comporte une première partie qui peut être résolue soit directement, soit en utilisant la forme réduite. La résolution directe demande, en plus des acquis indiqués ci-dessus, de savoir que

$$(A^2 = B^2) \Leftrightarrow (|A| = |B|),$$

la deuxième partie de la question 4 demande pratiquement l'utilisation de la forme simplifiée (ou l'utilisation de la deuxième partie de la question 2),

— la question 5 peut être traitée sous les deux formes (h(x) simplifié ou non). Evidemment le calcul sous la forme simplifiée est beaucoup plus simple.

c) *Acquis nécessaires à la résolution des diverses questions* (les numéros entre parenthèses renvoient à la liste ci-dessus).

Question 1. (1), (2), (7) et (6) à la fois.

Question 2. Première partie : (3), (4) et (7) à la fois.
Deuxième partie : (7) et résultat de la première partie (Notre sujet tiré du Bulletin APM n° 300 ne comportait pas g(x)).

Question 3. (5) et premier résultat de la question 2.

Question 4. a) (9) et premier résultat de la question 2
ou (9) et $[(A^2 = B^2) \Leftrightarrow (|A| = |B|)]$
b) (9), (10) et premier résultat de la question 2.

Question 5. (8), (11) et (3) (dans le sens :
$(a - b)(a + b) = a^2 - b^2$).

4. Conclusions

a) *Ce que nous aurions dû faire*

C'est *avant* de corriger les copies que nous aurions dû nous livrer à l'analyse sommaire indiquée ci-dessus. Connaissant la tendance des élèves à utiliser dans une question les résultats qu'ils viennent d'obtenir, nous aurions compris qu'il nous serait, sur cette épreuve, *impossible* d'apprécier les acquis des élèves qui n'auraient pas réussi à utiliser simultanément (3), (4) et (7).

b) *Une comparaison intéressante*

Le problème de BEPC donné en 1975 dans l'académie de Toulouse (voir annexe 3) présente avec le sujet étudié ici de grandes similitudes. Il n'est d'ailleurs pas certain que ces simili-

tudes soient autre chose que le produit du hasard (et du peu de variété de la quasi-totalité des sujets de BEPC) ; de toute façon, les variations sont suffisantes pour que l'honnêteté de l'examen soit respectée.

C'est d'ailleurs justement l'existence de ces variations qui conduit à une étude intéressante, à la lumière de notre expérience. En gros, on peut dire que l'épreuve de BEPC est certainement *beaucoup plus corrigeable* que le sujet étudié ici. Voyons pourquoi :

— Le sujet de BEPC propose

$$g(x) = (3x - 1)^2 - 4(x - 1)^2$$

Le coefficient 4 affecte ainsi le *second* terme au lieu du premier. Cette modification d'apparence anodine risque d'être importante au vu de notre correction.

— On demande de calculer f(0), f(1/2) et f(−1), c'est-à-dire des valeurs prises par une fonction polynome, alors que notre sujet abordait cette question directement pour une fraction rationnelle.

— Le résultat de la simplification de h(x) est *donné* (au passage, notons que le sujet de BEPC omet d'indiquer que D désigne l'ensemble de définition). Il est ainsi possible de tester les acquis d'élèves qui auraient commis une faute dans la question 2.

— A la fin du problème est ajoutée une question de représentation graphique, *indépendante* de ce qui précède.

Sommes-nous en droit de conclure que les constatations explicitées ici sont ressenties par nos collègues, et mises en pratique au moins quand il s'agit d'un examen réel ? Sans doute pas toujours (voir le Bulletin A.P.M. n° 297, p. 88).

c) *Quelques questions et quelques remarques*

Après l'expérience, et à la lecture de ce compte rendu, quelques réflexions ont été proposées par des membres du groupe de travail. Nous les présentons à la fin de cet article (la sagesse populaire dirait : "A quelque chose, malheur est bon").

Sur l'élaboration des sujets sous leur forme actuelle

— La confection des sujets ne pourrait que gagner en sérieux, si l'on ne faisait preuve d'une méfiance quelque peu déplacée envers les professeurs (en demandant 20 sujets pour n'en retenir qu'un).

— Un gain encore plus substantiel pourrait résulter du *test* auprès d'élèves de parties de sujets (bien entendu avec modification de données).

Nota : C'est possible pour le BEPC. Au niveau du Baccalauréat, la plus grande variété des sujets est peut-être un obstacle à cette procédure.

— Une autre possibilité pour gagner en signification est de se livrer, une fois le sujet posé, à *l'analyse des acquis* nécessaires à sa résolution. Au vu de cette analyse, des *retouches* du sujet pourraient être envisagées.

— Subsisterait-il des problèmes docimologiques, et lesquels, sur la correction d'une épreuve méthodologiquement correcte ?

Sur les types de sujets

— Au vu des sujets actuels de BEPC, on se demande pourquoi on n'envisage pas des questionnaires à choix multiples pour cette épreuve. Un avantage de ce type de questionnaires est qu'il peut être élaboré selon une *méthodologie prédéterminée*.

— La question est alors posée de savoir, plus généralement, quel type d'énoncé (problème, séquence de questions ouvertes, semi-ouvertes, à choix multiples) est le mieux adapté à un examen ou concours déterminé : BEPC, ex-Concours des Ecoles Normales, Baccalauréat (pour les différentes séries), etc...

N'oublions pas que l'élément premier et qui conditionne toute la suite des opérations est et sera toujours le *sujet proposé*. Toute faute de méthode dans l'élaboration et la rédaction du sujet est *irrattrapable*. Nous disons bien "de méthode", car on trouve souvent des critiques, portant sur des questions de correction d'écriture ou d'utilisation de terminologie, *non pertinentes vis à vis des réactions des candidats*. Comme le dit Vernon : "... Enfin, les critiques peuvent lire dans les questions des choses qui ne viennent même pas à l'esprit d'élèves intelligents" (cité par Landsheere : Evaluation continue et examens. Précis de docimologie — Labor et F. Nathan, 1975, page 100). Ce qui importe, ce n'est donc pas de satisfaire les désirs de critiques pointilleux, mais avant tout d'indiquer des procédures permettant :

1° de disposer de quelques *points de départ* pour l'élaboration,

2° de détecter et donc *d'éliminer* des défauts rédhibitoires, quant à la signification qui pourra être attachée ultérieurement aux résultats.

Les outils suivants, mentionnés par Landsheere, dans l'ouvrage cité ci-dessus, peuvent être des auxiliaires précieux.

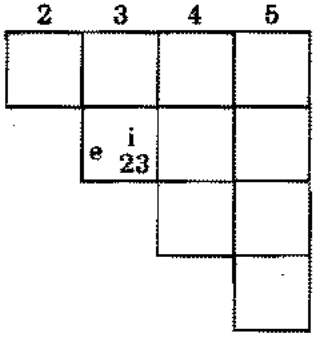
Pour le premier point (idées initiales d'élaboration), le modèle de Guilford risque de se révéler fructueux à l'utilisation.

Pour le second point, la classification NLSMA est certainement intéressante, surtout si on lui adjoint une liste d'acquis nécessaires à la résolution, comme celle que nous avons dressée ici.

Addendum

A la lecture de cet article, Marie-Claire Dauvisis, auteur avec Jean Cransac de l'article paru dans le Bulletin A.P.M. n° 300, a d'une part proposé quelques modificiations et précisions qui ont été prises en compte. D'autre part, elle n'est pas aussi stricte que nous ne le sommes dans l'introduction, vis à vis de l'acception du terme "docimologie" : pour elle, le problème de l'extrapolation à partir d'observations trop réduites *est un problème docimologique.* Autrement dit, on peut s'intéresser aux écarts de correction *même* pour des épreuves qui apportent peu de renseignements sur les acquis des élèves.

Annexe 1. Utilisation de la méthode INDSCAL



Une hypothèse qui peut paraître intéressante est la suivante : Devant une réponse à une question, tous les correcteurs observent les *mêmes phénomènes*, mais les font intervenir dans leur notation avec des *pondérations différentes.* Ainsi, l'un pourrait être plus sensible à l'obtention du résultat, un autre à l'exactitude des raisonnements ; de même l'exécution de calculs pourrait intervenir dans la cotation avec une importance plus ou moins grande selon les correcteurs.

Dans ces conditions, supposons que l'on demande à chaque correcteur de remplir un tableau d'écarts estimés, comme celui que montre la figure pour le cas de cinq copies. Par exemple, l'écart e_{23}^{i} placé dans le tableau sera l'écart estimé par le $i^{ème}$ correcteur entre les copies 2 et 3. Donner une estimation significative de l'écart entre deux copies sur tout un problème paraît hasardeux, mais en opérant question par question, on obtient un repérage plus sûr. Le principe de cette estimation est le suivant : soit n la note attribuée par le barème à la question envisagée. Si un correcteur estime que deux copies sont totalement dissemblables, sur cette question, il leur attribuera un écart de n. Si au contraire il les estime tout à fait semblables, il leur attribuera un écart de 0. S'il a déjà noté les copies, il devrait en principe toujours leur attribuer un écart au moins égal à l'écart des notes (sur la question considérée).

Ensuite, nous collecterons les tableaux ainsi remplis par les correcteurs sur les mêmes copies. L'examen de cette famille de tableaux à l'aide de la méthode INDSCAL permet :

1° de *détecter* les éléments d'appréciation multiples qui ont pu intervenir chez les correcteurs, sous l'hypothèse indiquée au début de cette annexe,

2° *d'apprécier* le degré avec lequel la famille des tableaux remplis par les correcteurs satisfait à cette hypothèse.

Autrement dit, la méthode essaye d'abord de satisfaire le mieux possible à l'hypothèse à partir des résultats fournis ; ensuite elle indique le degré de satisfaction ainsi obtenu avec l'hypothèse.

Annexe 2.

ALGEBRE

Soit les fonctions polynomes f et g définies dans R par :

$f(x) = 4(x-1)^2 - (x+1)^2$ et $g(x) = (2x+5)(x-3) - (x-3)^2$

1° Développer, réduire et ordonner f(x) (3 pts)

2° Ecrire f(x), g(x) et $11f(x) - 8g(x)$ sous forme de produits de facteurs du premier degré. (Respectivement : (2 Pts), (2 Pts), (3 Pts).)

3° Soit la fonction rationnelle h, définie dans R par :

$$h(x) = \frac{f(x)}{g(x)} \text{ ; simplifier } h(x) \quad (2 \text{ Pts})$$

4° Déterminer l'ensemble des réels tels que :

a) $h(x) = 0$ (2 Pts) b) $h(x) = \frac{8}{11}$ (4 Pts)

5° Calculer $h(6\sqrt{2})$ (2 Pts)

Annexe 3

Toulouse

① On considère les fonctions polynomes f et g de R dans R définies par :

$f(x) = (2x-1)^2 - (2x-1)(x-2)$

$g(x) = (3x-1)^2 - 4(x-1)^2$

1. Développer, réduire et ordonner f(x).

2. a) Mettre f(x) et g(x) sous forme de produits de facteurs du premier degré.

b) Calculer $f(0)$, $f\left(\frac{1}{2}\right)$, $f(-1)$.

3. Résoudre dans R l'équation $f(x) = g(x)$

4. Soit h la fonction rationnelle de R dans R définie par :

$$h(x) = \frac{f(x)}{g(x)}$$

a) Déterminer son ensemble de définition.

b) Démontrer que pour tout x élément de D

$$h(x) = \frac{2x-1}{5x-3}$$

c) Calculer $h(\sqrt{3})$. Donner le résultat sous forme d'un quotient ayant un dénominateur entier.

5. Soit un plan P muni d'un repère orthonormé $(O, \vec{i}, \vec{j})$.
(On pourra prendre pour unité 3 cm).

Dans ce repère, A et B sont les représentations graphiques respectives des fonctions a et b de $\mathbf{R}$ dans $\mathbf{R}$ définies par

$$a(x) = 2x - 1$$
$$b(x) = 5x - 3$$

a) Déterminer les coordonnées du point M tel que $A \cap B = \{M\}$

b) Construire A et B.