L'indice de certitude : un concept au service de l'apprentissage dans l'évaluation formative de l'élève.

Pierre-François COEN et René DENERVAUD

Haute école pédagogique de Fribourg (Suisse) — 2003

Introduction

Le débat sur l'évaluation n'est pas nouveau. Nombre d'auteurs ont, au cours de ces dernières années, mis l'accent sur les limites des procédures d'évaluations actuelles ou sur les difficultés de pratiquer une évaluation la plus rigoureuse possible. Ils ont tantôt mis l'accent sur les biais de l'évaluation (ROSENTHAL), sur l'influence des points de vue de l'évaluateur (ALLAL, CARDINET, PERRENOUD), sur la difficulté de la saisie et du traitement des données recueillies (HUBERMANN & MILES, De KETELE) ou même sur les enjeux complexes de la communication des informations et des décisions prises (HOUSSAYE, WEISS). En outre, l'édumétrie s'est penchée depuis longtemps sur cette question et plusieurs auteurs (HOFFMAN, 1964; GOSLIN, 1966; CRONBACH, 1950) se sont attachés à mettre en cause la fidélité de la réponse et par là même la validité de l'évaluation.

Or les doutes et les mises en garde des scientifiques du domaine laissent parfois perplexes les praticiens confrontés jour après jour à leur pratique d'évaluation. Une réflexion dans ce domaine nous semble dès lors opportune dans la mesure où elle pourrait s'inscrire dans la mise en place de nouvelles pratiques pédagogiques : interdisciplinarité, meilleures articulations entre théorie et pratique, intégration des technologies ou nouveaux moyens et dispositifs d'évaluation.

Dans cette perspective, l'évaluation doit vraiment apparaître comme un outil au service des apprentissages, même si elle n'échappe pas à son rôle certificatif notamment en fin de formation. Le but de cette étude est donc de lui donner une dimension supplémentaire en demandant aux apprenants d'y participer activement. En ce sens, cette pratique n'a rien de révolutionnaire et s'apparente d'une certaine manière à la mise en place d'un dispositif incluant l'autoévaluation de l'élève dans le "compte" de la note. Notre but est d'apporter du même coup, une validité plus grande à l'évaluation en y

associant obligatoirement le jugement de l'élève par son regard porté sur sa la qualité de sa propre réponse. Cela amène, à terme, à une prise de conscience plus grande des acquis et des lacunes des étudiants et cela peut modifier la manière d'envisager les apprentissages.

Aperçu théorique

A la suite de LECLERCQ (1993), et GILLES (1996), nous avons travaillé la théorie de la décision, le concept de réalisme et d'indice de certitude pour construire un dispositif susceptible d'engager plus intensément les apprenants dans le processus d'évaluation. Avant de décrire en détail notre expérimentation, il nous semble nécessaire de définir plusieurs concepts.

- *La mesure du score* : lorsqu'un apprenant répond à une question, il est possible de mesurer sa performance par rapport à des critères d'évaluation précis. Par exemple, un apprenant qui répondra correctement à une question totalisera le maximum de points.
- *La certitude* face à une réponse donnée : lorsque nous demandons à ce même apprenant de porter un jugement de valeur sur la réponse qu'il a produite, nous lui demandons d'exprimer du même coup la certitude qu'il accorde à sa réponse. Celle-ci peut être à 100 % si l'apprenant est tout à fait sûr de sa réponse, mais elle peut très bien diminuer, et cela indépendamment de la réponse elle-même. Ainsi, il sera possible de rencontrer des apprenants tout à fait certains d'une réponse erronée et à l'inverse, des apprenants doutant d'une réponse parfaitement correcte.
- *Le réalisme* est la valeur qui présente le rapport entre les deux paramètres décrits précédemment : performance et certitude. S'il est élevé, cela veut dire que l'apprenant est tout à fait conscient de la justesse de ses réponses, s'il est faible, cela signifie qu'il ne sait pas vraiment ce qu'il sait ou ce qu'il ignore. D'une façon schématique, le réalisme peut donc prendre quatre modalités différentes en fonction de la réponse des étudiants et de la certitude qu'il y accorde. La figure n° 1 présente ces quatre modalités.

		Certitude élevée	Certitude basse
Réponse	correcte	A	B
	fausse	D	C

Figure n° 1 : les quatre modalités du réalisme.

Dans le cas A et C, le réalisme est excellent puisqu'il y a concordance entre le résultat de l'apprenant et la certitude qu'il accorde à sa réponse (certitude forte pour réponse correcte / certitude faible pour réponse erronée). Dans le cas B et D, il y a discordance puisque les deux paramètres ne concordent pas (certitude forte pour réponse erronée / certitude faible pour réponse correcte). Selon nous, il faut encourager les apprenants à avoir un réalisme le plus élevé possible. C'est en effet en travaillant sur cet aspect qu'il est possible d'amener les apprenants à faire une réflexion sur leurs propres apprentissages (voire leurs stratégies d'apprentissage). Nous avons donc construit une première échelle (presque simpliste) dont la finalité était de "pénaliser" les apprenants ayant un faible réalisme. La figure n° 2 présente les valeurs chiffrées pour une échelle allant de 1 à 4.

		Certitude élevée	Certitude basse
Réponse	correcte	4	3
	fausse	1	2

Figure n° 2 : échelle du réalisme sur quatre points.

En termes concrets, un apprenant qui répond correctement à une question totalise 4 points si sa certitude est élevée, mais seulement 3 points en cas de certitude basse. Par ailleurs, un apprenant qui répond faux avec une certitude basse comptabilise 3 points contre un seul en cas de certitude élevée.

Reste à construire une échelle continue entre 1 et 4 tenant compte à la fois du niveau de certitude et de la justesse de la réponse (ces deux valeurs étant normalisés de 0 à 1). Ce faisant, il nous sera permis de traiter tous les types de questions, quel que soit le

nombre de points attribués et le niveau de certitude, exprimé quant à lui, sur une échelle ordinale allant de 1 à 10. Le tableau ci-dessous présente l'ensemble des valeurs que prend le *réalisme* en fonction de la justesse de la réponse et du niveau de certitude de l'élève.

Tableau n° 1 : valeur des scores pondérés en fonction de la performance et du réalisme (normalisés de 0 à 1)

		Performance										
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Certitude	1	1	1.39	1.76	2.11	2.44	2.75	3.04	3.31	3.56	3.79	4
	0.9	1.19	1.56	1.91	2.24	2.55	2.84	3.11	3.36	3.59	3.8	3.99
	0.8	1.36	1.71	2.04	2.35	2.64	2.91	3.16	3.39	3.6	3.79	3.96
	0.7	1.51	1.84	2.15	2.44	2.71	2.96	3.19	3.4	3.59	3.76	3.91
	0.6	1.64	1.95	2.24	2.51	2.76	2.99	3.2	3.39	3.56	3.71	3.84
	0.5	1.75	2.04	2.31	2.56	2.79	3	3.19	3.36	3.51	3.64	3.75
	0.4	1.84	2.11	2.36	2.59	2.8	2.99	3.16	3.31	3.44	3.55	3.64
	0.3	1.91	2.16	2.39	2.6	2.79	2.96	3.11	3.24	3.35	3.44	3.51
	0.2	1.96	2.19	2.4	2.59	2.76	2.91	3.04	3.15	3.24	3.31	3.36
	0.1	1.99	2.2	2.39	2.56	2.71	2.84	2.95	3.04	3.11	3.16	3.19
	0	2	2.19	2.36	2.51	2.64	2.75	2.84	2.91	2.96	2.99	3

Premier dispositif

Nous avons appliqué ce dispositif à trois reprises sur un effectif de 52 sujets de l'Ecole normale cantonale de Fribourg. A chaque question de l'évaluation était associée une échelle (de 1 à 10) sur laquelle l'étudiant devait fixer son degré de certitude. Le résultat de la question était ensuite pondéré par le degré de certitude et nous permettait d'exprimer le réalisme pour chaque question. Le réalisme était soumis à un seuil de maîtrise variable selon la question et les étudiants devaient reprendre obligatoirement toutes les questions situées en dessous du seuil. Cette remédiation obligatoire n'était en rien pénalisante, puisque la note globale était modifiée après la reprise des questions échouées. Les étudiants ont apprécié la méthode en particulier la possibilité de reprendre les questions ratées. Cependant, pour eux, le principal problème était de se représenter aussi bien que possible le réalisme (le résultat final) en fonction du score présumé à la question posée (en fonction de la certitude qu'il y accordait). Les tables étant relativement complexes à mémoriser, les étudiants avaient en fait relativement peu de moyens de se représenter le réalisme qu'ils allaient obtenir. Ainsi par stratégie, la majorité d'entre eux a choisi d'exprimer une certitude élevée estimant faire par là le meilleur calcul pour minimiser une péjoration éventuelle. Les résultats du tableau ci-

dessous font apparaître très clairement une focalisation autour du 0.9 de certitude. Par ailleurs, l'écart-type relativement bas (de 0.065 à 0.071) démontre une faible dispersion des sujets. Les étudiants ont donc peu douté de leurs résultats et ceci malgré le fait que ces derniers ne soient pas toujours bons.

Tableau n° 2 : moyennes et écarts-types de la certitude dans le 1er dispositif.

Descriptive Statistics

	C1T1	C1T2	C1T3
Mean	.877	.896	.916
Std. Dev.	.065	.071	.067
Minimum	.733	.715	.745
Maximum	1.000	1.000	.991

Deuxième dispositif

Une réflexion sur ces résultats nous a conduits à reconsidérer notre échelle et le mode d'attribution de la certitude. Dans un premier temps, nous avons modifié l'échelle de péjoration en reprenant les quatre cas de figure décrits plus haut en les appliquant à l'équation $f(x) = x^3$. Les valeurs 20, 11, 9 et 0 remplacent nos bornes 4, 3, 2, 1. Dans le carré supérieur droit apparaissent les deux valeurs que prend le réalisme si la réponse est correcte et la certitude est haute (20 points) ou basse (11 points); dans le carré inférieur gauche apparaissent les deux valeurs que prend le réalisme si la question est fausse et la certitude est basse (9 points) ou élevée (0 point).

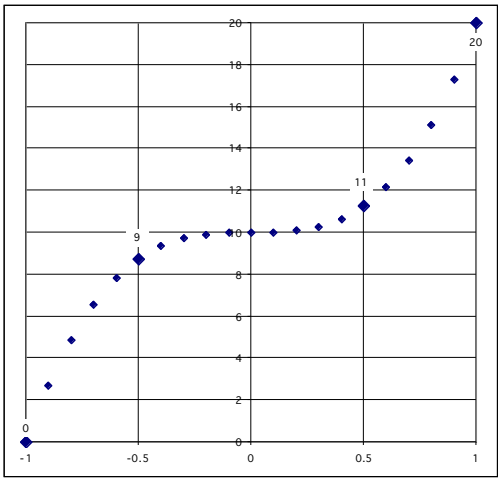

Figure n° 3 : répartition des bornes selon la fonction x^3.

Par ailleurs, pour des raisons pratiques, nous demandons aux étudiants de ne plus exprimer la certitude qu'ils accordent à la question, mais simplement le nombre de points qu'ils pensent avoir obtenus. La certitude se transforme donc en autoévaluation. Dans ce sens, on obtient trois cas de figure possibles : l'apprenant s'évalue très bien (la différence entre sa propre évaluation et celle de l'enseignant est nulle), l'élève se surévalue (si la différence est positive) et se sous-évalue (si elle est négative). Il devient ainsi possible d'exprimer le réalisme de l'apprenant sous une autre forme :

$$f(e,m,\max) = 1 - \frac{|e-m|}{\max}$$

e étant le nombre de points estimé par l'étudiant, *m* étant le nombre de points estimé par l'enseignant et *max* étant le nombre de points maximum attribué à la question. Le petit tableau ci-dessous illustre cette fonction.

Tableau n° 3 : exemple de valeurs que prend la fonction f(e,m,max).

	Cas 1 Bonne réponse Bonne autoévaluation	**Cas 2** Bonne réponse Mauvaise autoévaluation	**Cas 3** Réponse fausse Bonne autoévaluation	**Cas 4** Réponse fausse Mauvaise autoévaluation
e	4	0	0	4
m	4	4	0	0
Max	4	4	4	4
F(e, m, max)	1	0	1	0

Le cas n° 1 et n° 3 présente les variantes où l'apprenant s'évalue bien en réussissant la question (cas n° 1) ou en la ratant (cas n°3), et les deux autres variantes présentent un apprenant qui s'autoévalue mal malgré la réussite (cas n°2) ou l'échec à la question posée.

En résumé, nous disposons maintenant de plusieurs éléments qui concourent à construire le réalisme :

- une nouvelle mesure de la certitude basée sur l'autoévaluation de l'étudiant.
- une nouvelle échelle dont les bornes sont plus étendues (0 à 20) et plus péjorantes.

Le calcul du réalisme reste quant à lui inchangé. Le tableau ci-dessous en présente les nouvelles valeurs en fonction des nouvelles bornes.

Tableau n° 4 : valeur des scores pondérés en fonction du réalisme en utilisant les bornes (20, 11, 9 et 0).

		Performance										
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Réalisme	0	**0**	1.91	3.64	5.19	6.56	7.75	8.76	9.59	10.2	10.7	**11**
	0.1	0.09	2	3.91	5.64	7.19	8.56	9.75	10.8	11.6	12.2	12.7
	0.2	0.36	2.09	4	5.91	7.64	9.19	10.6	11.8	12.8	13.6	14.2
	0.3	0.81	2.36	4.09	6	7.91	9.64	11.2	12.6	13.8	14.8	15.6
	0.4	1.44	2.81	4.36	6.09	8	9.91	11.6	13.2	14.6	15.8	16.8
	0.5	2.25	3.44	4.81	6.36	8.09	10	11.9	13.6	15.2	16.6	17.8
	0.6	3.24	4.25	5.44	6.81	8.36	10.1	12	13.9	15.6	17.2	18.6
	0.7	4.41	5.24	6.25	7.44	8.81	10.4	12.1	14	15.9	17.6	19.2
	0.8	5.76	6.41	7.24	8.25	9.44	10.8	12.4	14.1	16	17.9	19.6
	0.9	7.29	7.76	8.41	9.24	10.3	11.4	12.8	14.4	16.1	18	19.9
	1	**9**	9.29	9.76	10.4	11.2	12.3	13.4	14.8	16.4	18.1	**20**

Les principaux avantages que nous voyons à la chose reposent 1° sur le fait que les étudiants parviennent mieux à exprimer leur certitude (sous forme d'autoévaluation); des calculs spéculatifs et hasardeux sont inutiles et superflus et 2° l'échelle de 0 à 20 présente des valeurs plus dispersées et nettement plus dissuasives ainsi les étudiants sont amenés à s'autoévaluer avec le plus grand soin possible pour éviter d'être trop sévèrement pénalisés.

En conclusion, nous pensons que la démarche qui consiste à demander la certitude accordée à une réponse présente un grand intérêt au niveau formatif. Si l'on inscrit la nécessité de revenir sur ces processus d'apprentissage dans le contexte de l'évaluation formative, l'utilisation de l'indice de certitude amène l'apprenant à une faire une réflexion sur le produit demandé (la réponse à fournir), mais également sur sa manière d'apprendre, car cette dernière influence indubitablement le premier. Nous nous proposons donc d'utiliser ce système dans nos prochains dispositifs d'évaluations et essayerons d'en mesurer l'efficacité.

Bibliographie

CRONBACH, L. J. (1950). Further Evidence on réponse sets and test design. *Psychol. Measurement, 10*. pp 3-31.

EBEL, L. R. (1965). Confidence-weighting and test reliability. *Journal of educational measurement, 2*, pp. 49-57

GOSLIN, D. A. (1966). *The search for ability, standardized testing in social perspective*. New-York : Willey.

HOFFMAN, B. (1964). *The tyranny of testing*. New-York : Collier

JANS, V. et LECLERCQ, D. (1997) Metacognitive realism : a cognitive style or a learning strategy ? *Educational Psychology, 17* (1 et 2), pp. 101-110.

LECLERCQ, D. (1983). *Confidence marking, its use in testing. In postlethwaite and choppin, Evaluation in Education*. Oxford : Pergamon Press, 6, pp. 161-287

LECLERCQ, D. (1993). Validity, reliability and acuity of self-assessment in educational testing. In D. Leclercq et J. Bruno (Eds), *Item banking : interactive testing and self-assessment. NATO ASI Serie*. Berlin : Springer Verlag, pp. 114-131.

LECLERCQ, D. et GILLES, J.-L. (1995). *Le kaléidoscope des techniques de questionnement*. Colloque National de l'Association Internationale de Pédagogie Universitaire. Colonster-Liège, 22 septembre 1995.