Journal of **E**ducational **M**easurement and **A**pplied **C**ognitive **S**ciences 1
http://www.jemacs.uni.lu/ 2007 volume 1, No 1

L'EVOLUTION DES QCM ET CRITERES DE QUALITE ETIC –PRAD

Dieudonné Leclercq, Université de Liège

A. Contexte

Comme tout pédagogue qui respecte ses principes, j'ai attribué deux objectifs au présent article. Le premier objectif est que le lecteur, au terme de cet article, ait perdu son innocence[1] quant aux QCM (ne plus croire qu'il existe une et une seule forme de QCM, la seule qu'il connaisse, qu'il pratique...ou qu'il comprenne).

Le second objectif est que le lecteur soit capable de dénouer les brins constituant la « tresse » de la qualité d'un Système d'Evaluation des Performances d'Individus en Apprentissage (SEPIA) incorporant des QCM. Nous parlons d'un système car c'est au regard des diverses composantes d'une évaluation (sa visée, certificative ou formative ; sa précision : diagnostique ou sommative ; ses objets : processus, produits ou les deux ; ses dimensions : unique ou multiples ; ses agents : externes ou internes, etc.) et de leurs interactions que l'on peut juger de la qualité d'une évaluation. Chacun de ces aspects a été décrit et illustré dans notre « Rose des vents des caractéristiques et fonctions de l'évaluation » (Leclercq, 2005). Cette **description** n'étant pas une **évaluation de la qualité** , nous avons par ailleurs développé une série de critères de qualité, série résumée par l'acrostiche ETIC PRAD.

Cette grille ETIC PRAD d'analyse de la qualité est inspirée de la grille VENTURE (1972) développée par le CSE (*Center for the Study of Evaluation)* de UCLA et et de l'association RBS (*Research for Better School*). Cette grille a été appliquée aux USA à plus de 1000 tests à usage scolaire, à l'exclusion de tests de pure connaissance de mémoire. Trois grands domaines ont été investigués : Les Higher Order Cognitive Skills, les Affective Skills et les Relational Skills. Chaque test était évalué selon 7 critères :

Validity (décomposée en construct validity, content validity)
Examinee appropriateness (facilité d'utilisation pour l'élève, adéquation à ses possibilités intellectuelles)
Normed excellence (existence de normes statistiques permettant de situer une performance parmi celles des autres étudiants du même âge ou du même niveau scolaire)
Teaching feedback (intérêt du résultat ou feedback pour l'utilisation en classe, en vue d'améliorer la performance)
Usability (commodité d'administration : rapidité, faible coût, sans matériel spécial, par une seule personne, à correction aisée)
Retest potential (possibilité de réutiliser le même test plusieurs fois, notamment pour mesurer les progrès)
Ethics (absence de problème éthique, notamment de discrimination).

[1] Dans le sens que B. Bloom (1975) donnait à cette expression dans son article « L'innocence en pédagogie » : ne plus pouvoir dire « je ne savais pas », parade des incompétents pour justifier tous leurs errements. Désormais, la personne formée sera bien obligée de plaider coupable.

Chacun de ces critères est composé d'indicateurs permettant d'attribuer des points.

Ainsi, pour la validité, les scores peuvent aller de 0 à 13. Par contre, l'Examinee appropriateness ne va que de 0 à 6 et l'Ethics se résume à un Oui/non.

Pour chacun des critères (sauf pour l'Ethics), trois catégories de scores sont déterminées : Poor (P), Fair (F) et Good (G). Ainsi, pour la Validité, un score inférieur à 6 vaut un P, un score entre 6 et 10, un F et un score de 11 à 13 un G. Chacun des 1000 tests reçoit donc un label final en 6 lettres consécutives, comme PPFPGF par exemple, dans l'ordre de VENTURE. En cas de problème éthique, s'ajoute un astérisque. Par exemple **GGFPGF*** Les auteurs ont établi le « portrait robot » du test modal constitué des modes de chaque critère Cela a donné **PGPPGP**

En d'autres termes, en majorité, les tests considérés ne posent pas de problèmes éthiques (Eth), sont bien adaptés aux élèves (Ex), sont faciles à utiliser par les enseignants (U), **mais** on ne sait pas trop bien ce qu'ils mesurent (V), ils ne permettent pas de situer par rapport aux autres élèves (N), ils ne servent à rien pour la formation (T) et ils ne peuvent pas être réutilisés (R).

B. Les 8 qualités ETIC PRAD d'un SEPIA

B1. Validité Ecologique

Cette expression, due à Egon Brunswick (1943), représente la mesure dans laquelle la situation d'évaluation correspond à la situation de la vie réelle qu'elle est sensée représenter ou prédire. Ainsi, pour mesurer la capacité de manœuvrer chez un conducteur de camion, lui demander d'introduire au clavier une grandeur angulaire de braquage ne rencontre pas le critère de validité écologique car, dans la réalité, c'est au moyen d'un volant et à 1,5m du sol qu'il aura à manœuvrer. Certaines personnes ont en effet de grandes capacités de « manœuvres dans l'espace », mais dans certaines situations. Ceci rejoint l'idée de Howard Gardner (1996), le promoteur de l'idée des « intelligences multiples » qui note *« La mesure d'une intelligence donnée ... devrait mettre en lumière les problèmes susceptibles d'être résolus dans les données et les outils propres à cette intelligence »*....(1996, 48) et « *Quand les individus sont évalués dans des conditions proches de « véritables situations de travail », il est possible de prédire leur résultat final avec beaucoup plus de précision* » (1996,158). On voit que les qualité de validité écologique et prédictive (pour plus loin) sont souvent associées.

Plus les arguments en faveur de la validité écologique du test seront puisés dans la vie courante, plus l'épreuve aura une « validité apparente » (en anglais ***face validity***).

B2. Validité Théorique

Elle se décompose en validité de **contenu** (ou de « couverture » du contenu : tout ce qu'il faut tester l'est-il et rien que cela ?) et validité de **construct** (le Système d'Evaluation des Performances d'Individus en Apprentissage ou SEPIA est-il fondé sur un modèle crédible, scientifiquement fondé, des **Processus Mentaux** ?). Les auteurs d'un test doivent établir ce type de validité par des arguments empruntés aux grandes théories et par des résultats expérimentaux jugés fiables. La notion de validité

de construct fut introduite par Cronbach et Meehl en 1955. Par exemple, on peut créer un test de compréhension d'une matière en se basant sur la théorie de Bloom (1956)[2], fortement inspirée du behaviorisme des années 50 ou par celle de Smedslund (1997), inspirée, elle, du socio-constructivisme piagétien et du cognitivisme. Elle est d'ailleurs publiée dans les Cahiers de Psychologie Cognitive. (NB les parenthèses sont de nous) :

> *« Une personne E (comme Elève) comprend correctement ce qui est signifié par une expression X produite par une personne P (comme Professeur) si les deux personnes s'accordent sur ce qui, pour P : (1) est équivalent à X, c-à-d signifie la même chose que X (2) est impliqué par X, c-à-d découle de X, (3) est contredit par X, ou nié par X, (4) est non pertinent par rapport à X, c-à-d n'a rien à voir avec X. Un désaccord sur un seul de ces critères indique une mauvaise compréhension (misunderstanding). Cela suppose que E est logique. On ne peut en effet distinguer un manque de compréhension d' un manque de logique. »*

Nous avons montré ailleurs (Leclercq, 2005, chap 4) comment construire des épreuves à QCM sur base de l'une et de l'autre théorie.

[2] Ex : 10 : Transposition : (11) Traduire, (12) transformer, (13)dire avec ses mots, (14) illustrer, (15) préparer, (16) lire, (17) représenter, (18) changer, (19a) réécrire, (19b) redéfinir……(a) des significations, (b) des exemples, (c) des définitions, (d) des abstractions, etc. 20 : Interprétation : (21) Interpréter, (22) réorganiser, (23) réarranger, (24) différencier, (25) distinguer, (26) faire,etc.. 30 : Extrapolation : (31) estimer, (32) inférer, etc

B3. Validité Informative ou Diagnostique

C'est la multiplicité des données et leur distinctivité, leur capacité d'être précises (porter sur une capacité et non sur la voisine). Par exemple, l'examen oral sollicite chez l'étudiant à peu près tous les niveaux de la taxonomie des processus cognitifs de Bloom. Cependant, il est fréquent que la communication vers l'étudiant se résume à un « c'est satisfaisant », sans plus de commentaire. C'est que plusieurs indices concordants (qui se consolident l'un l'autre) confortent l'interrogateur dans son jugement. Par contre, souvent, il ne commente pas chaque dimension, entre autres parce qu'il n'a pu en observer qu'un échantillon…et en combinaison avec d'autres dimensions. Souvent, il ne peut pas, sur la base des données à sa disposition, indiquer la (ou les) causes exactes du manque de qualité d'une performance (manque d'étude ? manque de capacité de mémoriser ? manque de capacité de communiquer ? stress paralysant, etc.).

Un exemple de précision diagnostique est fourni par les feedbacks donnés, tous les 3 mois aux étudiants des 6 années de la Faculté de Médecine de Maastricht qui ont subi le même examen (Progress Test) sur toute la médecine. Ce feedback (Leclercq & Vandervleuten, 1998, 200-201) précise à chaque étudiant pour chacune des 29 disciplines et des 14 catégories, non seulement son nombre de réponses correctes sur le nombre de questions mais sa position (note z) par rapport aux autres étudiants de son année. Et ce 24 fois au cours de toutes les études.

B4. Validité Conséquentielle

La validité conséquentielle (Green, 1998 ; Moss, 1998 ; Reckase, 1998 ; Talepros, 1998 ; Yen, 1998)) d'une évaluation pédagogique s'apprécie aux suites que cette évaluation a sur les représentations, les actes (ex : réviser ou non la matière, changer ou non de méthode d'étude) des apprenants, des formateurs ou d'autres personnes.

Un exemple d'une telle validité est fourni par l'opération RESSAC (Résultats à des Epreuves Standardisées au Service des Apprentissages en Candidatures). Les étudiants d'une première année en psychologie (Leclercq et al., 2003) ont reçu, pour chacune des épreuves de janvier subies dans 4 cours, deux notes : l'une portant sur leur performance de « restitution de Mémoire » et l'autre sur leur Compréhension en profondeur ». Une moitié environ des étudiants en échec ont dit avoir modifié leur méthode d'étude (plus de mémorisation, plus d'étude en profondeur ou les deux) pourpréparer les examens de juin et septembre. Leurs taux de réussite à ces examens a été nettement et systématiquement (dans chacun des 4 cours) nettement supérieur à ceux qui ont déclaré ne pas avoir changé leur stratégie d'étude.

Autre exemple : nous avons démontré à des étudiants de 1° année universitaire la tendance générale à surestimer sa propre compréhension de mots d'un texte technique. Après avoir passé un test qui le montrait sur leurs propres données, nombreux étudiants ont dit avoir, en conséquence, consulté beaucoup plus le dictionnaire. Ce qu'un post-test a confirmé. (Leclercq, Simon, Marotte et Lacaille, 2002.).

B5. Validité Prédictive ou Concurrente

Les mesures obtenues permettent-elles de prédire efficacement (c'est-à-dire avec précision) d'autres mesures souvent ultérieures (par exemple la réussite scolaire ou professionnelle, le rendement à une autre épreuve, etc.). C'est la corrélation entre les mesures prédictives et les mesures critères (ou à prédire) qui permet de répondre à cette préoccupation. Le cas échéant, la validité prédictive peut être établie en l'absence de validité de « construct »; c'est le cas lorsqu'un instrument prédit efficacement sans que l'on comprenne pourquoi. Ce type de situation n'est pas propre à l'éducation.

Un exemple célèbre est celui d'André Inizan (1972) qui a établi une batterie de 8 tests prédictifs passés en dernière année de l'école maternelle[3] . Un score total pour chaque enfant est calculé par la somme de ses scores à ces 8 tests. Par un suivi longitudinal de dizaines d'enfants, il a établi le nombre de mois nécessaires pour savoir lire (atteindre un score de 38 points à son test de lecture) en fonction du score total au test prédictif et de l'âge à l'entrée en primaire.

Un autre exemple est fourni par la Régression Multiple entre les notes obtenues à des tests à l'entrée (prédicteurs) sur la réussite en première année universitaire (critère) dans les diverses facultés des universités de Belgique francophone, dans l'opération MOHICAN (Leclercq et al, 2003), MOnitoring HIstorique des CANdidatures. .

B6. Replicabilité ou stabilité – Reliability (fidélité)

Une formule (Ebel, 1969, 566) 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 ?

B7. Acceptabilité - Praticabilité

Pour le professeur, plusieurs composantes sont à envisager : l'adhésion et l'applicabilité. Un exemple relatif à l'adhésion : les évaluations pédagogiques doivent-elles servir plus à sélectionner qu'à former ? Autre exemple : certains enseignants considèrent que la capacité d'évaluation (le niveau le plus élevé de la taxonomie des objectifs cognitifs de Bloom) doit intervenir dans la notation des performances des étudiants, en accordant des points (supplémentaires, donc positifs !) au réalisme dans l'auto-évaluation de leurs compétences. Ce réalisme peut être confronté à la réalité et on peut calculer objectivement la surévaluation et la sous-évaluation . D'autres enseignants, cependant, n'acceptent pas, par principe, de combiner deux mesures dont l'une fait appel à la subjectivité (pourtant mesurée objectivement).

[3] Copie de Figures Géométriques (FG), Répétition de Rythmes sonores (RR), Copie de Rythme écrit (CR), Articulation (A), Mémoire de Dessin (MD), Mémoire de Récit, Copie de lettres ou test de Horst (H), Manipulation de cubes de Kohs (K).

Les critères d'applicabilité sont faciles à énumérer : durée, matériel requis (ordinateur ou seulement papier ?), concentration, précautions (antifraude par ex.), de quel lieu, à quels moments, etc.

Pour l'étudiant aussi plusieurs critères sont à considérer : l'adhésion et la familiarité. Ainsi, fréquemment un nombre élevé d'étudiants n'adhèrent pas au principe de passer des tests (optionnels) formatifs et informatifs : ils ne profitent pas de cette occasion qui leur est offerte.

Quant à la familiarité, il a été démontré (Leclercq, 1986, 108-115) que plus l'étudiant est familier avec les procédures de testing, avec les barèmes de cotation, etc. plus il est « aguerri aux tests » (en anglais « test wiseness »), plus ses chances de réussite sont élevées, tout spécialement avec les QCM.

B8. Validité Déontologique (ou Ethique)

Depuis longtemps la docimologie négative (Piéron, 1963) a montré que les corrections de copies par des juges sont l'objet d'effets regrettables (de contraste, de sévérité du correcteur, de halo, de non concordance interjuges, de non constance intrajuge, 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.

C.L'évolution des QCM à la lumière des critères de validité ETIC PRAD

Nous décrirons ci-après seulement 7 « moments », sur un bien plus grand nombre possibles.

Une mise **en gras** d'une des 8 lettres signifie une amélioration et un **soulignement** un affaiblissement.

0

1C1. 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), c-à-d : « Une seule des solutions proposées est correcte ; 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 d'impartialité (Civil Rights) dans la notation, de non-discrimination, 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. Ces trois 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 leur livre «La Docimologie », Piéron et ses collaborateurs (Laugier et Weinberg) avaient 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. En Tunisie, zone d'influence pédagogique française, certains types de copies du Bac (en philosophie, en français) font déjà l'objet d'une double correction systématique.

C2. Une attaque théorique sur le hasard et sa parade

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. Très tôt (Mc Call, 1920), cette objection a été contrée par la « correction for guessing classique : on retire 1/(k-1) pour chaque réponse incorrecte à un QCMC. Tout aussi tôt (West, 1923), cette procédure a été critiquée à son tour, mais ces voix ont été eu entendues (ou peu lues ? ou peu crues ?). Nous prétendons, aujourd'hui encore que cette procédure manque de validité **Déontologique** car elle est injuste : elle pénalise

aveuglément les personnes à qui ont a interdit d'exprimer leur degré de doute. En outre, elle est basée sur un modèle **théorique** dépassé (et faux) comme l'a montré Choppin (1975) ; quelle est basée sur un modèle théorique faux. De plus, elle manque de validité **diagnostique** pour les enseignants puisqu'elle ne leur apprend rien de plus. Enfin, et pour les mêmes raisons, elle n'a pas de validité **conséquentielle** car, à part « omettre plus souvent », elle n'a pas d'effet sur le comportement des étudiants. Cross & Frary (1977) ont démontré pourquoi cette procédure dissuade peu de « deviner ». Les tenants des Degrés de Certitude (voir plus loin) soutiennent que cette procédure répond aux quatre manques signalés ci-dessus.

Une autre parade a consisté à recourir à la consigne des QCRM (Questions de Choix à Réponses Multiples) où l'étudiant est invité à répondre Vrai / Faux / Omission à chacune des solutions proposées, n des k solutions étant correctes et k-n étant incorrectes). Cette procédure, qui corrspond au modèle 2 de l'activité mentale décrit par Choppin) a donné lieu à d'intéressants développements (Coombs, Milholland & Womer, 1966) permettant d'approcher pour la première fois la « connaissance partielle » chère à De Finetti (1965). On attribue alors un point positif par réponse incorrecte et un point négatif par incorrecte. Le score à une question peut donc varier de –k à +k.

C3. Une rafale de critiques théoriques sur les processus mentaux mesurés

C3.1. Les QCM ne mesurent pas l'évocation de mémoire

Il est évident que les QCM 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 est bien plus facile (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, 2000). Ajouter la solution « Aucune » (ou « Autre ») aux solutions possibles permet de palier, quoiqu' imparfaitement, cette faiblesse et constitue une amélioration de la validité **théorique**.

C3.2. Les QCM 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 du type « Dans quel pays a été décrété pour la première fois l'Habeas Corpus ? 1. La France 2. L'Italie 3. L'Allemagne 4. Autre. », l'étudiant a tendance à d'abord considérer (et éliminer) les solutions proposées, puis seulement choisir la solution 4 (la réponse correcte est « L'Angleterre »). C'est le modèle 2 de l'activité mentale décrit par Choppin. Or, ce que l'on veut mesurer, c'est sa capacité à EVOQUER la sa solution, puis seulement à la confronter à des solutions possibles. C'est le principe des QCL (Leclercq, 1999) : l'étudiant reçoit une liste de plusieurs centaines de solutions rangées par ordre alphabétique, ce qui rend cette liste proche de l'index d'un livre. 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 de construct

dans la mesure où le processus est respecté. Ainsi, à la question «Quel nom donne Gavriel Salomon au processus médiatique qui fait faire le travail par le média au lieu de le laisser faire par le cerveau du spectateur ?..... », l'étudiant doit d'abord penser « supplantation », PUIS consulter la liste alphabétique et noter 482, le code du terme « supplantation ». Il n'aurait pas le temps de reparcourir à chaque fois la totalité des 700 termes pour travailler par « reconnaissance » plutôt que par « évocation ». Cette procédure contribue à la 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. Bien sûr, il suffit, dans ce cas-là de taper le mot (ou l'expresssion courte) correcte. Schurwirtz (1998) a cependant développé un principe de réponse combinant la réponse ouverte et la QCM, voir la QCL. Ce médecin chercheur de l'Université de Maastricht développe, dans le cadre de l'approche PBL (Problem Based Learning) des tests interactifs sur ordinateurs sur base de cas médicaux. Le programme d'ordinateur commence par présenter à l'étudiant testé une « vignette clinique » (un cas de patient), et l'étudiant doit diagnostiquer la maladie. Pour ce faire, il commence à introduire sa réponse au clavier. Sur base des quatre premières lettres, le programme affiche alors la liste des diagnostics possibles (en tout 2500) rangés par ordre alphabétique comme le fait Windows quand on « cherche » sur base d'un mot entré au clavier. Ainsi, si l'étudiant a tapé « Diab », le programme affiche «diabetis (type 1)» et l'étudiant peut « dérouler le menu » accompagnant ce mot pour choisir, parmi toutes les formes de diabète celle qui constitue sa réponse. Ceci constitue une amélioration de la validité d'**Acceptabilité** - Applicabilité de la technique. En effet, quand il existe des milliers de réponses possibles la technique QCL est incommode : chaque testé devrait alors avoir devant lui non pas une feuille A4 (recto-verso), mais un dictionnaire.

2

C3.3. Les QCMC ou QCMR renforcent le curriculum caché de l'école

Le curriculum caché[4] 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 jne peut pas la retrouver par le raisonnement. Bref, le curriculum habituel (il y a de merveilleuses et de plus en plus nombreuses exceptions) n'exerce pas à la vigilance cognitive, à le 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 !

---

[4] Michael Haralambos ("Sociology: Themes and Perspectives", 1991): "*The hidden curriculum consists of those things pupils learn through the experience of attending school rather than the stated educational objectives of such institutions*".

Pour toutes ces raisons, nous avons développé (Leclercq, 1986, 127-144) les QCM à Solutions Génrales 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 est) 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.

C4. La rencontre entre QCM et DC (Degrés de Certitude)

Rappelons d'emblée que 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 bataille de Waterloo ? » et demander à l'étudiant d'accompagner sa réponse d'un degré de certitude. Shufford (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%.
Avecles auteurs précités, nous pensons que ce procédé a une plus grande validité **écologique** 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. (DeFinetti, 1965). Des sentiments du genre « j'irai relire le cours, j'irais voir dans le dictionnaire, sont résumés dans le DC), comme nous l'avons montré expérimentalement (Leclercq & Boskin, 1990).
Avec les QCM, les DC résolvent en outre (mais c'est un heureux effet secondaire, pas le but principal) le problème du « guessing », ce qui contribue à la validité d'**acceptabilité** (par les enseignants) des QCM.

Enfin, les DC montrent leur importante contribution à la validité **informative (diagnostique)** des QCM dans des recherches comme celle de Baragabiribije (2005) où les questions qui font l'objet de conceptions erronées (*misconceptions*), ici en physique, présentent des solutions erronées reçoivent une certitude moyenne plus élevée que les réponses correctes, ce qui est anormal.

La place étant limitée, 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 n'est as finie. Nous invitons les lecteurs qui s'en sentent le désir...et le courage, d'écrire quelques pages de cette passionnante histoire.

Références

Alkin, M. , Thomas, J, Hill, R., Levin, J., Scanlo,n, R., Boyer, G & Simon, A. (1972) VENTURE. Los Angeles : Center for the Study of Evaluation at UCLA and Research for Better Schools, Inc..

Baragabirije, D. (2003) Testing des représentations erronées de la physique à l'université. DES en Technologie de l'Education. Universités de Liège et Namur.

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

Bloom, B. et al. (1956), Taxonomy of educational objectives. Handbook 1 : Cognitive domain, New-York, McKay, traduit par M. Lavallée sous le titre « Taxonomie des objectifs pédagogiques », Montréal, Education nouvelle (1969).

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.

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, Proceedings of the 12th ADMEE Conference, 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.

Haralambos, M. (1991). Sociology: Themes and Perspectives.

Inizan, A., Le temps d'apprendre à lire, Paris : Bourrelier, 1963.

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

Leclercq D. & Boskin A.(1990), Note taking behavior studied with the help of hypermedia, in Estes, Heene & Leclercq (Eds), Proceedings of the 7th ICTE , Brussels, vol 2, 16-19, Edimburgh : CEP Consultants.

Leclercq, D., Simon, F., Marotte, P., Lacaille, C. (2002). Former des étudiants de première candidature universitaire à des compétences transversales : lesquelles et comment ?, 2e Congrès des chercheurs en éducation, Louvain-la-Neuve, Mars.

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., Detroz, P., Dupont, C., Gilles, J-L. (2003). Changer de méthode d'étude suite aux feedbacks : l'opréation RESSAC, in Leclercq, D. (Ed.) (2003), Diagnostic cognitif et métacognitif au seuil de l'université. Liège : Editions de l'Université de Liège, 155-170.

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

Linn, R. L. (1998). Partitioning responsibility for the evaluation of the consequences of assessment programs. *Educational Measurement, 17,* 28-30

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

Lune, S., Parke, C. S., & Stone, C. A. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement, 17,* 24-28.

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

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement, 17,* 6-12.

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

Reckase, M. D. (1998). Consequential validity from the test developer's perspective. *Educational Measurement, 17*, 13-16.

Schurwirth, L.,(1998) An approach to the assessment of medical problem solving : Computerised Case-based Testing, Ph. D., Rijksuniversiteit Maastricht : Datawyse Universitaire Pres, 1998.

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

Smedslund, J. (1997), The forgotten variable of understanding . Cahiers de Psychologie Cognitive – Current Psychology of Cognition, 16 (1-2), 217-221.

Taleporos, E. (1998). Consequential validity: A practitioner's perspective. *Educational Measurement, 17,* 20-23, 34.

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.

Yen, W. M. (1998). Investigating the consequential aspects of validity: Who is responsible and what should they do? *Educational Measurement, 17,* 5.