ETIC PRAD et SEPIAs

Systèmes d'Evaluation des Performances d'Individus en Apprentissage

D. Leclercq (2005)

Extrait du Chapitre 5 « Analyse édumétrique des données de recherche » de D. Leclercq (2005) Edumétrie et docimologie pour Praticiens chercheurs. Editions de l'Université de Liège

En proposant 8 critères, nous proposons un instrument pour juger des qualités de SEPIAs (Systèmes d'Evaluation des Performances d'Individus en Apprentissage**).**
Le chapitre consacré à « La rose des vents des fonctions et caractéristiques de l »'évaluation pédagogique » ne parlait pas de **qualité** de ces caractéristiques. La notion de qualité n'a de sens que dans un système. Nous renverrons à des cas développés dans d'autres chapitres.

La grille ETIC PRAD a été utilisée dans Leclercq, D. (2006). L'évolution des QCM. in G. Figari & L. Mottier Lopez (Dir). *Recherches sur l'évaluation en éducation* . Paris : L'Harmattan, 139-146.

Partie 1

1.1. D'un test vers un SEPIA

Quand Carver (1974) a défini les deux types de propriétés des **tests**, à savoir leurs propriétés psychométriques d'une part et édumétriques de l'autre, il est resté assez vague. Bien sûr, il a attribué aux **propriétés psychométriques** la capacité de fournir des mesures valides, fidèles et sensibles entre les individus, bref de rendre compte des variations interindividuelles. Symétriquement, il a attribué aux **propriétés édumétriques** la capacité de rendre compte de manière valide, fidèle et sensible des modifications intraindividuelles, des changements, cognitifs ou affectifs ou sensori-moteurs ou encore relationnels par exemple, chez une même personne, soit dans le temps pour un même type de performance et de contenu (on parle alors de gains et de perte, ou de modification qualitative), soit pour des performances et des contenus différents (on parle alors de flexibilité de stratégies d'apprentissage, de profil adaptatif personnel), et , dans les deux cas, d'ambivalence et de polyvalence mathétiques[1]).

Ce ne sont pas les indices eux-mêmes qui peuvent être classés comme psychométriques ou comme édumétriques, mais un **système** d'évaluation, incluant les **propriétés** des tests, les **caractéristiques** des mesures et l'**utilisation** de l'information qui en est faite.

Le titre de Carver se concentrait sur le concept de test. Nous partons d'un concept apparemment plus complexe, parce qu'il se veut plus analytique : celui de **Système d'Evaluation d'Individus en Apprentissage**. Par « d'individus », nous voulons exclure de notre propos les évaluations des institutions ou des enseignements, même si ceux-ci ont bien des points communs avec l'évaluation des étudiants. Nous parlons de performances, permettant de faire, éventuellement des inférences sur les compétences (qui assurent la reproductibilité des performances). Enfin, et surtout, nous parlons de **système.**

1.2. Système

Par système, nous entendons d'un ensemble d'éléments (ici de caractéristiques et de qualités) interdépendants.

Ainsi, les scores des étudiants dépendent évidemment de leurs réponses qui, elles-mêmes, dépendent des consignes (y compris les modes et barèmes de correction) et des types de questions (QROC, QCM, QRM, SGI, DC[2]) qui, elles-mêmes dépendent des objectifs ou visées (formative ? certificative ?) de l'évaluation.
Enfin, par le pluriel au mot « qualités », nous voulons aussi indiquer qu'il va falloir prendre en compte toutes les composantes du système et les juger à plusieurs aunes, selon plusieurs critères.
Nous pensons que les systèmes édumétriques ou d'évaluation des performances cognitives (SEPIA) des étudiants peuvent être **décrites** par « la rose des vents » de l'évaluation et **évaluées** (en qualité) par la grille ETIC PRAD que nous développerons ci-après.

Avant de pénétrer dans le vif de notre sujet, nos exposerons l'approche VENTURE qui a été appliquée aux USA à des milliers de test. Notre grille ETIC PRAD s'inspire dd plusieurs des principes sous-tendant cette grille VENTURE.

[1] Du verbe « manthanau » (en grec ancien) : « j'apprends ».

[2] Nous aurons l'occasion ci-après de définir chacune de ces expressions.

1.3. Les 7 critères de VENTURE

Le projet VENTURE est l'œuvre du CSE (*Centre for the Study of Evaluation*) de l'UCLA (*University of California at Los Angeles*) et de l'association RBS (*Research for Better School*) et qu'ils ont décrit dans une publication de 1972. Ces deux organismes ont examiné 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).

1.4. Les indicateurs de chaque critère

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 :

VENTURE
GGFPGF*

Les auteurs ont établi le « portrait robot » du test modal constitué des modes de chaque critère VENTURE.

VENTURE
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),
-ne permettent pas de situer par rapport aux autres élèves (N),
- ne servent à rien pour la formation (T) et
-ne peuvent pas être réutilisés (R).

Nous nous sommes étendus ailleurs sur ces résultats et sur l'approche VENTURE elle-même (Leclercq, 2005d). Ce qui importe ici est d'illustrer la multiplicité des critères et la complexité de certains d'entre eux (la validité par exemple). Bien que nous ayons donné la seule vue d'ensemble (le portrait robot modal, bien décevant par ailleurs), il est évident que ces critères permettent de repérer des tests (plutôt rares) ayant des qualités telles qu'ils méritent d'être adoptés dans la pratique scolaire et diffusés. Grâce à la radiographie VENTURE, on sait au moins sur quelle base ils ont été choisis.

Partie 2

Les 8 qualités ETIC PRAD d'un SEPIA

2.1. La Validité Ecologique

Cette expression est due à Egon Brunswick (1943). Le SEPIA est fondé sur un modèle de la situation, de l'Environnement dans lequel l'évaluation a un sens, ne dénature pas la mesure. En ce compris le mode de réponse. 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 possibede prédire leur résultat final avec beaucoup plus de précision* » (1996,158)

Quand nous défendons (Leclercq, 1982, 1993, 2003) le recours aux Degrés de Certitude (DC) acompagnant une réponse, c'est entre autres parce que nous pensons que ce procédé a une plus grande validité écologique que le testing habituel qui empêche les étudiants d'exprimer leur doute. Choppin (1971) 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) danbs 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).

2.2. La 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'évaluation des performances cognitives (SEPIA) est-il fondé sur un modèle crédible (scientifiquement fondé) des **Processus Mentaux** ?

Le test correspond-il à la théorie concernant la variable mesurée ? 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. Plus les arguments seront puisés dans la vie courante, plus l'épreuve aura une « validité apparente » (en anglais *face validity*).

La notion de validité de construct fut introduite par Cronbach et Meehl en 1955.

Un exemple

Voici un plan de répartition des questions (100 QCM, dont 40 à livres fermés = LF, donc 100 degrés de certitude) et 2 questions orales pour un examen universitaire en grand groupe.

chapitres

Compétences	1	2	3	4	5	6	TOT
Mémoire de restitution (LF)	**5**	**5**		**5**	**5**		**20**
Mémoire de reconnaiss. (LF)	**5**	**5**		**5**	**5**		**20**
Compréhension QCM SGI L Ouv	**5**	**5**		**5**	**5**		**20**
Application à cas particuliers	**5**	**5**		**5**	**5**		**20**
Analyse par QCM SGI Double face Sur ordinateur	**5**	**5**		**5**	**5**		**20**
Synthèse- expression (oral)			**1**			**1**	**2**
Evaluation –jugt Métacognitif-cert	**25**	**25**		**25**	**25**		**100**

On voit que cette répartition est inspirée par le **construct** (la théorie) « Taxonomie de Bloom » (en lignes)

On voit aussi le soucis de « couverture » de tous les chapitres (en colonnes), de manière à assurer la validité de **contenu.**

2.3. La Validité Informative ou Diagnostique

On peut s'intéresser à cette validité

-**par étudiant** : c'est la multiplicité des données et leur distinctivité, leur capacité d'être précises (porter sur une capacité et non sur la voisine).

Un exemple
L'examen oral sollicite chez l'étudiant à peu près tous les niveaux de la taxonomie des processus cognitifs de Bloom évoquée en K3 ci-avant. 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 autre exemple
Les QCM, avec leurs variantes (SGI, DC, etc.) permettent de fournir des **précisions séparées** sur certains aspects de la performance tels que la mémoire de reconnaissance, celle de rappel, la compréhension, la vigilance, la Confiance, la Prudence, Nuance, etc.

-par questions ou par matières

Un exemple
Le plan de répartition des questions de la page précédente permet d'imaginer des scores par chapitre.

Un autre exemple
On trouvera au chapitre 7 la description de la recherche de Baragabiribije où est mise en évidence une façon de repérer les questions qui font l'objet de conceptions erronnées (misconceptions) dans une population d'élèves.

2.4. Validité Conséquentielle

La validité conséquentielle d'une évaluation s'apprécie aux suites quecette évaluation a sur les représentations, les actes (ex : réviser ou nonb la matière, changer ou non de méthode d'étude) des apprenants, des formateurs ou d'autres personnes.

Un exemple
Nous avons démontré à des étudiants de 1° année universitaire la tendance 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.

La validité conséquentielle a été discutée et illustrée par plusieurs auteurs (Green, 1998 ; Linn, 1998 ; Lune, Parke & Stiome, 1998 ; Moss (1998 ; Reckase, 1998 ; Talepros, 1998 ; Yen, 1998).

2.5. Validité Prédictive ou Concurrente

a) Le principe

Les mesures obtenues permettent-elles de prédire efficacement (c'est-à-dire avec précision) d'autres mesures ultérieures (par exemple la réussite professionnelle, le rendement sportif, etc.). A nouveau, 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.

b) Un exemple

Inizan a aussi établi une batterie de 8 tests prédictifs passés en dernière année de l'école maternelle :

-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)

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.

Ainsi, à 6 ans, avec un score de 90 à 100, 3 mois suffisent. Avec un score de 86 à 90, 4 mois ; pour 72 à 86, 6 mois ; de 61 à 72, 9 mois et en-deçà de 61, l'apprentissage est jugé « inopportun » car irréalisable au cour d'une année scolaire complète.

Vérifier la validité prédictive de ce modèle (ou plus exactement des dates-mesures prédites) consiste à comparer les prédictions et les observations.

Autres exemples :

-la Régression Multiple des prédicteurs sur la réussite de la 1° candi dans MOHICAN (Leclercq, 2003)..

-Analyse d'items (JLGilles : plus prédictive d'un défaut).

2.6. Replicability – Reliability (fidélité)

Une formule (Spearman Brown) détermine le nombre de questions et le nombre de distracteurs nécessaires pour obtenir un niveau de fidélité donné (0,8 par exemple).

On peut se poser 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 ? On y répond par la formule ci-dessous :

$$n = \frac{rnn(1 - rll)}{rll\,(1 - rnn)}$$

(Guilford et Fruchter, 1978, p. 432)

Les formules correspondantes pour la validité[1] sont les suivantes. La corrélation entre un critère (désignons-le par y) et un test x allongé a fois se note ry (ex); la fidélité du test de longueur initiale (a = 1) est notée rxx.

$$(ry\,(ax))^2 = \frac{(ryx)^2}{\frac{1 - rxx}{a} + rxx}$$

(Guilford et Fruchter, 1978, p. 449)

On voit que la fidélité de départ intervient dans le calcul de l'accroissement de la validité.

$$a = \frac{1 - rxx}{\frac{(ryx)^2}{(ry\,ax)^2} - rxx}$$

(Guilford et Fruchter, 1978, p. 450)

[1] Rappelons qu'en termes statistiques la validité des mesures (scores à un test) est la corrélation entre les mesures et les mesures obtenues au moyen d'un instrument de référence, ou critère. Ce coefficient de validité ne peut excéder le coefficient de fidélité : on ne peut imaginer un test mieux corrélé avec un autre test qu'avec une formule parallèle de lui-même.

2.7. Acceptability - Practicability

-pour le professeur :

Plusieurs composanes sont à envisager :

-l'adhésion
Certains enseignants considèrent que leurs évaluations doivent servir plus à sélectionner qu'à former. Il est vrai que les enseignants doivent faire les deux. Mais dans quelles proportions ? 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, par exemple en accordant des points (supplémentaires, donc positifs !) au réalisme dans l'auto-évaluation des 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'accepetent pas, par prinicpe, de combiner deux mesurs dont un e fait appel à la subjectivité (pourtant mesurée objectivement).

-l'applicabilité :

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

-**pour l'étudiant** :
-l'adhésion
Il arrive fréquemment que la possibilité soit offerte aux étudiants de passer des tests pour se faire une idée de leur niveau de compétence et que beaucoup d'étudiants ne profitent pas de cette occasion.

-la familiarité
Plus l'étudiant est famlilier 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 (voir chapitre 9).

2.8. Deontology - Ethics

-Inter-étudiants :
Certains types d'épreuves favorisent certains types d'étudiants. Ainsi, les examens oraux favorisent les « extravertis », les écrits ceux qui ont « une belle écriture » ou « une bonne orthographe ».

ceux qui se sous-estiment (argument rejeté car la connaissance est pour s'en servir (Hunt) et pour la communiquer à d'autres

-Transparence
, recalculabilité, contrôlabilité, imputabilité, (pas comme le Bac français ;
NB : en Tunisie, les doubles corrections).

-libre des discordances de jugement inter-juges (docimologie négative).

Références

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

CSE (*Centre for the Study of* Evaluation) - UCLA (*University of California at Los Angeles*) & RBS (*Research for Better School*) (1972). The VENTURE Project.

Carver, R.P. (1974), Two dimensions of tests : Psychometric and edumetric, American Psychologist, 29, 512-518.

Choppin, B. (1975), Guessing the Answer on Objective Tests, Brit. Journal of Educational Psychology, 45, 206-213.

Cronbach, L. J. and Meehl, P. M. (1955) "Construct Validity in Psychological Tests," *Psychological Bulletin* 52: 281-302.

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.

Gardner, H. (1996). Les intelligences multiples, Paris : Retz, trad de Multiple intelligences. The theory in practice. A reader. (1993) Basic Books.

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. (1982), Confidence marking, its use in testing,. in Postlethwaite & Choppin, Evaluation in Education, vol. 6, 161-287, Oxford : Pergamon Press.

Leclercq, D. & Boskin, A. (1990), Note taking behavior studied with the help of hypermedia, in Estes, Heene & Leclercq (Eds), Proceedings of the 7th International Conference on Technology and Education, Bruxelles. Edimburgh : CEP Consultants, 2, 16-19.

Leclercq D.(1993). Validity, Reliability and Acuity of Self-Assessment in Educational Testing, in Leclercq D. & Bruno J. (1993), Item Banking : Interactive Testing and Self-Assessment, NATO ASI Series, F 112, Berlin : Springer Verlag, 114-131.

Leclercq, D. (Ed) (2003). Diagnostic cognitif et métacognitif au seuil de l'université. Le projet MOHICAN mené par les 9 uniersités 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.

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

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

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

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

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

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