Chapitre 9

Analyses psychométriques des questions des 10 check-up MOHICAN

Vue d'ensemble

D. Leclercq, *Président du Groupe de travail CIUF « Réussite en candidatures »*
J.-L. Gilles, *Directeur du SMART ULg*

1. Les indices de popularité des différentes solutions

Ces indices sont à la fois psychométriques et édumétriques. Ainsi, la popularité de la solution correcte à une QCM ou pourcentage de répondants ayant choisi cette solution, ou encore pourcentage de Réponses Correctes (%RC) est un indicateur psychométrique : plus ce taux de réponse correcte est proche de 50%, plus cette question contribue à discriminer les répondants entre eux. Il peut aussi s'interpréter édumétriquement en termes de distance à la perfection. Ainsi, un taux de réponses correctes de 80% indique une « non maîtrise » concernant 20% des étudiants et donne une idée de la taille du public concerné par l'utilité éventuelle d'une remédiation.

1.1 Les popularités des solutions correctes ou Facilité Objective de la question : Quel est le niveau objectif de réussite et d'erreur ?

Cette **popularité** de la solution correcte, ou taux de réussite, est l'indice de la **Facilité Objective de la question**, qui va de 0% à 100%. Dans la littérature, on trouvera souvent l'expression « p » ou « indice de difficulté », expression malheureuse puisque quand cet indice numérique est maximal (100%), la facilité est maximale mais la difficulté, elle, est minimale.
C'est sur une échelle verticale de facilité qu'ont été positionnées les questions dans le chapitre 3. Ils sont donc ainsi situés (édumétriquement) en valeur absolue, c'est-à-dire par leur distance à la perfection (100%), et relativement les uns aux autres (le plus facile, le plus difficile, etc.).

1.2 Les popularités des distracteurs : quelles erreurs sont les plus fréquentes ?

Un distracteur est une solution incorrecte proposée parmi les choix de réponses, dont les solutions générales[32] Aucune et Toutes. Nous appelons **« distracteur-vedette »** le plus populaire, le plus choisi, et son « challenger », celui qui est choisi en deuxième lieu, ou plutôt en deuxième proportion. Ces popularités renseignent sur l'ampleur des erreurs et surtout sur leur nature, voire leur cause. La **valeur diagnostique** d'un distracteur est due au talent du créateur de la question. Une méthode efficace de génération de distracteurs consiste à utiliser les réponses erronées données antérieurement par des étudiants lors de l'administration de la question de façon ouverte, c'est-à-dire demandant à l'étudiant de produire une réponse et non de la choisir. C'est de cette façon que les distracteurs de plusieurs des check-up de MOHICAN ont été générés (pour plus de détails, voir les auteurs de chaque check-up).

[32] Ces solutions sont appelées «générales» parce qu'elles sont d'application à toutes les questions d'une épreuve. Elles seraient en outre appelées «Implicites» si elles n'apparaissaient **pas** (dactylographiées) parmi les solutions de chaque question. Comme la grande majorité des étudiants n'avaient quasiment aucune expérience de cette technique, nous avons décidé de rendre explicites ces deux solutions, ce qui en diminue la valeur formative qui vise à entraîner la vigilance cognitive (Leclercq, 1986, 127-143).

2. Les indices (psychométriques) de discrimination de chaque solution de chaque QCM

2.1 Les erreurs et les réponses correctes viennent-elles d'étudiants de compétences différentes ?

Très souvent, dans une situation de testing, et c'est le cas pour MOHICAN, la meilleure information dont on dispose sur la compétence de chaque étudiant dans la matière est son score au total à l'épreuve. Chaque fois que l'on utilisera ci-après l'expression « étudiant fort » ou « étudiant faible », il faudra comprendre « dans la matière mesurée par le score total à l'épreuve dont la question fait partie ». On peut déterminer si le choix d'une solution (correcte ou distracteur) appartenant à une QCM est lié au score total à l'épreuve par le calcul d'une corrélation, la **corrélation point bisériale** (en raccourci **r. bis**). Cet indice prend des valeurs comprises entre -1 et +1.
Est-ce un groupe d'**étudiants « forts »** (au total de l'épreuve) qui ont choisi une solution ? Si oui, la corrélation est positive car le score moyen au total de l'épreuve de ceux qui ont choisi cette solution est supérieur au score moyen de ceux qui ne l'ont pas choisie. C'est la situation attendue pour la solution correcte d'une QCM.
Si ce sont des **étudiants « faibles »** (au total de l'épreuve) qui ont choisi cette solution, la corrélation (r.bis) est négative. C'est la situation à laquelle on s'attend pour les solutions incorrectes.
Quelle valeur du rpbis de la réponse correcte est satisfaisante ? Cela dépend d'une épreuve à l'autre, comme on va le voir.

2.2 La valeur-seuil ou valeur repère du r.bis : La question contribue-t-elle mieux que le hasard au score total à l'épreuve ?

On doit s'attendre à ce que la corrélation entre le choix d'une solution correcte à une question et le score total à toute l'épreuve soit positive. C'est une conséquence du fait que chacune des NQ questions de l'épreuve contribue, pour un poids de 1/NQ questions à ce score total. Autrement dit, chaque score (0/1) à une question est une partie du score total à l'épreuve. On parle aussi de « **recouvrement** » entre le score à une question et le score au total de l'épreuve, ou d'**inclusion** du premier dans le second. On peut calculer la corrélation « repère » qui découle de ce raisonnement par la formule $1/\sqrt{NQ}$ (Guilford & Fruchter, 1978). Cette corrélation « automatique » ou « repère » positive doit être atteinte et si possible dépassée par la valeur du r.bis de la réponse correcte pour considérer que la question « va dans le même sens », « mesure la même chose » que l'ensemble du test. Cet indice est d'ailleurs appelé **indice de cohérence interne**.
Dans une épreuve de 25 questions, la corrélation repère vaut $1/\sqrt{NQ}$ soit $1/\sqrt{25}$ soit1/5 soit 0,20. Avec 45 questions, la corrélation automatique ou repère pour la réponse correcte vaut $1/\sqrt{45}$=1/6,7=0,15. En deçà de cette valeur repère, le r.bis indique que la question pose un problème. Le r.bis n'indique pas la cause du problème. La raison peut en effet être une mauvaise formulation de la question, ou le fait qu'une question (de mathématique par exemple) n'est pas du même type que les autres (dans une épreuve de géographie par exemple) ou même que des erreurs ou des ambiguïtés ont été véhiculées par le cours oral ou par le support écrit.

2.3 Les valeurs négatives des r.bis des distracteurs : sont-ils choisis par des étudiants plus faibles que la moyenne générale ?

On s'attend à ce que le choix d'un **distracteur** ait, lui, une **corrélation négative** avec le score total parce que choisi par des étudiants dont le score total moyen à l'épreuve est inférieur au score total moyen des autres étudiants (dont les étudiants qui ont réussi).
Les valeurs des rpbis des distracteurs peuvent être comparées entre elles, ce qui permet de voir que certains distracteurs sont choisis par des élèves plus faibles (en moyenne) au total de l'épreuve que d'autres distracteurs. C'est souvent l'indication d'une différence de profondeur ou de **gravité des erreurs**.

L**'omission** est, elle, souvent le fait d'étudiants « intermédiaires » : moins forts que ceux qui ont choisi et moins faibles que ceux qui se sont trompés.

Les indices de discrimination r.bis, empruntés à la psychométrie, sont utilisés depuis des décennies par les gestionnaires d'épreuves pédagogiques, notamment ceux qui doivent garantir les propriétés métriques des tests et des examens. Quand ils ne prennent pas les valeurs attendues (voir ci-avant), ils servent de signal d'alarme qui invitent le constructeur de l'épreuve à se demander pourquoi le signal s'est déclenché. Voici la vue d'ensemble des situations es 10 check-up à ce sujet.

	r.bis de la Réponse Correcte			r.bis des Rép. Incorrectes ou distracteurs		
	N	positif et non supérieur au seuil	positif ET supérieur au seuil	N	non négatifs	négatifs
Vocabulaire	**45**	2	43.	**206**	23	183
Syntaxe	**12**	0	12	**72**	0	72
Compr.Texte	**6**	1	5	**36**	1	35
Géo	**10**	1	9	**56**	3	53
Math	**22**	0	22	**132**	4	128
Physique	**10**	0	10	**60**	1	59
Chimie	**8**	1	7	**48**	2	46
Bio	**10**	3	7	**60**	3	57
Histoire	**25**	2	23	**150**	14	136
Art	**25**	2	23	**150**	25	125
	173			**970**		

Ce tableau récapitulatif montre qu'à quelques exceptions près, discutées en détails dans l'analyse de chaque épreuve, ces indices psychométriques (r.bis) sont satisfaisants.

3. Un indice psychométrique de la fidélité des tests : l'alpha de Cronbach

Cet indice mesure la cohérence interne des questions par rapport au score total au test, bref dans quelle mesure chaque question mesure la même chose que les autres.

La formule est la suivante : $\alpha = (NQ/ (NQ-1)) . (1- (\leq\sigma_q / o^2_q)$

Voici les valeurs des alpha de Cronbach pour chacun des check-up MOHICAN (extraite de Gilles, 2002, 308) :

	Voca-bulaire	Syntaxe Compréh.	Compréh. Textes	Compréh. Géo	Math	Physique	Chimie	Biologie	Hist-Act-Eco	Conn. Art
Alpha de Cronbach	0,833	0,572	0,393	0,530	0,769	0,473	0,414	0,410	0,668	0,707
Nbre de Questions	45	12	6	10	22	10	8	10	25	25
Allongement Nécessaire pour F = 0,80	-9	24	31	25	4	35	37	48	25	16

Pour le check-up **Vocabulaire**, on pourrait se permettre de supprimer 9 questions, et la fidélité de 0,8 ne serait pas remise en cause.

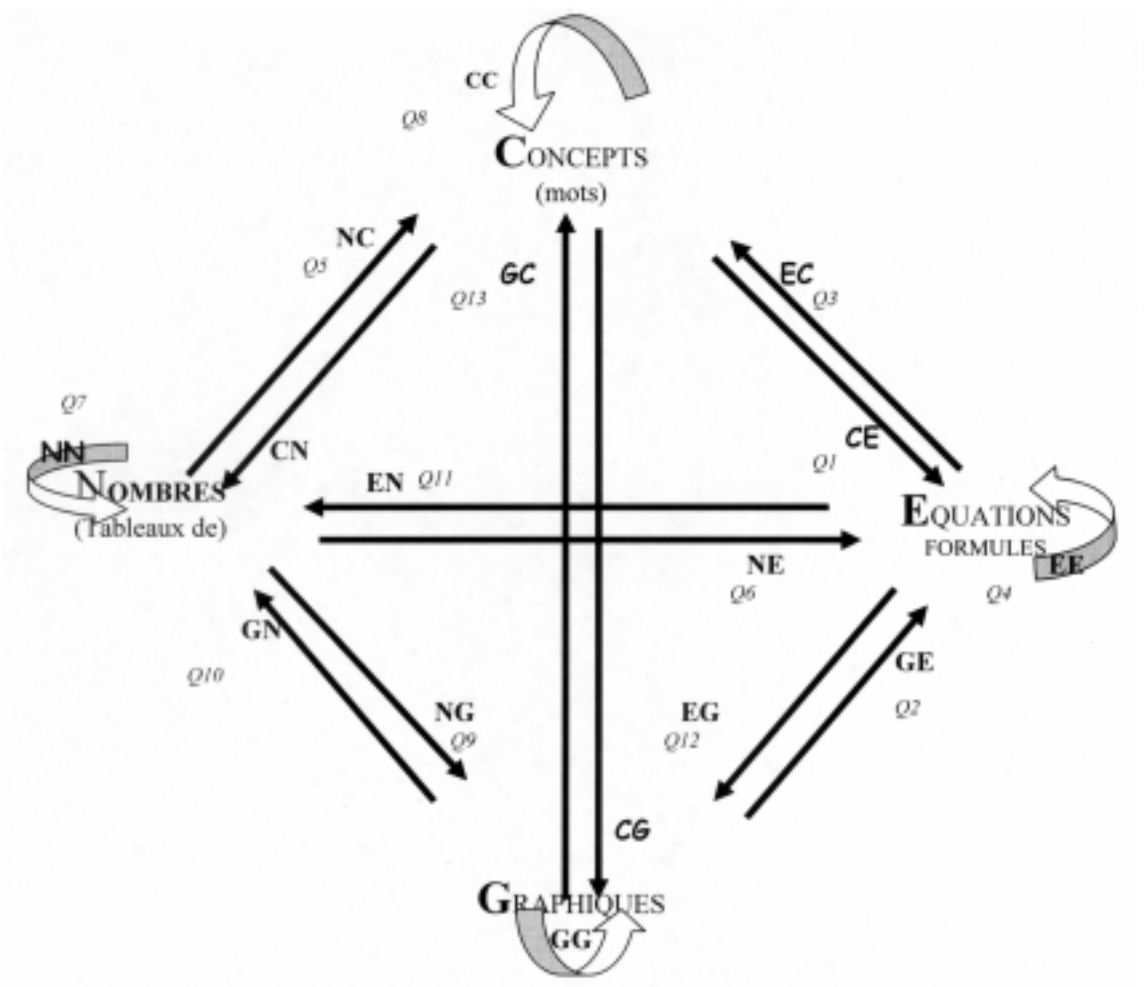
Pour les autres check-up, on voit que le nombre de questions devrait être largement augmenté, sauf pour le check-up de Math qui avec 26 questions (au lieu de 22) permettrait d'atteindre une fidélité de classement des étudiants entre eux (sur base de leur score total) avec une fidélité de 0,80.

Rappelons, cependant, que les check-up n'avaient pas pour ambition d'opérer un tel classement.

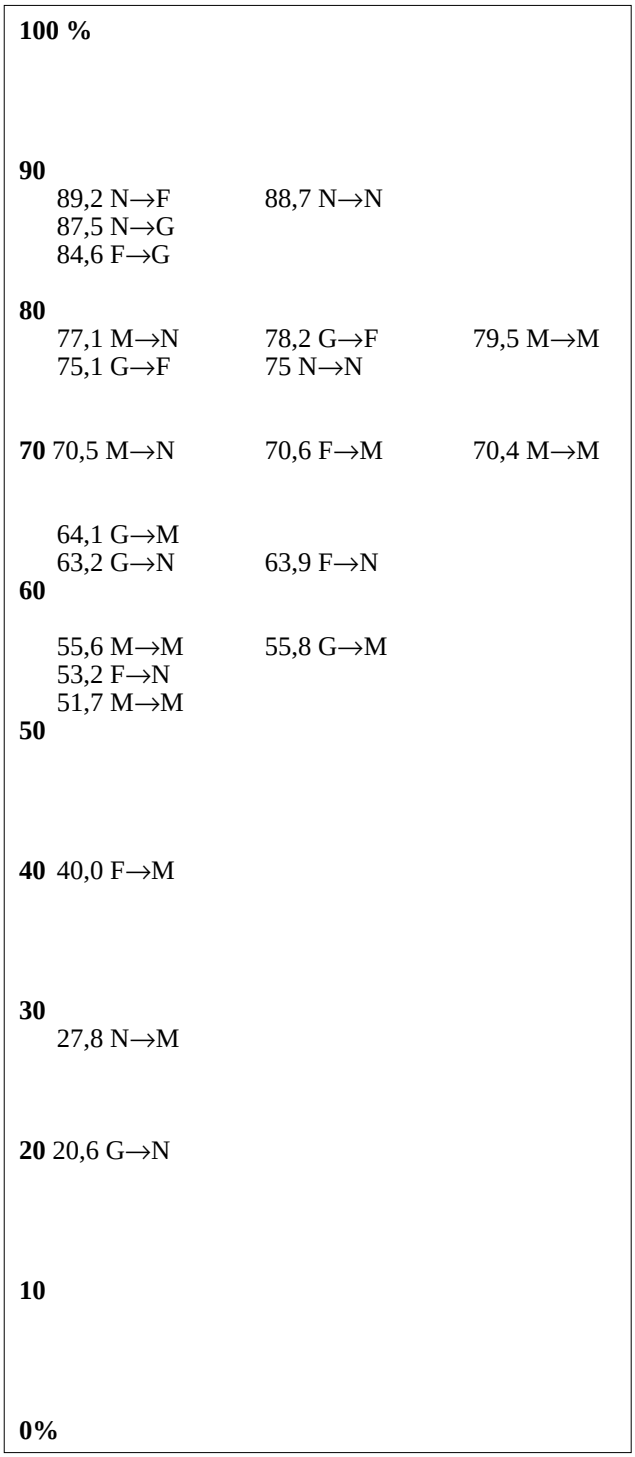
La préoccupation première était celle de la validité (ou pertinence de la mesure en fonction des objectifs de mesure) et non de la fidélité de classement.

Les critères de validité ont varié selon les check-up. Ainsi, pour le check-up **Vocabulaire**, la validité des questions présentes repose sur les observations systématiques faites par l'auteur (M. Monballin) des vocables non techniques de la langue française faisant difficulté pour les étudiants de 1° année universitaire (sur base d'interviews de professeurs).
Pour les check-up de Syntaxe, Compréhension de Textes, Géographie et Lecture de cartes et graphiques, Physique, Chimie, biologie, Art et Histoire-Economie-Actualité, c'est la multiplicité (l'hétérogénéité) des domaines touchés qui a été le critère de choix, principe antagoniste avec le concept de fidélité.

Le check-up de **mathématique**, lui, porte sur l'application de principes, de connaissances, mais surtout sur la compréhension, dans le sens « traduction d'un langage dans l'autre » car ses questions peuvent être représentées comme dans un carrefour représentant les divers « tournants » possibles (du concept exprimé en mots au graphique, et vice versa, etc.).
Un tel « carrefour » comporte 16 « chemins », dont la plupart (voir les numéros correspondants des questions) ont été reflétés dans le « Check-up » Mathématique.



C'est cette signification, conçue dès la création des questions (ou plus exactement la sélection dans une banque de questions plus vaste des auteurs) **qui en fait la validité diagnostique** qui apparaît clairement au chapitre 3 page 14 (voir sa copie ci-contre) où il apparaît que ce sont les questions de traduction vers les MOTS qui sont les moins bien réussies ! ! !



Le lecteur jugera de l'intérêt diagnostique à la fois de la structuration en « carrefour des traductions » et de la présentation sur une échelle unique des questions par ordre de Facilité objective (FO) décroissante. On imagine que les 22 questions pourraient être situées dans un espace à deux dimensions, l'autre dimension étant celle des Facilités Subjectives, ou, mieux, des facilités subjectives des réponses Correctes, bref des Indices de Confiance, comme au chapitre 2, en section B.

4. Le Réalisme par Calibration par question ?

Il est tout à fait pertinent de calculer un indice de Réalisme par Calibration par question ; c'est ce qu'a fait Gilles (2002), auquel nous renvoyons.

5. Bibliographie

Cronbah, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16

Gilles, J.-L. (2002). Qualité spectrale des tests standardisés universitaires, Thèse de doctorat en Sciences de l'Education, Université de Liège.

Guilford, J.-P. & Fruchter (1978). *Fundamental Statistics in Psychology and Education*. New York : Mac Graw Hill.