Leclercq, D. (2006) L'évolution des QCM. In G. Figari et L. Mottier-Lopez.
Recherches sur l'évaluation en Education. Paris : L'Harmattan, 139-146.

L'EVOLUTION DES QCM

Dieudonné Leclercq

Université de Liège D.Leclercq@ulg.ac.be

La présente contribution a un double but. Tout d'abord faire perdre éventuellement au lecteur son innocence[1] quant aux QCM : il n'existe pas une et une seule forme de QCM. Le second objectif est de montrer en quoi l'évolution des QCM est une succession de modifications visant à rencontrer des critères de qualité d'un système d'évaluation résumables sous le sigle **ETIC PRAD** (Leclercq, 2005) que nous énumérons d'emblée. Nous marquerons ces 8 mots-clés du signe * et nous mettrons leur première lettre en majuscule.

La validité **Ecologique*** (Brunswick, 1943) de l'évaluation pédagogique, ou « validité apparente » (en anglais *face validity*), est d'autant plus grande que la situation correspond à la situation de la vie réelle qu'elle est sensée représenter ou prédire.

La validité **Théorique*** (Cronbach & Meehl, 1955) se décompose en validité de contenu (ou de « couverture » : tout ce qu'il faut tester l'est-il ?) et en validité de construct (le Système d'Evaluation est-il fondé sur un modèle crédible, scientifiquement fondé, par exemple des processus mentaux ?).

La validité **Informative*** (si possible diagnostique) est la multiplicité des informations résultant de l'évaluation, leur distinctivité (capacité de porter sur une capacité et non sur la voisine), leur précision (sensibilité), leur valeur explicative.

La validité **Conséquentielle*** (Green, 1998) s'apprécie aux suites que l'évaluation a sur les représentations, les actes (ex : réviser ou non la

[1] Dans le sens que B. Bloom (1972) donnait à cette expression: ne plus pouvoir dire « je ne savais pas », parade des incompétents pour justifier tous leurs errements. Désormais, la personne formée devra plaider coupable.

matière, changer ou non de méthode d'étude) des apprenants, des formateurs ou du système.

La validité **Prédictive*** ou concurrente des mesures obtenues est leur capacité de prédire efficacement (c'est-à-dire avec précision et exactitude) d'autres mesures souvent ultérieures, par exemple la réussite scolaire ou professionnelle, le rendement à une autre épreuve, etc.

La **Replicabilité*** ou stabilité (**fidélité**) d'une mesure est sa stabilité dans le temps ou entre corrections. Une formule (Ebel, 1969) précise le nombre de questions d'une épreuve et, pour les QCM, le nombre de solutions proposées (distracteurs) nécessaires pour obtenir un niveau de fidélité donné (0,8 par exemple). Des formules répondent à la question de la façon inverse : quel doit être le coefficient d'allongement n du test pour atteindre une fidélité donnée (par exemple 0,80 ou 0,90) d'un test qui existe déjà et dont on connaît la fidélité actuelle ?

L' **Acceptabilité*** ou praticabilité d'une évaluation, pour le professeur, concerne l'adhésion aux principes et l'applicabilité des méthodes (durée, matériel et lieux requis, concentration, précautions antifraude, moments possibles, etc.).

Pour l'étudiant, l'Acceptabilité* concerne l'adhésion et/ou la familiarité. Ainsi, il a été démontré (Leclercq, 1986) que plus l'étudiant est familier avec les procédures de *testing*, avec les barèmes de cotation, plus il est « aguerri aux tests » (en anglais *test wiseness*) et plus ses chances de réussite sont élevées, tout spécialement avec les QCM.

La validité **Déontologique*** (ou éthique) prend diverses formes. L'équité est probablement la plus connue. Depuis longtemps, la docimologie négative (Piéron, 1963) a montré que les corrections de copies par des juges sont l'objet non seulement de non concordance interjuges, de non constance intrajuge, mais d'autres effets regrettables (de contraste, de sévérité du correcteur, de halo, d'effet Posthumus, etc.) qui sont largement évités par le recours aux QCM. Par ailleurs, les droits des étudiants étant de plus en plus (et à bon droit) reconnus, les systèmes d'évaluation garantissent de plus en plus la transparence de l'évaluation en termes de recalculabilité de la note à partie de la copie brute, de contrôlabilité du processus, ce que les QCM permettent. Nous décrirons ci-après seulement quelques moments charnières de l'histoire des QCM.

1. La gloire de la consigne classique dès sa naissance

Pressés de sélectionner les officiers parmi les appelés à la guerre de 1914-1918, les Etats-Unis font confiance aux *Army tests* conçus par Otis. Il s'agit de tests constitués de Questions à Choix Multiple fonctionnant avec la consigne Classique (QCMC), i.e. : « Une seule des solutions proposées est correcte et vous avez droit à une seule réponse ». Le fait que les USA se sont retrouvés parmi les vainqueurs n'a pas peu fait pour assurer une crédibilité à ce mode de *testing* (validité Prédictive*). Au cours des années suivantes, les modalités de *testing* systématique ont encore accru l'exigence d'efficience (rapport coût/efficacité) tant appréciée par les Américains du Nord (validité d'Acceptabilité*). Après la guerre 1940-1945, aux USA toujours, l'exigence grandissante de non-discrimination raciale (Civil Rights) dans la notation a fait apprécier ce que les américains ont appelé les « objective tests » (validité Déontologique*), alors que ces tests n'ont d'objectif que la correction. L'ajout de critères d'analyses a posteriori des réponses par les indices de discrimination ou corrélations point bisériales (Davis, 1946) ont donné aux épreuves par QCM une validité Théorique* (de construct) via la psychométrie. Ces quatre types de validité expliquent, à notre avis, le plus grand attachement des Américains aux QCM que ne l'ont été et le sont les Français par exemple.

Pourtant, dès 1963, dans son livre « La Docimologie », Piéron avait montré les discordances importantes pouvant exister entre les notes de différents juges d'une même copie « rédigée », et même l'instabilité de la note d'un même juge à une même copie. Au courant de ce problème, les autorités françaises ont cependant maintenu la notation subjective, sur la base du raisonnement selon lequel le correcteur ne connaissant pas l'identité de l'auteur de la copie, les injustices se répartissent au hasard selon un bon vieux principe (français lui aussi) d'égalité. Il se pourrait que la pratique de plus en plus courante de recours en justice (mode venue elle aussi des USA) des étudiants contre la note obtenue amène à reconsidérer la situation.

2. Une attaque théorique sur le hasard et la *correction for guessing* classique

Tversky (1964) définit la puissance d'un test par « 1 – la probabilité

d'atteindre la performance parfaite par chance ». Or on sait qu'à chaque QCMC qui comporte k solutions, l'étudiant a 1/k chances de fournir la solution correcte par chance. Plusieurs parades ont été développées pour pallier ce défaut. Dès 1920, Mc Call recourt à la *correction for guessing* classique qui consiste à fixer comme suit les tarifs : le Tarif en cas de Réponse Correcte (TC) vaut + 1 point, le Tarif en cas d'Omission (TOM) vaut 0 et le Tarif en cas de réponse Incorrecte (TI) vaut -1/(k-1) où k est le nombre de solutions proposées.

Tout aussi tôt, West (1923) critique cette procédure. Nous prétendons, aujourd'hui encore, que cette procédure est inadéquate tout d'abord parce qu'elle est basée sur un modèle théorique dépassé de l'activité mentale d'un étudiant en train de répondre à une QCMC : le premier des trois modèles décrits par Bruce Choppin (1975).

Dans ce modèle 1, quand l'étudiant « sait », il choisit la réponse correcte et quand il ne « sait pas », il choisit au hasard parmi les réponses proposées. D'où la *correction for guessing* classique.

Le modèle 2 commence comme le premier, mais au lieu de répondre au hasard quand il « ne sait pas », l'étudiant commence par éliminer les solutions qu'il sait être fausses et choisit au hasard parmi celles qui restent. Ce modèle 2, dont le 1 n'est qu'une variante extrême, reconnaît la notion de connaissance partielle défendue par De Finetti (1965). Il a donné lieu à des consignes du type QCRM (Questions de Choix à Réponses Multiples) consistant à inviter l'étudiant à éliminer les solutions incorrectes, donnant lieu, dans une QCMC à des scores allant de -(k-1) à +(k-1), rendant la mesure plus subtile, plus diagnostique*.

Le modèle 3 de Choppin va jusqu'au bout du concept de connaissance partielle et dit que quand un individu est placé devant une question (à Choix multiple ou non), il commence par ranger les solutions possibles par ordre de plausibilité décroissante et, si la consigne l'oblige à ne fournir qu'une d'entre elles, alors il choisit celle dont la probabilité (subjective) est a plus élevée (à ses yeux) Ce modèle débouche sur le recours aux degrés de certitude car, comme le dit De Finetti (1965), « Seule la probabilité subjective peut donner une signification objective à toute méthode de mesure et de *scoring*. » (p. 111).

Des modèles de Choppin, on aura compris les faiblesses de la *correction*

for guessing classique. (1) Elle manque de validité Déontologique* car elle est injuste : elle pénalise aveuglément les personnes à qui ont a interdit d'exprimer leur degré de doute. De plus (2), elle manque de validité Informative* pour les enseignants puisqu'elle ne leur apprend rien de plus. Enfin (3), et pour les mêmes raisons, elle manque de validité Conséquentielle* pour les étudiants car, à part « omettre plus souvent », elle n'a pas d'effet sur leur comportement. Cross et Frary (1977) ont en effet démontré (voir détails dans Leclercq, 1986) pourquoi cette procédure dissuade peu de « deviner ».

Les tenants des Degrés de Certitude (voir plus loin) soutiennent que cette procédure répond aux trois manques signalés ci-dessus.

3. Une rafale de critiques théoriques sur les processus mentaux mesurés et non mesurés

3.1. Les QCMC ne mesurent pas l'évocation de mémoire

Il est évident que les QCMC ne peuvent prétendre mesurer la capacité d'évoquer des connaissances, mais bien celle de les « reconnaître », ce qui n'est pas la même chose. Depuis longtemps, en effet, on sait (Luh, 1922) que la performance de recognition a un taux de réussite plus élevé que la performance d'évocation. Ces observations ont été maintes fois confirmées dans des contextes aussi différents que l'apprentissage de langues étrangères (Bahrick, 1984) ou de la médecine (Schurwirtz, 1998). Ajouter la solution « Aucune » (ou « Autre ») aux solutions possibles améliore la validité Théorique*.

3.2. Les QCMC invitent au raisonnement à rebours

Même avec la solution «Aucune » ou « Autre », les QCMC induisent un processus mental ne correspondant pas à celui que les étudiants doivent pratiquer dans la vie courante. Avec une QCMC, l'étudiant a tendance à d'abord considérer (et éliminer) les solutions proposées, puis seulement choisir la solution au lieu d'en évoquer une personnelle. C'est le modèle 2 de l'activité mentale décrit par Choppin. Or, ce que l'on veut mesurer, c'est sa capacité à évoquer la solution, puis seulement à la confronter à des solutions possibles. C'est le principe des QCL (Leclercq, 2005) ou Questions à Choix Larges : l'étudiant reçoit une liste de plusieurs

centaines de solutions rangées par ordre alphabétique (comme l'index d'un livre) parmi lesquelles il doit choisir. Chaque solution possible est affectée d'un numéro d'ordre (par exemple de 1 à 700) et c'est par ce numéro en trois chiffres (lisible par le lecteur optique de marques) que l'étudiant répond. On garde ainsi les avantages de l'automatisation de la correction, en donnant une plus grande validité Théorique* (de construct) L'automatisation de la correction permettant de poser beaucoup de questions (plus d'une par minute, par exemple 100 en une heure) contribue d'une autre façon encore à la validité Théorique* mais dans son aspect « validité de contenu ».

Les QCL seront cependant de plus en plus abandonnées avec le recours aux réponses par clavier. Il suffit, dans ce cas-là de taper le début du mot et le système propose la suite en choix large. Schurwirtz (1998) a appelé cela *Long Menu Questions*. Ceci constitue une amélioration de la validité d'Acceptabilité* **-** applicabilité de la technique.

3.3. Les QCMC ou QCMR renforcent le curriculum caché de l'école

Le curriculum caché est ce que personne n'enseigne mais que tout le monde apprend à l'école. On y apprend, notamment, que quand une question est posée, il faut y répondre ; or certaines questions, parce qu'elles sont absurdes ou excessivement intrusives, ne doivent ou ne peuvent recevoir de réponse ! On y apprend que quand l'autorité pose une question, elle est forcément pertinente, bien posée, etc. On y apprend que toute question à une réponse et que si on ne la connaît pas, on ne peut pas la retrouver par le raisonnement. Bref, le curriculum habituel (il y a heureusement de plus en plus d'exceptions) n'exerce pas à la vigilance cognitive, à la détection des anomalies, des incohérences, etc. notamment par ses modalités de *testing*, les QCMC en étant la plus représentative. Grave lacune dans la validité Théorique* de cette technique ! Pour toutes ces raisons, nous avons développé (Leclercq, 1986) les QCM à Solutions Générales Implicites ou QCM SGI. Ces solutions sont au nombre de quatre : Aucune, Toutes, Manque de données dans l'énoncé, Absurdité dans l'énoncé. Elles sont Générales par ce qu'elles sont valables (et identiques) pour toutes les questions d'un test par QCM SGI. Elles sont implicites parce qu'elles ne sont présentées qu'une seule fois (au début du test) et ne sont pas répétées dans chaque question : l'évalué doit y

penser tout seul. Du coup, cette procédure a aussi un impact sur la validité Informative* (ou diagnostique) car elle permet de distinguer deux niveaux de la taxonomie de Bloom : la compréhension (sans piège) et l'analyse (avec piège). Gilles (1999) a montré que les QCM SGI dont la réponse correcte est une SGI avaient une validité Prédictive* supérieure à celles dont la réponse correcte est une solution « visible » pour la réussite d'étudiants en médecine.

4. La rencontre entre QCM et DC (Degrés de Certitude)

Le recours aux Degrés de Certitude est indépendant des QCM. On peut très bien poser une question ouverte (du genre « En quelle année a eu lieu la révolution française ? » et demander à l'étudiant d'accompagner sa réponse d'un degré de certitude[2]. Shuford (1966), Van Naerssen (1965) et De Finetti (1965) ont montré que la consigne ne devait pas être verbale (« peu sûr », « moyennement sûr », « très sûr ») mais probabiliste (en pourcentages de chances). Nous avons en outre montré (Leclercq, 1982, 1993) qu'une précision plus grande que 20% était illusoire, d'où notre consigne en 6 degrés : 0%, 20%, 40%, 60%, 80%, 100%.

Avec les auteurs précités, nous pensons que ce procédé a une plus grande validité Ecologique* que le testing habituel qui empêche les étudiants d'exprimer leur doute. Choppin (1975) a décrit ce problème dans ses modèles 1, 2 et 3. Il dénonce la vision manichéenne (tout ou rien) de phrases telles que « Répondez uniquement si vous savez ; omettez si vous ne savez pas », alors que nous sommes très souvent (et en particulier lors de situations d'apprentissage) dans des états de connaissance partielle. (De Finetti, 1965). Le degré de doute explique les comportements de vérification (dans le dictionnaire par exemple) comme nous l'avons montré expérimentalement (Leclercq & Gilles, 1993, 45).

Avec les QCM, les Degrés de Certitude résolvent en outre (mais c'est un heureux effet secondaire, pas le but principal) le problème du *guessing*,

[2] Dans le cadre de l'opération MOHICAN (Leclercq, 2003), qui a posé des QCM (+ Autre et Toutes) dans dix matières à 4000 étudiants entrant dans les universités de la Communauté française de Belgique, nous avons posé cette question sur la date de la révolution Française. Il est intéressant de connaître non seulement le taux de réponse correcte, mais aussi la certitude moyenne (ou Confiance) accompagnant les réponses correctes ainsi que la certitude moyenne (Imprudence) accompagnant chacune des réponses incorrectes.

ce qui contribue à la validité d'Acceptabilité* (par les enseignants) des QCM.

Enfin, les Degrés de Certitude montrent leur importante contribution à la validité Informative* des QCM quand les solutions erronées sont plus choisies et surtout avec une certitude plus élevée que la (ou les) solution(s) correcte(s), ce qui est anormal. Cette situation est révélatrice de conceptions erronées (*misconceptions*).

Nous arrêterons ici cette dialectique entre les améliorations apportées aux QCM et les critiques qui continuent à leur être faites, les deux contribuant à améliorer divers aspects de la validité des mesures. L'histoire des QCM n'est pas finie. Nous invitons ceux qui s'en sentent le désir... et le courage d'en écrire quelques pages.

Références

Barhrick, H.P.(1984). Semantic memory content in permastore : 50 years of memory for Spanish learned in school. Journal of Experimental Psychology : General, 120, 1-29.

Bloom, B.S. (1972). L'innocence en pédagogie, Education - Tribune Libre, 135, 14-20.

Brunswick, E. (1943). Organismic achievement and Environment Probability, Psychological Review, 50, 255-272.

Choppin, B.H. (1975). Guessing the answer on objective tests, British Journal of Educational Psychology, 45, 206-213.

Cronbach, L. & Meehl, P. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

Cross, L. & Frary, (1977). An empirical test of Lord's theoretical results regarding formula scoring of multiple choice tests, Journal of Educational Measurement, vol. 14, 313-321.

Davis, F.B. (1946). Analyse des items. Louvain : Nauwelaerts (1966)

De Finetti, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item, British Journal of Mathematical and Statistical Psychology, 18, 87-123.

Ebel, R.L. (1969). Expected reliability as a function of choices per item, Educational and Psychological Measurement, 29, 565-570.

Gardner, H. (1996). Les Intelligences Multiples (traduit de Multiple Intelligences, 1993). Paris : Retz.

Gilles, J.L. (1999). Apports des mesures métacognitives lors d'un test de compréhension d'un article scientifique, in C. Depover & B. Noël (Eds), Approches plurielles de l'évaluation des compétences et des processus cognitifs, Actes de la 12e Conférence de l'ADMEE Mons : UMH-FUCAM, 19-30.

Green, D. R. (1998). Consequential aspects of the validity of achievement tests: A publisher's point of view. Educational Measurement, 17, 16-19, 34.

Leclercq, D. (1986). La conception des QCM. Bruxelles : Labor.

Leclercq, D. & Gilles, J.L.(1993). Hypermedia : Teaching Through Assessment, in D. Leclercq & J. Bruno, (1993), Item Banking, Interactive Testing and Self Assessment. NATO ASI Series F112. Heidelberg : Springer Verlag, 31-48.

Leclercq, D. (Ed.) (2003). Diagnostic cognitif et métacognitif au seuil de l'université. Le projet MOHICAN mené par les 9 universités de la Communauté française Wallonie Bruxelles. Liège : Editions de l'Université de Liège

Leclercq, D. (2005). Edumétrie et docimologie pour praticiens chercheurs. Editions de l'université de Liège.

Luh, C.W. (1922). The conditions of retention. Psychol. Monograph, 31, 142, 401-410.

Mc Call, W.A. (1920). A new kind of school examination. Journal of Educational Research, 1, 33-46.

Messick, S. (1988, 3° edition). Validity. In Linn R. (Ed), Educational Measurement. NY : Macmillan

Pieron, H. (1963). Examens et docimologie, Paris : Presses Universitaires de France.

Schurwirth, L.,(1998). An approach to the assessment of medical

problem solving : Computerised Case-based Testing, Ph. D., Rijksuniversiteit Maastricht : Datawyse Universitaire Pres.

Shufford, E., Albert, A. & Massengill, N.E. (1966). Admissible probability measurement procedures, Psychometrika, 31, 125-145.

Tversky, A. (1964). On the Optimal Number of Alternatives at a Choice Point', Journal of Mathematical Psychology 1(2): 386-391.

Van Naerssen, R.F (1962). A scale for the measurement of subjective probability, Acta Psychologica, 20, 2, 159-166.

West, P.V. (1923). A critical study of the right minus wrong method. Journal of Educational Research, 8, 1-9.