Les points ou cotes attribués aux élèves de nos classes et à tous les apprenants ont-ils les propriétés métriques des mesures effectuées avec une balance, ou avec un thermomètre, ou avec une règle graduée, ou avec un chronomètre, etc. ?

Que la réponse soit OUI pour certaines mesures et NON pour d'autres est une donnée cruciale que tous ceux qui « notent » les autres et tous ceux (les parents par exemple) qui interprètent ces notes devraient avoir bien comprise avant de se lancer dans des débats stériles ou des pratiques contestables.

Comment s'exprime la relation fondamentale entre la compétence d'un apprenant et la difficulté d'une question ? Comment juger de la qualité d'une question ou d'un test à la lumière des réponses des étudiants ? Quelle est l'influence du hasard dans l'attribution des points lors du recours à des QCM ? Quels sont les systèmes de « compensation par retraits de points » et que valent-ils ?

Autant de questions auxquelles cet ouvrage répond en théorie et en pratique, partir d'exemples concrets et d'analyses de cas. Un livre que tout enseignant, tout formateur se doit d'avoir lu.

IMPRIME EN BELGIQUE D/1987/258/31 ISBN 2-8040-0245-4

EDITIONS LABOR ÉDUCATION 2000

L 90 80 77

D. LECLERCQ QUALITE DES QUESTIONS ET SIGNIFICATION DES SCORES EDUCATION 2000 Editions LABOR

Dieudonné LECLERCQ

QUALITE DES QUESTIONS ET SIGNIFICATION DES SCORES

AVEC APPLICATION AUX QCM

EDITIONS LABOR ÉDUCATION 2000

D/1987/258/31
ISBN 2-8040-0245-4
L 90 80 77

D. LECLERCQ

Qualité des questions et signification des scores

avec application aux QCM

Collection
EDUCATION 2000
dirigée par Gilbert DE LANDSHEERE

Editions LABOR, Bruxelles

Préambule

Dans un ouvrage précédent, intitulé « La conception des QCM », nous avons tenté de fournir une définition de la nature profonde et des aspects formels d'une QCM, nous avons analysé cette technique, ses défauts et ses avantages, et proposé une classification des consignes ainsi que vingt règles de rédaction fondées sur la recherche expérimentale.

Le lecteur a pu percevoir un certain nombre de problèmes évoqués à cette occasion dépassant largement la seule technique des QCM et concernant l'évaluation pédagogique en général. C'est le cas par exemple pour les modèles de l'activité mentale d'un étudiant en train de répondre, pour la problématique de l'uniformisation des individus d'une société, pour la mémorisation de propositions incorrectes, pour la formulation des questions, pour le caractère explicite ou implicite des sollicitations, etc.

Il en ira de même dans le présent ouvrage qui aurait pu s'intituler « L'utilisation et la correction des QCM ». Or, comme le premier ouvrage, il aborde des problèmes concernant toute l'évaluation pédagogique, qu'elle procède par QCM ou non. C'est pourquoi le titre évoque les grandes préoccupations et le sous-titre rappelle que le cas des QCM est traité tout spécialement.

Rappelons enfin que ces deux ouvrages techniques s'inscrivent dans le prolongement du « Précis de Docimologie » de G. DE LANDSHEERE (1985), dont ils constituent, dans leur domaine particulier, un approfondissement.

Il peut paraître bizarre d'avoir réuni deux préoccupations, l'une centrée sur les instruments (la qualité des questions), l'autre sur les scores (la qualité des élèves). C'est que, comme l'avait déjà très justement fait remarquer B. MATALON (1971), on obtient des informations sur les personnes par les questions et sur les questions par les personnes. La matrice des réponses peut être lue dans les deux sens, mais, on le verra, avec des préoccupations et des techniques différentes.

CHAPITRE I

A l'heure où les micro-ordinateurs permettent de corriger automatiquement des épreuves scolaires, de calculer des scores, d'établir des bulletins, de gérer des banques de questions, de générer des tests « sur mesure » (pour le professeur ou pour chacun des élèves), etc..., il importe que les utilisateurs (et évidemment, les concepteurs) de ces programmes informatiques maîtrisent une série de concepts fondamentaux qui y sont mis en œuvre.

Ils ne sont pas les seuls à être concernés, d'ailleurs, car tout enseignant qui donne une note ou une cote à une série d'étudiants devrait jongler avec les mêmes notions.

Quelles sont ces notions ? Nous n'évoquerons que les principales laissant au lecteur, par l'index en fin de volume, le soin de découvrir les autres.

A. La difficulté d'une question

Il est évident que la forme de la question (QCM ou question ouverte) a un impact sur le taux de réussite, et il importe de se demander comment on peut comparer des résultats obtenus avec des questions de formes différentes.

Ce problème de COMPARABILITE sera évidemment traité, mais on ne peut s'en tenir là. Plus généralement, quelle que soit la forme de la question, il est évident que la difficulté est fonction de la compétence de la personne à qui la question est posée.

La situation est tout à fait comparable à la notion d'information : un même événement (ici une même question) n'apporte pas la même quantité d'informations (ici une même difficulté de répondre) à chaque personne, car l'information préalable (ici la compétence) varie de personne à personne.

Les conséquences d'une telle évidence ne sont pas toujours bien perçues. Il en découle en particulier qu'augmenter le nombre de personnes à qui on pose la question n'apporte pas forcément plus de précisions aux statistiques concernant le taux de réussite.

Pour illustrer ce principe, imaginons qu'une même question ait été posée à des élèves de 3e, 4e, 5e et 6e années primaires (100 élèves dans chaque année) et que les taux de réussite respectifs pour cette question soient 40 %, 50 %, 70 % et 80 %.

Remplacer ces 4 valeurs par leur seule moyenne (60 %) sous prétexte qu'elle est calculée sur quatre fois plus de sujets constitue une regrettable perte d'information.

Idéalement, il faudrait pouvoir disposer des taux de réussite de sujets de compétence connue dans la matière. Or cette préoccupation a déjà fait l'objet de beaucoup de travaux en édumétrie, et est même le point de départ du célèbre Modèle de RASCH qui permet de calculer la probabilité qu'*une* personne donnée réponde correctement à *une* question donnée.

De telles notions rendent caduques certaines accumulations « aveugles » de statistiques dans des banques de résultats.

B. La qualité d'une question

La qualité d'une question n'est, en principe *pas* liée à sa facilité, du moins en *édu*métrie, car certaines pratiques *psycho*métriques, elles, éliminent des questions « trop faciles » et les questions « trop difficiles », le but de ces pratiques étant non pas de mesurer la compétence absolue de chaque étudiant, mais de créer des différences au niveau des résultats pour pouvoir classer les personnes entre elles.

Il est possible de se donner des règles de construction de questions, mais ces règles a priori, si elles contribuent à la qualité, ne la garantissent pas. Elles doivent être complétées par une analyse a posteriori des résultats. Il existe en effet des techniques d'analyse mathématique des réponses des examinés indiquant « l'état de santé » de la question. Plus précisément, le calcul d'indices de « discrimination » pour chaque question (et pour chaque solution d'une QCM) fournit en quelque sorte la « température » de la question... et le professeur (ou le constructeur) peut savoir si la question « fait de la fièvre » et même à quel endroit précis. Pas plus que le thermomètre du médecin, ces techniques ne fournissent le diagnostic précis, la raison pour laquelle la question « ne va pas bien ». Cela peut-être parce qu'elle a été mal rédigée, ou parce qu'elle n'est pas à sa place dans cette épreuve, etc. Il reste à l'expert humain, qui interprète les indices, à procéder à une analyse clinique, mais celle-ci est largement orientée par les indices mathématiques !

C. La signification d'un score

Faut-il distinguer notation (proche de l'anglais « rating ») de cotation (belgicisme, proche de l'anglais « scoring ») ? Nous pensons que oui, afin d'augmenter la rigueur de nos débats et de notre terminologie.

Ainsi, il importerait de ne pas confondre notation et cotation, pas plus que nombres et chiffres (ex. : 427 est *un* nombre de *trois* chiffres). Or, c'est ce que certains font allègrement lorsqu'ils entament un débat sur la « notation par lettres ou par chiffres », alors qu'il faudrait plutôt parler de « la notation par lettres opposée à la cotation par nombres ».

Ces mots ont-ils une telle importance ? Oui, car ils recouvrent des différences fondamentales dans les propriétés mathématiques des notations d'une part et des cotations d'autre part... et corrélativement, dans les fonctions des unes et des autres.

A l'intérieur des notations, il importe de savoir si les *notes* sont ordinales ou simplement nominales. A l'intérieur des cotations, il importe de savoir si les *mesures* (ou cotes) sont issues d'une échelle métrique à zéro absolu ou non.

Ces considérations, qui peuvent paraître théoriques, sont au cœur des débats déjà évoqués ci-dessus, débats tant de fois stériles parce qu'éludant les principes fondamentaux qui les ont fait naître et les justifient. Sans se référer aux concepts théoriques ci-dessus, il est impensable d'aborder des questions du genre de celles qui suivent :

« Cela a-t-il un sens d'attribuer un score négatif à une performance ? Et un score supérieur au maximum ? Quel est le sens exact du score 0 ? »

Autant de problèmes que ceux qui, dans leur profession, cotent d'autres personnes, doivent pourtant avoir considéré de façon approfondie.

D. La rectification pour devinette

Il est affligeant de constater qu'un grand nombre de personnes (y compris des auteurs américains publiant sur le sujet) confondent « deviner » (en anglais *guessing*) et « répondre au hasard » (en anglais *random guessing* ou *blind guessing*). Les expressions anglaises parlent d'elles-mêmes : si un choix aléatoire est *aveugle*, c'est qu'un choix deviné *ne l'est pas* !

Faire la différence entre ces deux concepts est crucial, car elle permet d'éclairer la pertinence non seulement de formules de rectification des scores (aussi appelées formules de « correction-for-guessing ») mais d'approches originales en évaluation.

Alors que tous les enseignants qui « cotent » les travaux utilisent des barèmes de tarifs, nombreux sont ceux qui n'ont pas une vue suffisamment claire des divers barèmes possibles et des fondements du calcul de divers tarifs.

Nous avons consacré un long chapitre à la trop célèbre « correction-for-guessing » classique, et ce pour plusieurs raisons. Tout d'abord, parce que, à notre connaissance, aucun ouvrage en français ne traite en profondeur ce sujet qui a fait l'objet de centaines de publications (surtout aux Etats-Unis).

Il nous a semblé important d'en dévoiler les fondements, d'en faire la critique, et d'exposer les (nombreuses) solutions de rechange, avec leurs avantages et leurs inconvénients.

Une autre raison de se pencher sur la « correction for guessing » est qu'une fois de plus, la notion de devinette doit être (comme celle d'information transmise, ou de difficulté d'une question) rapportée *à une personne précise*.

La personnalité de l'étudiant (sa propension générale ou occasionnelle à prendre des risques) influence notablement le « guessing ». Divers chercheurs ont investigué et ont proposé des systèmes de cotation tenant compte de cette facette de la personnalité. Qu'il les adopte ou non, aucun évaluateur ne peut ignorer l'existence de ces systèmes.

E. La qualité d'une épreuve

La qualité d'une épreuve se juge à plusieurs critères, notamment à la validité ou à la fidélité des notes ou des mesures qu'elle permet d'obtenir.

Que la validité (propriété des résultats de mesurer réellement ce qu'ils prétendent mesurer) se définisse elle-même de plusieurs façons, et que certaines de ces définitions puissent faire l'objet de mesures quantitatives (indices de validité), voilà qui ne peut laisser indifférent un constructeur de tests.

Quels facteurs affectent la fidélité de l'épreuve (propriété de classer les individus toujours dans le même ordre et avec les mêmes écarts bref, la fiabilité du classement) ? En quoi l'augmentation du nombre de questions affecte-t-elle la fidélité ? Et le nombre de solutions proposées dans une QCM ? Tous problèmes qui ont reçu des réponses très précises : il existe des formules mathématiques permettant de calculer le gain de fidélité en fonction de l'allongement du test et vice-versa.

Comment prévenir la fraude (ou, à défaut la détecter a posteriori), tout spécialement dans les épreuves recourant aux QCM ? Ce problème rejaillit lui aussi sur la qualité des résultats d'une épreuve.

* * *

A la lecture de cette introduction, on aura compris le caractère technique des questions – et des réponses – auxquelles s'attaque ce livre. Il s'agit plus d'un ouvrage à consulter sur certains problèmes précis qu'un livre à lire de bout en bout.

C'est pourquoi nous avons fourni un index assez détaillé en fin de volume.

CHAPITRE II

Les indices de facilité et de difficulté d'une question

A. Les trois grandes approches.
B. L'effet du hasard sur l'indice de facilité d'une QCM.
C. L'effet des contextes de passation du test sur la facilité.
D. L'effet de la compétence des étudiants sur la facilité.

Introduction

On distingue trois grandes approches de la facilité d'une question.

Dans l'approche théorique, des experts portent des jugements sur les relations hiérarchiques entre plusieurs questions d'une même matière. La facilité de chaque question est *déduite d'un modèle.*

Dans l'approche introspective, les étudiants sont invités à exprimer la probabilité de succès qu'ils accordent à leur réponse. Ces probabilités représentent donc pour chaque question la facilité *ressentie* par chaque sujet qui y répond.

Dans l'approche expérimentale, de loin la plus pratiquée, les réussites et les échecs des étudiants constituent les données de base. La facilité de la question est donc ici *observée à partir des réponses* fournies par un groupe d'étudiants.

Des relations existent entre ces approches et on peut valider l'une par l'autre.

Divers facteurs affectent la facilité d'une QCM :
- la possibilité de faire des choix heureux en situation d'ignorance totale ;
- les circonstances de passation de l'épreuve (avant ou après l'étude de la matière) ;
- la compétence des élèves qui y répondent.

A. Trois approches

1. Les approches théoriques

a) Le principe

Des experts (enseignants, inspecteurs, chercheurs, etc.) analysent la structure et le contenu d'une question, et la comparent à d'autres questions, plus ou moins complexes.

Ces spécialistes de la discipline déterminent ainsi un ordre de complexité entre les diverses questions. Par exemple, la multiplication de deux nombres entiers de trois chiffres est supposée plus difficile que la multiplication de deux nombres entiers de deux chiffres. On énumère les capacités (comprendre des concepts, effectuer des opérations, etc.) requises par chaque question et on attribue une importance relative à chaque capacité. La somme ainsi pondérée des capacités est une des façon d'estimer la facilité (ou de la difficulté) d'une question.

Les enseignants et les chercheurs consacrent beaucoup de temps à ces analyses *a priori* qui permettent d'élaborer des hiérarchies hypothétiques sur lesquelles seront fondées les progressions des programmes, des manuels scolaires, des leçons, etc.

La manière la plus élémentaire d'élaborer de telles hypothèses pour obtenir un « indice de facilité », consiste à demander à des juges (ou « experts ») de classer les questions dans un ordre croissant de difficulté. Sur une échelle *ad hoc*, on adopte alors pour indice de facilité de la question la moyenne des scores avancés par les divers experts[1]. Par cette méthode, des indices numériques peuvent être obtenus rapidement, mais ils ne sont qu'hypothétiques et devront être confrontés aux approches introspectives et expérimentales. On trouvera un exemple de cette approche dans le travail coordonné par B. Tistaert (1975).

[1] Dans ce genre d'approche, il convient de préciser le *degré d'accord* entre experts, en mentionnant soit l'écart type de la distribution des jugements (c'est l'indice approprié dans le cas présent), soit des pourcentages d'accord.

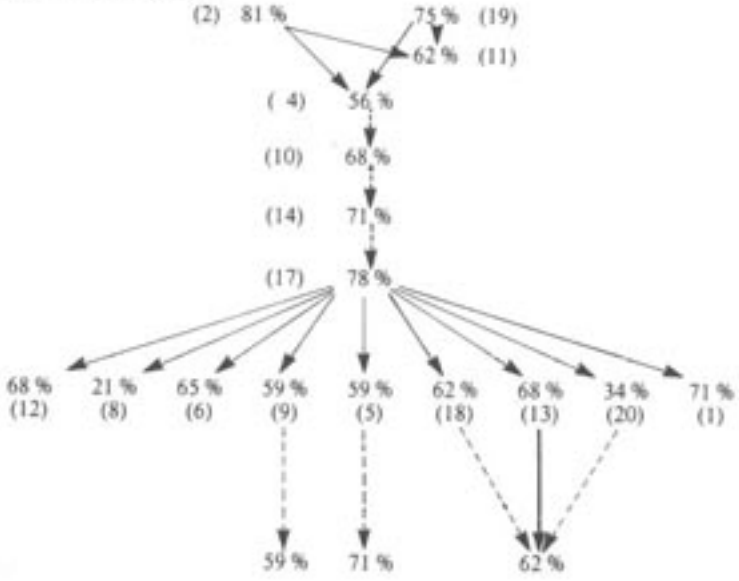
b) Un exemple

Sous notre direction, B. Lumingu (1973) a mené une étude préalable à la construction d'un test diagnostique *d'utilisation de dictionnaires*. Dans cette étude, il a fait l'analyse (en arbre) des divers comportements impliqués, puis les a structurés sous la forme d'un graphe fléché. Les flèches signifient : « Cet objectif doit être maîtrisé avant de pouvoir maîtriser cet autre objectif. (Cette capacité est prérequise à cette autre) ».

Il a ensuite soumis ces hiérarchies théoriques à l'épreuve de la réalité en présentant toutes les questions de la hiérarchie (4 tests et 20 questions par test) à des élèves de l'enseignement secondaire inférieur. Ceux-ci devaient répondre sans aide dans un premier temps, puis en disposant de trois dictionnaires dans un second temps. Les résultats expérimentaux (pourcentages de réussite avec dictionnaires) ont été portés sur le graphe hiérarchique. Lumingu a ainsi pu mettre en évidence des paires de résultats *compatibles* avec la hiérarchie théorique (représentées par des traits pleins) et des paires de résultats *incompatibles* avec cette hiérarchie (représentées par des traits discontinus). Les numéros des questions du test figurent entre parenthèses.

Dans l'exemple ci-dessous, les résultats compatibles avec la théorie sont au nombre de 14, les résultats incompatibles au nombre de 7.

Avec dictionnaires :



Grâce aux pourcentages de « réussites *sans* dictionnaire », il était possible de déceler les questions où la culture générale permettait de se passer du dictionnaire. L'analyse fine des résultats (voir techniques au chapitre 3) a permis de mettre en évidence des raisons pour lesquelles la hiérarchie théorique et les résultats divergent.

Diverses questions présentaient des particularités ignorées dans l'arbre des objectifs, qui a pu être revu à la lumière de l'analyse des résultats. En retour, la conception des questions a elle aussi été améliorée.

On voit donc que l'approche théorique et l'approche expérimentale se complètent, ou plutôt s'éclairent tour à tour. C'est le modèle théorique préexistant au test qui donne aux résultats leur grande valeur informative : on les interprète au travers d'une « grille de lecture », et les observations permettent en retour d'améliorer la grille elle-même.

Néanmoins, il est évident que si un bon « modèle » a pu être élaboré c'est notamment grâce à une bonne expérience quotidienne de l'enseignement ou de l'évaluation dans la discipline.

2. Les approches introspectives

a) Le principe

Ces approches sont basées sur les impressions de l'étudiant. Celui-ci est invité à dire dans quelle mesure il trouve la question facile. En plus de sa réponse, il indiquera à quel degré il est convaincu de son exactitude. Il utilisera, pour ce faire, soit des « pourcentages de chances », soit des « probabilités subjectives », soit des « degrés de certitude », soit des « rapports » (ex. : une chance sur trois).

Dans le cas des QCM, on peut demander à l'étudiant d'apprécier personnellement le caractère attractif de chacune des solutions proposées, c'est-à-dire leur probabilité de constituer la réponse correcte. Si l'on a annoncé à l'étudiant qu'une (et une seule) solution est correcte, la somme des probabilités doit valoir 1 (la somme des pourcentages doit valoir 100).

La moyenne des indices introspectifs fournis par une population d'étudiants pour la solution correcte peut être utilisée comme indice de facilité d'une question. De tels indices subjectifs (caractère attractif ou *attractivité*) peuvent être confrontés aux indices expérimentaux (ou popularité) et, bien entendu, aux indices théoriques (ou *complexité*). L'idéal, évidemment consiste à combiner les trois approches.

b) Un exemple

Nous avons présenté à six étudiants dix questions en leur demandant, pour chacune :

1. de distribuer des probabilités (somme = 1) sur les solutions proposées,
2. de choisir ensuite une seule solution (le choix a été souligné dans les exemples qui suivent).

Voici les résultats obtenus pour la question 6 de notre test (cette question propose 3 solutions ; la solution 1 est correcte) :

	Etudiant n°						Probabilité moyenne ou attractivité de la rép. cor.	Popularité (% de choix) de la rép. cor.
	1	2	3	4	5	6		
Solution 1	0,80	0,45	0,55	0,0	0,50	0,60	0,483	83 %
Solution 2	0,10	0,10	0,15	0,30	0,10	0,20	0,158	0 %
Solution 3	0,10	0,45	0,30	0,70	0,40	0,20	0,358	17 %
Total des probabilités	1	1	1	1	1	1		

On voit (aux valeurs soulignées) que chaque étudiant a choisi la solution à laquelle il avait accordé la probabilité la plus élevée. Ce comportement des étudiants est une preuve de leur cohérence : ils agissent dans leur intérêt et (sans le savoir) conformément à la théorie des décisions. Cette dernière recommande en effet de « maximiser » les bénéfices attendus ou espérés. Les termes « attendus » et « espérés » indiquent qu'il faut prendre en considération la probabilité (estimée subjectivement) que ces bénéfices surviennent[1].

Les étudiants d'un certain âge (plus de douze ans, stade des opérations formelles) agissent spontanément de la sorte. Des expériences récentes laissent croire que des enfants plus jeunes encore peuvent s'auto-évaluer et prendre des décisions de façon pertinente (c'est-à-dire ici conformes à leurs intérêts).

On constate que, sur de petits nombre d'étudiants, les indices d'*attractivité* et de *popularité* des solutions proposées peuvent être très dissemblables. Ainsi, pour la question 6, la solution correcte a une attractivité moyenne de 0,48 et une popularité moyenne de 0,83.

[1] Si l'on appelle TC le tarif prévu pour une réponse correcte, ps/x probabilité subjective que la réponse soit correcte étant donné que l'on a choisi la solution x et SAQ/x le score attendu à la question étant donné qu'on a choisi la solution x, on peut écrire que SAQ/x = ps/x. TC. La théorie des décisions prévoit que si ps/y est plus élevé que ps/x, alors SAQ/y est plus élevé que SAQ/x et l'on *doit* préférer le choix de la solution y au choix de la solution x.

Ce même phénomène peut aussi être observé pour la question 10 de notre test (cette question propose 3 solutions, la solution 2 est correcte) :

	Etudiant n°						Probabilité moyenne (attractivité) de la rép. cor.	Popularité (% de choix) de la rép. cor.
	1	2	3	4	5	6		
Solution 1	0,15	0,10	0,20	0,25	0,15	0,05	0,166	0 %
Solution 2	0,70	0,80	0,75	0,50	0,5 0	0,90	0,691	100 %
Solution 3	0,15	0,10	0,05	0,25	0,25	0,05	0,142	0 %

Dans les deux exemples ci-dessus, les deux colonnes de droite (probabilité moyenne et popularité) ne correspondent pas. La première est le reflet d'un *processus*, la seconde celui d'un *résultat*. Selon que l'on s'intéressera à l'un ou à l'autre, on préférera l'une ou l'autre procédure d'analyse, l'un ou l'autre indice numérique.

Seules les probabilités subjectives fournies par l'étudiant peuvent rendre compte à la fois du choix et de l'attractivité des solutions pour les individus. C'est pourquoi nous les recommandons.

3. Les approches expérimentales

Ces approches ont eu un grand succès dès la naissance de la psychométrie. L'indice le plus utilisé fut – et reste – la proportion (ou fréquence relative) *des réponses correctes* fournies à la question par une population donnée d'étudiants. Cet indice est le plus souvent désigné par p dans la littérature classique, mais nous utiliserons pc par souci d'uniformité :

pc = proportion de réponses correctes.
pi = proportion de réponses incorrectes.
po = proportion d'omissions.
pe = proportion d'échecs (pi + po).

Toutes ces proportions (ou fréquences relatives) varient entre 0 et 1.

Dans la littérature habituellement, on désigne p (pour nous pc) comme l'indice de la *difficulté* d'une question. Récemment, divers auteurs comme Wood (1977, p. 241) et Ebel (1979, p. 263) ont dénoncé le caractère paradoxal d'une telle expression ; en effet, la difficulté s'*élève* à mesure que la fréquence des réponses correctes *baisse*. Ces auteurs proposent de rompre avec quatre-vingts ans de littérature et d'appeler « *indice de facilité* » la proportion de réponses correctes (pc) et « indice de difficulté » la proportion d'échecs (réponses incorrectes + omissions = pe). Bien entendu, pc et pe sont complémentaires : leur somme vaut 1.

Nous adopterons cette récente convention terminologique, d'ailleurs rendue souhaitable par les nouveaux modèles qui seront analysés en fin de chapitre.

B. L'effet du hasard sur l'indice de facilité d'une QCM

Il est évident que si trente étudiants sur cent réussissent (pc = 0,3) une QCM à quatre solutions proposées, on ne peut conclure que 30 % d'entre eux connaissent la réponse correcte, car un certain nombre de succès sont dus à des choix heureux faits au hasard. La valeur pc pour une QCM n'est pas directement comparable à la valeur pc pour une question ouverte.

Diverses formules « de correction » tentent de rendre ces deux valeurs comparables.

1. La correction « de Davis »

Davis (1946) a proposé une célèbre formule de correction de l'indice de facilité (basé sur pc). Nous en exposerons le principe de base, puis le développement mathématique. Signalons que les symboles N désignent des nombres (de réponses) et les symboles p des proportions ou fréquences.

Le raisonnement de départ est le suivant :

- Supposons une QCM, à k solutions proposées, pour laquelle on dispose de N réponses d'étudiants. NC (= réponses correctes) ont répondu correctement, NI incorrectement et NO ont omis (avec NE = nombre d'échecs ; NE = NI + NO).
- NC doit être décomposée en NCC (nombre de sujets ayant fourni la réponse correcte par connaissance) et NCH (nombre de sujets ayant choisi la solution correcte par hasard ou « *guessing* » aveugle).
- La facilité serait mieux estimée par NCC que par NC ; il suffirait de calculer NCC = NC – NCH. Malheureusement, ni NCC, ni NCH ne sont connus.
- On va donc essayer d'*estimer* NCC au départ des seules valeurs connues (NC, NI et NO) : $\boxed{NCG = NC - \frac{NI}{k-1}}$ (1) formule 1 (classique)

[1] *Détails du calcul de NCG, estimation de NCC*

Pour estimer NCE, Davis propose un indice que nous appellerons NCG (réponses correctes corrigées « pour *guessing* ».
Cet indice NCG est fondé sur les développements suivants :
- Si NH est le nombre (inconnu) de personnes qui ont répondu au pur hasard, NCH vaut $\frac{NH}{k}$ et $NI = NH - \frac{NH}{k}$

  car on part du principe que toutes les erreurs (NI) sont le résultat de réponses au hasard.
- Il suffit d'exprimer NH au moyen de NI pour trouver NCH :

On peut appliquer la formule de DAVIS aux fréquences :

$$pcg = pc - \frac{pi}{k-1}$$

avec pcg = proportion de réponses correctes corrigées « pour devination » (*estimation* de pcc).

2. La formule de Davis « surcorrige »

Rappelons que cette méthode de correction de la facilité d'une QCM est basée sur le principe que *toutes* les réponses incorrectes (NI) sont fournies au hasard. Or un certain nombre (inconnu du professeur) de réponses incorrectes sont fournies « en toute bonne foi ». La formule de Davis constitue donc une *surcorrection.*

a) Un exemple

Ainsi, considérons une QCM à 5 solutions proposées présentée à cent étudiants. On a observé soixante réussites (NC = 60) et quarante erreurs (NI = 40). Imaginons que, parmi ces quarante personnes, trente se soient trompées de bonne foi (NI = 30) et dix en répondant au hasard (NIH = 10). Dans la formule, c'est NIH et non NI qu'il faudrait prendre en compte :

Avec la formule initiale, utilisant NI, on obtiendra :

$$NCG = 60 - \frac{30}{4} = 60 - 7{,}5 = 52{,}5$$

Avec NIH dans la formule, on obtiendrait :

$$NCG = 60 - \frac{10}{4} = 60 - 2{,}5 = 57{,}5$$

Malheureusement, NIH est inconnu.

$NH - NH/k = NE$

$NH\,(1 - \frac{1}{k}) = NE$

$NH = RE / (1 - \frac{1}{k})$

$NCH = NH/k = NE / (1 - \frac{1}{k})\,k = NE / (k - \frac{k}{k}) = NE / (k - 1)$

– On peut donc calculer NCG, estimateur de NCC, par la formule de DAVIS :

$$NCG = NC - \frac{NE}{k-1}$$ formule 1 (classique)

b) Une hypothèse

Le pourcentage de sujets qui connaissent la réponse correcte à la question serait donc inférieur à la fréquence de réponses correctes observées, mais supérieur à la fréquence de réponses correctes « corrigées pour guessing » (NCG) calculée par la formule de Davis (conclusion 1).

Est-il possible d'en préciser la valeur ?

c) Une solution

On peut partir d'un raisonnement inspiré des travaux de Ziller (1957) : un étudiant qui présente des omissions est moins suspect de fournir des réponses au hasard qu'un étudiant qui n'en présente pas. On trouvera la justification et le développement de la formule de Ziller au chapitre 5.

Ce raisonnement s'applique aussi pour des ensembles des réponses fournies par plusieurs étudiants.

Si, pour une question, on appelle NO le nombre d'étudiants qui se sont abstenus, NI / (NI + NO) peut estimer le rapport du nombre de réponses au hasard au nombre d'individus ne connaissant pas la réponse. Cette valeur représente la tendance du groupe d'élèves à répondre au hasard. On peut s'en servir pour nuancer NCG calculé ci-devant.

Le nouvel indice ainsi obtenu, que nous appellerons NCZ (avec le Z de Ziller), est, comme NCG, une tentative pour estimer NCC, le nombre de réponses correctes fournies par « compétence ».

$$NCZ = NC - \left(\frac{NI}{k-1} \cdot \frac{NI}{NI + NO} \right)$$

Cette formule ne s'applique pas quand NI et NO sont nuls.

Si, dans notre exemple, NC = 60, NI = 20 et NO = 20, alors

$$NCZ = 60 - \left(\frac{20}{4} \cdot \frac{20}{40} \right) = 60 - \frac{5}{2} = 57{,}5.$$

Il peut arriver que NCZ (donc pcz) soit négatif. On peut penser dans ce cas que les sujets sont plutôt mal informés sur la question que non informés.

3. La formule de Davis « sous-corrige »

On peut tout aussi bien prétendre que la formule 1 aboutit à une *sous-correction* : les étudiants ont très souvent des connaissances partielles qui leur permettent d'éliminer certaines solutions incorrectes et de choisir au hasard non plus parmi k, mais parmi k−1 ou k−2 solutions. Or, la formule classique ci-dessus suppose que la solution correcte a une attractivité égale à 1/k.

a) Un exemple

Ainsi, revenons à notre exemple d'une QCM (à 5 solutions proposées) présentée à cent étudiants. On a observé que QRC = 60 et QRI = 40. Supposons maintenant qu'il soit facile d'éliminer deux des cinq solutions ; k serait modifié (il vaudrait 3). On obtiendrait alors :

$$NCG = 60 - \frac{40}{2} = 60 - 20 = 40.$$

Cette nouvelle valeur est largement inférieure à NCG calculé par la formule de Davis (NCG valait 50) qui surestime donc la capacité et sous-corrige !

b) Une hypothèse

La proportion sujets qui connaissent la réponse correcte serait donc inférieur à pcg calculé par la formule de Davis (conclusion 2).

4. Une solution d'ensemble

Les deux raisonnements ci-dessus aboutissent donc, à des conclusions (1 et 2) contradictoires, et à une impasse. La seule issue consiste à demander aux étudiants d'exprimer les probabilités subjectives qu'ils attribuent à l'exactitude de leur réponse.

Le taux de connaissance du groupe sur une question pourrait alors être estimé par la moyenne de ces probabilités subjectives. Cette moyenne ne devrait cependant être calculée que sur les données fournies par les seuls individus qui s'auto-évaluent correctement. Nous avons proposé dans un autre ouvrage, intitulé « Connaissance partielle et auto-évaluation » des moyens de mesurer le réalisme.

C. L'effet des contextes de passation du test sur la facilité

Habituellement, une épreuve scolaire est présentée en *bloc* : au moment où l'étudiant répond à une question, il n'a pas reçu d'information sur l'exactitude de ses réponses aux questions antérieures. Si cette façon de tester est la plus répandue, elle n'est cependant qu'une des nombreuses possibilités d'administration d'une épreuve. Aussi examinerons-nous les effets d'autres modalités telles que :

- la présentation fractionnée de l'épreuve ;
- la correction immédiate des réponses ;
- les méthodes d'apprentissage et de *testing* ;
- le moment du *testing* (avant et après l'apprentissage).

1. La présentation fractionnée de l'épreuve

Il est possible d'administrer *en plusieurs étapes* une épreuve portant sur une matière déterminée. Ainsi, dans les modules à augmentation de la difficulté, les questions de compréhension et d'application sont mieux réussies si les questions de connaissances préalables ont été corrigées.

Nous avons fait une telle expérience à partir de deux modules d'auto-évaluation de physique. Les étudiants ont d'abord répondu à l'ensemble des questions. Ils ont ensuite pris connaissance des réponses correctes aux questions de niveau « connaissance de mémoire ». Après cela, les étudiants ont été invités à répondre à nouveau aux seules questions des autres niveaux (compréhension, application, processus supérieurs). Ils ont donc répondu deux fois aux mêmes questions, avant et après une information. L'impact moyen de cette information est appréciable quand on compare les pourcentages moyens de réussite *avant* et *après* l'information.

Séquence	Avant	Après
Notion de pression (42 étudiants)	53,82 %	71,41 %
Principe d'Archimède (34 étudiants)	32,15 %	55,88 %

2. La correction immédiate des réponses

On peut pousser à l'extrême la procédure de présentation fractionnée. Il suffit de révéler à l'étudiant la réponse correcte à une question dès qu'il y a répondu. Cette information immédiate rend, bien entendu, la réponse aux questions qui suivent plus facile. Cette amélioration du rendement est encore plus nette quand on invite l'étudiant à s'auto-corriger en cas d'erreur.

Dans un Programme AutoCorrectif à Embranchements sous forme de Livre Brouillé[1] à Réponses Ouvertes (PACELBRO), nous avons appliqué ce principe[2]. De plus, quand l'étudiant était incapable de trouver son erreur ou de la corriger, des aides de plus en plus importantes lui étaient fournies.

Une telle procédure conduit à des taux de réussite plus élevés que la présentation en bloc, où une erreur du même type peut se reproduire dans plusieurs questions. Grâce à la *correction immédiate*, les erreurs corrigées lors des questions antérieures ont une moindre probabilité de se reproduire dans la suite de l'épreuve.

C'est à cette constatation que nous avons abouti quand nous avons testé la technique PACELBRO sur des questions d'arithmétique en sixième primaire. Ces questions étaient reprises du test d'A. Bonboir (1960), test qui avait été présenté en *bloc* à près de 3.000 élèves[3]. Les résultats d'A. Bonboir constituaient donc une bonne base de référence pour savoir si la correction immédiate augmente les taux de réussite des questions.

Voici les résultats obtenus pour cinq questions portant sur la notion de pourcentage. Nos résultats sont calculés sur 200 élèves de sixième année primaire (testés en 1967).

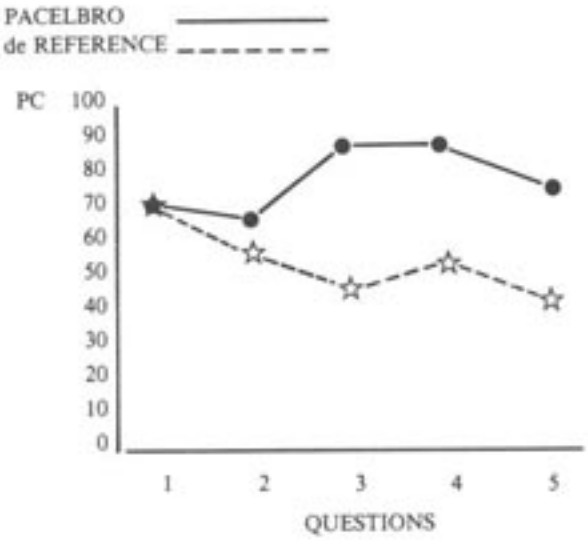
[1] Dans un *livre brouillé*, l'étudiant est renvoyé d'une page à une autre non pas dans l'ordre séquentiel, mais en fonction de ses réponses. Par *réponse ouverte*, on signifie que l'étudiant doit rédiger sa réponse et non pas la choisir (réponse fermée).
[2] Ce programme a été décrit dans D. Leclercq, J. Donnay et R. De Bal, 1977.
[3] Les 686 questions du test (regroupées en 13 catégories) ont été présentées à 2871 élèves.

Ordre	Question	Taux de réussite (en % de réponses correctes)	
		En bloc (sur 2871 élèves) (A. Bonboir)	PACELBRO (sur 200 élèves)
1	5 % de 600 f =	71,05	71,00
2	440 f x 2 % =	55,72	66,80
3	600 f x 2,5 %	48,18	85,91
4	412 f x 6 % =	53,72	85,12
5	5 % des 4 % de 300 f =	40,16	74,45

(Taux de référence)

On le voit, les taux de réussite aux cinq questions varient selon que l'épreuve est administrée « en bloc » ou avec correction immédiate.

Le graphique ci-dessous fait en outre apparaître que l'amélioration n'est pas immédiate.



3. Les méthodes d'apprentissage

Les popularités des solutions d'une QCM varient *selon les moments de l'apprentissage et les méthodes d'instruction.* Ainsi, D'Hainaut (1971, vol. 2, p. 179) a proposé la question suivante à quatre échantillons d'étudiants.

> Pour faire dévier un corps en mouvement en terrain plat, il faut lutter contre :
> 1. Son poids.
> 2. Sa masse.
> 3. Son poids et sa masse.
> 4. Ni l'un ni l'autre.

Echantillons	La question est appliquée à des élèves qui...	Taux de réussite (pc)
A (12 ans)	... sont « de niveau nul » (selon l'expression de d'Hainaut) c'est-à-dire n'ont quasiment eu aucune chance d'avoir été exposés à la matière.	2 %
B (14 ans)	... n'ont pas reçu de cours.	13 %
C (14 ans)	... viennent de recevoir un cours magistral sur le sujet (notion de masse).	48,5 %
D (14 ans)	... viennent de recevoir un cours programmé sur le sujet.	78,5 %

On voit combien il importe de préciser sur quelle population, à quel niveau d'apprentissage (et même avec quelle méthode) des indices mathématiques comme le taux de réussite (pc) ont été obtenus.

On voit aussi combien il serait dangereux de « mêler » plusieurs groupes de résultats sous le prétexte de travailler sur de plus grands nombres.

4. Les méthodes de testing

La facilité d'une question pour un groupe déterminé d'étudiants peut aussi varier *avec les modalités de testing.* Ainsi, Johnson et Rosenthal (1978) ont présenté un texte puis 42 questions de compréhension à trois groupes d'étudiants. Un groupe A (99 étudiants) a reçu les questions immédiatement après le texte et un groupe B (48 étudiants) vingt et un jours après. Un groupe C (96 étudiants) a reçu les questions sans avoir lu le texte. Les facilités moyennes des questions furent respectivement 78 %, 48 % et 24 %.

D. *L'effet de la compétence des étudiants sur la facilité de la question*

Dans la section A du présent chapitre, le problème de la taille de l'échantillon a été soulevé. Plus le nombre d'étudiants considérés augmente, plus l'on risque de rassembler des étudiants de compétences différentes. La présente section est consacrée à l'examen d'indices établis d'une part sur des critères nationaux, d'autre part sur des critères de compétence. Cette section tentera de montrer que le *taux de réussite* (qui, on l'a vu, pose tant de problèmes d'interprétation) est loin d'être le seul indice possible de la facilité d'une question. Pour aboutir à ces nouveaux indices, issus du modèle de Rasch, il est utile de faire encore un petit bout de chemin avec le *taux de réussite* dont on percevra de plus en plus les inconvénients.

1. Un indice pour tous les individus d'un même pays

Il est possible de calculer pour une question un indice de facilité valable pour toute une population nationale, en prenant par exemple le taux de réussite à la question par un échantillon représentatif de tous les individus de ce pays. Une telle mesure peut être utile pour comparer divers pays entre eux, comme c'est le cas dans les recherches de l'I.E.A.[1].

La valeur d'un indice « national » peut être instructive pour les responsables de l'enseignement.

2. Un indice pour tous les individus d'un même niveau scolaire

A l'intérieur d'un pays, on peut poser une même question à divers niveaux scolaires et calculer un indice pc pour chacun d'eux. On peut aussi présenter une matière à différents niveaux scolaires et enregistrer les divers taux de succès ; comme pour une question, la matière aura alors plusieurs indices, un par niveau scolaire testé.

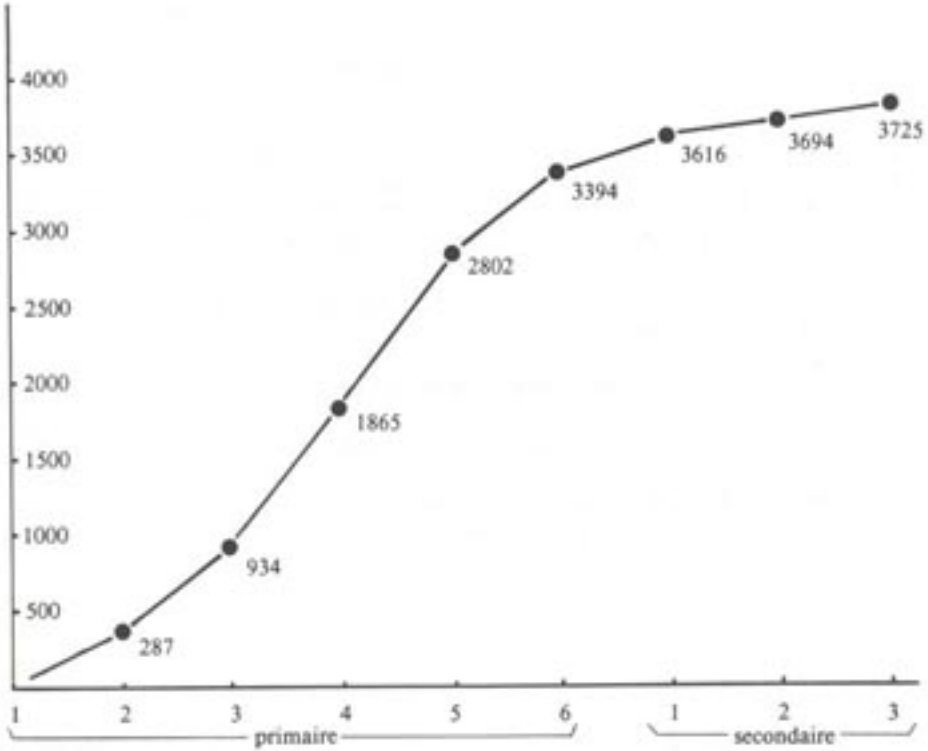
[1] L'International Evaluation of Educational Achievement est une association dirigée par des experts internationaux (HUSSEN, BLOOM, DE LANDSHEERE, POSTLETHWAITE, etc.). Le lecteur intéressé par les résultats de recherche menées par cette association lira avec intérêt DE LANDSHEERE, (1980), GRISAY (1974), HENRY (1974), LORET (1980), HENRY et MASSOZ (1979).

a) Un exemple

Ters, Mayer et Reichenbach (19, p. 24) ont considéré 4000 mots de l'échelle d'orthographe de Dubois-Buysse.

Lorsque 75 % des enfants d'une année scolaire réussissent à écrire un mot sans erreur, ce mot est *considéré* comme *acquis* au cours de cette année scolaire[1].

Le tableau ci-dessous présente en ordonnée le nombre de mots acquis et en abscisse les huit niveaux scolaires différents étudiés. Les huit résultats, représentés par de petits cercles, ont été reliés par une ligne dont l'allure générale prend la forme d'une S majuscule.



[1] La valeur-repère de 75 % a été choisie arbitrairement, un peu par tradition : dans le test de Binet-Simon d'intelligence générale, une question est considérée comme typique d'un âge lorsqu'elle est réussie par 75 % des enfants de cet âge.

b) Interprétation de la courbe

L'allure de la courbe indique l'existence de trois phases d'apprentissage.

Phase 1 (ici avant la deuxième primaire)

La pente de la courbe est douce. Il faut consacrer beaucoup de temps d'apprentissage pour enregistrer un progrès relativement faible (nombre de mots nouveaux orthographiés correctement).

Phase 2 (ici, entre la deuxième primaire et la sixième primaire).

La pente de la courbe est forte. A tout effort supplémentaire (ici, une année d'étude) correspond un important accroissement de rendement (ici, l'acquisition d'un grand nombre de mots supplémentaires).

Phase 3 (ici, de la sixième primaire à la troisième secondaire).

L'apprentissage plafonne, la courbe tend vers une asymptote horizontale : on ne fait plus beaucoup de progrès en un an (peu de mots nouveaux sont acquis).

Des courbes semblables pourraient être tracées pour :

- l'apprentissage de la lecture élémentaire, en mettant en abscisse les mois des trois premières années et en ordonnée le nombre de mots que l'enfant est capable de lire ;
- l'apprentissage du vocabulaire d'une langue étrangère ;
- etc.

c) Critique de la méthode

La notion d'année scolaire présente cependant certains inconvénients. On sait que dans une même classe sont rassemblés des individus forts et des individus faibles dans une même matière, et que la différence qui les sépare peut parfois correspondre à plusieurs années scolaires.

C'est pourquoi on a cherché à calculer un indice de facilité d'une question pour tous les individus qui ont *la même compétence dans un domaine,* quel que soit leur niveau scolaire ou leur année d'étude.

3. Un indice pour tous les individus d'un même niveau de compétence dans le domaine considéré

a) Le principe

Si l'on dispose d'une épreuve valide, donnant un score total représentatif de la capacité de l'individu dans le domaine considéré, on peut constituer

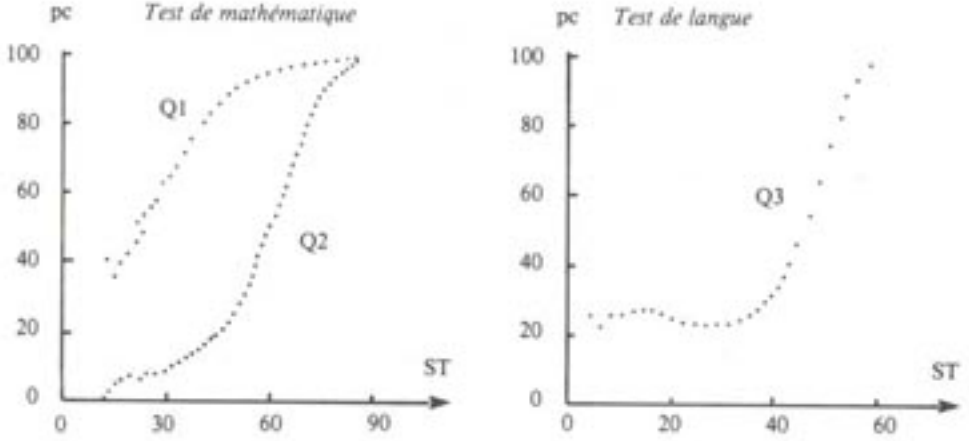
des groupes d'étudiants sur la base de ce score total. Ainsi, considérons une épreuve (valide) de 20 questions présentée à plusieurs milliers d'étudiants. On peut appeler « groupe O » le groupe constitué de tous ceux qui ont obtenu un score nul au total de l'épreuve. De même, on constituera les « groupe 1 », « groupe 2 », « groupe 3 »... jusqu'au « groupe 20 », groupes dont la capacité dans le domaine va en augmentant.

Si l'on considère *une des vingt questions*, on peut calculer ses taux de réussite (pco pcl... pc20) *pour chaque groupe*. Bien entendu, ces valeurs de pc ne sont calculées que si chaque groupe compte un nombre de sujets suffisant (50, par exemple).

b) Des exemples

Lord et Novick (1968, p. 364) ont procédé à de tels calculs.

Ils ont ensuite porté les scores totaux en abscisse (axe de la compétence des groupes) et les valeurs de pc en ordonnée (axe de taux de réussite des questions). Bien entendu, les pc sont toujours compris entre 0 et 1 (ici entre 0 et 100 %). Voici des graphiques qu'ils ont ainsi obtenus.



ST = score de chaque groupe au total du test.

Pour la question 1, le groupe le plus faible (13 points sur 90 au total du test) réussit à 40 %. Pour la question 2, le même groupe obtient 0 % de réussite.

La question 3 est une QCM à 4 solutions. Elle fait partie d'un autre test. Le groupe le plus faible (3 points sur 60) réussit la question à 25 %.

La disposition des points constitue dans les trois cas, une courbe en S majuscule. (Plus ou moins complète).

Dans ses travaux en collaboration avec Birbaum et Novick, Lord (1972) développe une « théorie des traits latents ». Dans cette théorie, *la facilité d'une question est représentée* non plus par une valeur unique (un point) mais par *une courbe caractéristique de cette question*, ce que les américains appellent *item characteristic curve (I.C.C.)*.

4. Une courbe caractéristique pour chaque question

a) Le principe

Cette courbe est du type logistique :

- Sa forme est celle d'un S dont les extrémités tendent vers des horizontales.
- Elle présente un *point d'inflexion* (endroit où la courbe change de sens).
- Sa formule est du type $\frac{x}{x+1}$ (cette expression est le *logit de x*, d'où l'expression « courbe logistique »).

L'ordonnée de cette courbe est soit le taux de réussite observé (pc compris entre 0 et 1) soit la probabilité de réussite calculée ou estimée (prob. comprise de la même façon entre 0 et 1).

L'abscisse de cette courbe reflète la compétence de l'étudiant (exprimée sur une échelle à définir). On voit que pour un individu dont la compétence est faible, la probabilité de réussite est faible.

Dans ce type de théorie, chaque question a *sa* courbe caractéristique.

Nous pensons qu'il faut persévérer dans cette voie ; c'est ainsi que l'on s'exprime dans d'autres disciplines, par exemple en psychologie développementale.

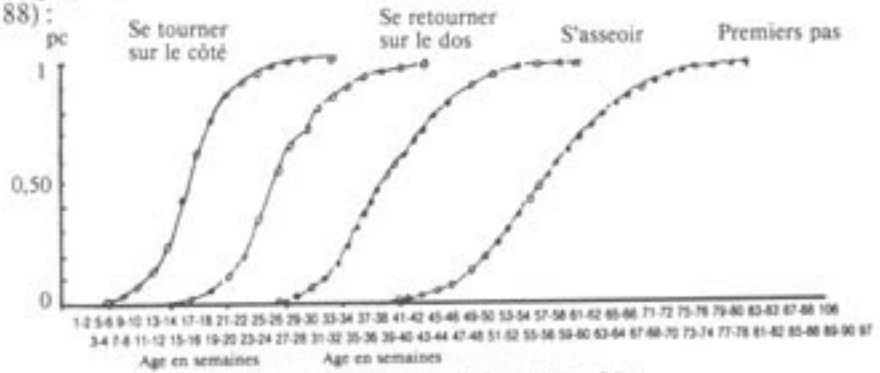
b) Les courbes caractéristiques du développement

Si l'on considère non plus des questions, mais des étapes du développement psychomoteur, telles que se tourner sur le côté, se retourner sur le dos, s'asseoir, effectuer ses premiers pas, on peut aussi observer des courbes caractéristiques.

Si l'on note pour chaque âge (exprimé en semaines) le pourcentage d'enfants de cet âge qui maîtrise le comportement considéré, alors on observe que chaque comportement a *sa* courbe en S, caractérisée par sa pente (son escarpement) et surtout sa position sur l'axe des âges (plus la courbe est à

droite, plus le comportement est « difficile » à acquérir (ou d'apparition tardive).

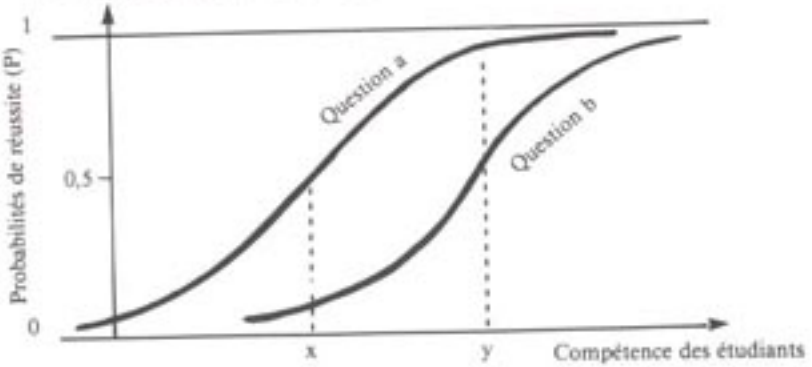
Le graphique ci-dessous est établi à partir des graphiques de Pikler (1971, p. 88) :



Courbes de distribution des âges d'apparition des différents mouvements (d'après les données des enfants à poids normal à la naissance).

c) *Implications pour le modèle*

Nous avons placé sur un même graphique les courbes caractéristiques de deux questions différentes (exemple fictif). On verra que c'est l'interaction de la difficulté de la question et de la compétence de l'étudiant qui produit la probabilité de réussite.



On remarque que la courbe caractéristique de la question *a* est située plus à gauche que celle de la question *b*. L'étudiant X (moins compétent que l'étudiant Y) a 50 % de chances de réussir la question *a* et seulement 10 % de chances de réussir la question *b*. L'étudiant Y a 90 % de chances de réussir la question *a* et 50 % de chances de réussir la question b.

c) *Quatre règles*

Règle 1 :

> Plus une courbe caractéristique est placée à droite (sur l'échelle horizontale), plus la question est *difficile*.

Manipuler l'équation d'une courbe pour réfléchir sur la facilité d'une question est peu commode. Il est plus économique d'en manipuler certains paramètres seulement, comme l'abscisse de son point d'inflexion et sa pente (son escarpement).

Règle 2 :

> Plus une courbe caractéristique d'une question a son point d'inflexion à droite sur l'échelle horizontale, plus cette question est difficile.

Si l'on veut s'exprimer schématiquement, on peut dire d'une question que sa *difficulté vaut X* pour signifier que son point d'inflexion est à la verticale de la compétence X. La question *a* a une difficulté X tandis que la question *b* a une difficulté Y.

Règle 3 :

> Quand on présente une question de difficulté X à un étudiant de compétence X, la probabilité de réussite vaut 0,5 (une chance sur deux).

On voit que l'on peut exprimer les difficultés des questions (plus exactement les points d'inflexion de leur courbe caractéristique) dans la même échelle que la compétence des étudiants. C'est *la règle 4* :

Règle 4 :

> Echelle des compétences des étudiants
> =
> Echelle des difficultés des questions.

On voit enfin pourquoi nous avons pris soin de parler ici de difficultés et non de facilités : plus l'indice est élevé, plus la question est difficile.

L'ouvrage de Wright et Stone (1979) expose comment sont calculées ou estimées d'une part, la compétence d'un sujet particulier et d'autre part, la difficulté d'une question particulière. Il est impossible d'expliquer ici ces longues procédures souvent confiées à l'ordinateur[1] ; nous ne décrirons que le modèle sous-jacent.

[1] Lire aussi B. Choppin (1975), J. Stene (1977) et G. Fischer (1977).

5. Le modèle de Rasch

a) Le principe

C'est le statisticien danois, G. Rasch (1960, 1980), qui a, le premier, eu l'idée d'exprimer la probabilité de succès d'un étudiant donné à une question donnée en fonction *de la différence entre deux paramètres* seulement : la compétence de l'étudiant et la difficulté de la question, ces deux valeurs étant exprimées sur une même échelle de mesure. C'est *la règle 5* :

> La probabilité de succès à une question, étant donné une compétence C et une difficulté D, ne dépend que de la différence entre C et D.

b) La formule de base

Si l'on appelle C la compétence de l'étudiant et D la difficulté de la question, alors :

$$P(1 \mid C,D) = \frac{e^{C-D}}{1 + e^{C-D}}$$

(formule classique du modèle de Rasch).

La différence (C-D) apparaît en exposant du nombre e (2,7183). P (1 | C,D) signifie « probabilité de succès (symbolisé par le chiffre 1) étant donné (barre verticale) une compétence C d'étudiant et une difficulté D de question ».

Nous invitons le lecteur à calculer les valeurs de P puis à dresser la courbe caractéristique d'une question pour D = 4 et pour les valeurs de C allant de 1 à 9, par exemple. Ces neuf valeurs de P permettront de retrouver la courbe logistique attendue.

On voit maintenant pourquoi si C = D, alors P = 0,5. En effet, si C-D = 0, on a $e^0 = 1$ et la formule générale vaut $\frac{e^0}{1+e^0} = \frac{1}{1+1} = 0,5$.

Le choix du nombre e est purement arbitraire. Toute autre valeur numérique peut convenir, puisque C et D sont exprimés dans les mêmes unités. D'autre part, quelle que soit la valeur de x, la valeur X^0 est toujours 1.

Les pédagogues ont cherché à remplacer e par un nombre dont les propriétés numériques seraient commodes et qui permettrait de travailler sur des échelles familières.

6. Les Wits

a) Le principe

B. Choppin (1975) utilise, en remplacement du nombre dans la formule de Rasch, un nombre, désigné par la lettre W, qui vaut 1,24573 et qui présente d'intéressantes propriétés numériques.

En effet, $W^0 = 1$ (évidemment)
$W^5 = 3$
$W^{10} = 9$

Si l'on se souvient que $W^{-a} = \frac{1}{W^a}$, on voit que $W^{-5} = \frac{1}{3}$ et que $W^{-10} = \frac{1}{9}$.

L'échelle des compétences des étudiants ainsi que l'échelle des difficultés des questions s'expriment en *wits*, mot créé sur le modèle du mot *bit (binary digit)*, à partir de la lettre W[1].

Les compétences des étudiants et les difficultés des questions varient entre 0 et 100 wits. Dans un groupe de questions, celles qui ont une difficulté « centrale » ont un indice proche de 50 wits.

Grâce aux valeurs remarquables de W, la formule de Rasch est très commode à interpréter :

$$P(1 \mid C.D) = \frac{W^{C-D}}{1 + W^{C-D}}$$

Ainsi, si un étudiant reçoit une question dont la difficulté est inférieur de 10 wits à sa compétence, on a :

$$P = \frac{W^{10}}{1 + W^{10}} = \frac{9}{1+9} = 0,9$$

Cet étudiant a donc 90 chances sur 100 de réussir cette question.

A l'inverse, si la difficulté de la question est supérieure de 10 wits à sa compétence,

$$P = \frac{W^{-10}}{1+W^{-10}} = \frac{\frac{1}{9}}{\frac{9}{9} + \frac{1}{9}} = \frac{\frac{1}{9}}{\frac{10}{9}} = \frac{1}{10} = 0,1.$$

il a dix chances sur cent de réussir cette question.

[1] . Woodcock et . Dahl (1971) avaient présenté une « échelle W » et Wilmott et Fowles (1974) une échelle de Wits. En Australie, cette même échelle est utilisée sous le nom de Brytes : ce sont des wits dont la valeur 50 est arbitrairement associée à un certain niveau conceptuel de la discipline.

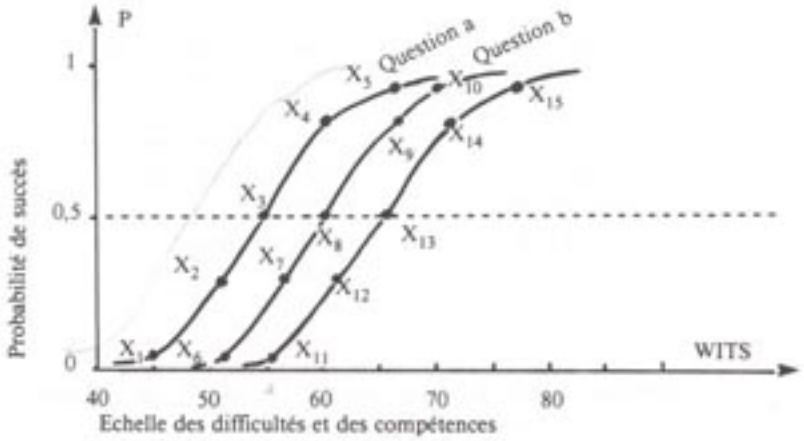
De même, si C-D = 5, alors $W^{C-D} = 3$ et $P = 3/4 = 0{,}75$.

si C-D = −5, alors $W^{C-D} = 1/3$ et $\dfrac{\frac{1}{3}}{\frac{3}{3}+\frac{1}{3}} = \dfrac{\frac{1}{3}}{\frac{4}{3}} = \frac{1}{4} = 0{,}25$.

si C-D = 0, alors $W^{C-D} = 1$ et $P\ \dfrac{1}{1+1} = \frac{1}{2} = 0{,}5$.

b) Des exemples

Voici quelques courbes caractéristiques de questions.



Le tableau suivant indique les diverses probabilités (étant donné C et D) représentées sur le graphique ci-dessus par des lettres x numérotées.

Difficultés des questions (en Wits)	Compétences des étudiants (en WITS)						
	45	50	55	60	65	70	75
a = 55	x_1=.10	x_2=.25	x_3=.50	x_4=.75	x_5=.90		
b = 60		x_6=.10	x_7=.25	x_8=.50	x_9=.75	x_{10}=.90	
c = 65			x_{11}=.10	x_{12}=.25	x_{13}=.50	x_{14}=.75	x_{15}=.90

On voit que toutes les courbes ont la même pente et que leur point d'inflexion est toujours à la même hauteur (.50). *Le seul paramètre* est donc le point d'inflexion ; nous désignerons ce paramètre par la lettre D (symbole de la difficulté de la question).

7. Le modèle de Lord et Novick

a) Le principe

Au lieu de ne considérer que D (comme dans le modèle de base de Rasch), Lord et Novick utilisent trois paramètres :

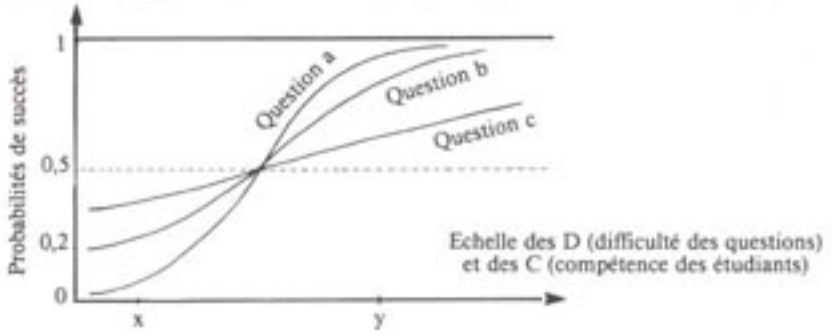
- Le paramètre D, correspondant au point d'inflexion de la courbe et représentant la difficulté de la question.
- Le paramètre A, correspondant à la pente de la courbe et représentant la valeur discriminative de la question (voir chapitre suivant).
- Le paramètre G, correspondant à la valeur plancher de la courbe (ordonnée à l'origine) et représentant la probabilité d'obtenir une réponse correcte grâce au seul hasard.
  La lettre G rappelle le mot *guessing*.
  Dans une QCM, G = 1/k.

La formule proposée est : $P(1 \mid C ; A, D, G) = G + \dfrac{1 - G}{1 + C^{-1{,}7\ A\ (C-D)}}$ [1]

Ici, l'axe des abscisses va, en principe, de $-\infty$ à $+\infty$. En pratique, on considère essentiellement l'intervalle allant de − 2,5 à + 2,5.

b) Des exemples

Voici les courbes caractéristiques (selon le modèle de Lord et Novick) de trois questions a, b et c.
La question a présente une pente proche de 45° (A=1).
La question c présente une pente beaucoup plus faible (discrimination faible).
La question b, QCM à 5 solutions proposées, a une valeur G = 0,20.



[1] La valeur 1,7 est une constante qui amène l'ogive logistique approximativement en correspondance avec l'ogive normale (B. Wright and M. Stone, 1979, p. 21).

c) Critique

Le modèle de Lord et Novick présente une conséquence paradoxale : un individu de compétence X a plus de chances de réussir la question *b* que la question *a*, alors que s'il accroit sa compétence jusqu'à Y, la question *b* risque d'être moins bien réussie que la question *a*. On ne voit pas bien ce qui pourrait justifier une telle inversion des facilités des questions !

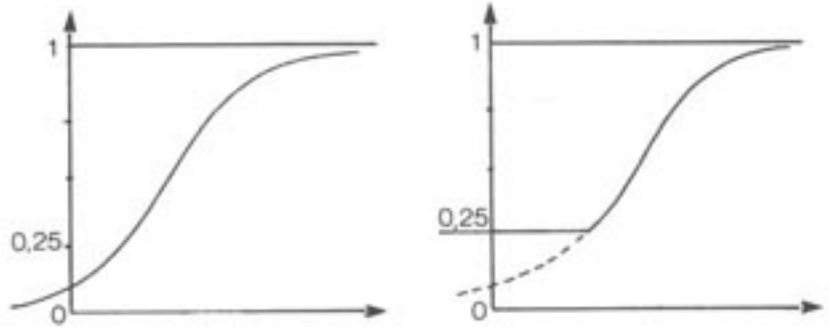
Le modèle adopté par Choppin, plus simple, n'en est pas moins rigoureux. En effet, on n'introduit dans les analyses que des questions dont l'indice de discrimination est excellent (pente A satisfaisante). Les autres questions sont impitoyablement éliminées.

8. Conclusions

En théorie, le modèle de Lord et Novick, à trois paramètres par question, est plus séduisant que le modèle de base à un seul paramètre. Celui-ci pourrait facilement être adapté au cas des QCM (où la valeur G = $\frac{1}{k}$ doit être considérée comme la « probabilité plancher »).

La procédure la plus simple consiste à tronquer la portion de la courbe inférieure à G.

Si, par exemple, la courbe caractéristique d'une QCM à 4 solutions proposées est la suivante (à gauche), on la « transformera » comme dans le schéma de droite.



Les objections (voir point 7C ci-avant) que l'on peut faire au cœfficient de pente nous paraissent sérieuses : il vaut mieux ne retenir que des questions ayant un indice de discrimination élevé (voir chapitre suivant), plutôt que de traiter par des modèles très sophistiqués des résultats à des questions fort critiquables. De plus, le modèle de Lord et Novick requiert la mise en œuvre de programmes d'ordinateur complexes et coûteux à l'utilisation. Pour toutes ces raisons, nous préférons actuellement le modèle à un seul paramètre par question.

Des recherches prometteuses sont actuellement menées à Vienne (Fischer, 1977) visant à calculer D (indice de difficulté de la question), non plus selon des principes statistiques, mais sur la base d'une analyse logique de la matière et du contenu de la question.

L'approche de la difficulté d'une question a pris un tournant décisif avec le modèle de Rasch. Cette entreprise est, à notre avis, appelée à connaître d'autres développements considérables.

CHAPITRE III

Les indices du pouvoir discriminatif d'une question

A. Les deux composantes du pouvoir discriminatif d'une question.
B. La correlation biseriale de point.
C. Préparation de l'analyse d'une épreuve.
D. L'analyse d'une épreuve à l'aide d'indices de discrimination.
E. Méthodes rapides de calcul d'indices de discrimination.
F. Comparaison des divers indices.
G. Les données absentes.

Introduction

Une question réussie par la moitié des élèves et ratée par l'autre moitié permet de discriminer les élèves. Mais quelle est la valeur de cette coupure ? Les élèves qui ont réussi sont-ils bien les plus compétents de la classe dans la matière ? La question sépare-t-elle les élèves comme le ferait un outil plus valide, par exemple un test qui est constitué de nombreuses questions, qui a déjà fait ses preuves, et qui est jugé pertinent par les experts en la matière ? Bref, la question discrimine-t-elle à *bon escient* ?

Le chapitre suivant abordera ce problème. On étudiera non seulement les indices mathématiques appropriés à la mesure du pouvoir discriminatif d'une question, mais aussi l'utilisation de ces indices pour juger de la qualité intrinsèque d'une question, de sa rédaction, voir de la pertinence de son insertion dans un test donné.

A. Les deux aspects du pouvoir discriminatif d'une question

Le pouvoir discriminatif d'une question revêt deux aspects, l'un quantitatif, l'autre qualitatif :

- le *pouvoir séparateur* de la question (aspect quantitatif) ;
- la *concordance* entre la séparation (les élèves qui ont réussi d'une part et ceux qui ont échoué de l'autre) effectuée par la question et d'autres coupures ou classements effectués par d'autres moyens réputés valides (aspect qualitatif).

Ces deux aspects peuvent être combinés ; si leurs effets se conjuguent, la question aura alors un pouvoir discriminatif élevé.

1. Le pouvoir séparateur d'une question

a) Le principe

Une question réussie par tous les élèves (facilité = 100 %) ne crée aucune distinction entre les individus. Elle ne permet pas de les répartir en deux groupes : son pouvoir séparateur est nul. Il en va de même pour une question à laquelle tous échouent (facilité = 0 %).

Quand un seul élève sur 100 réussit une question, les 99 autres constituent une masse *indistincte* : la question sépare peu. Il en va de même pour une question de facilité égale à 99 %. D'après le même raisonnement, c'est lorsque sa facilité est de 50 % qu'une question a un pouvoir séparateur maximal.

b) Notation

Rappelons que nous désignons par pc la fréquence, à une question donnée, de réponses *correctes* et par pe la fréquence, à une question donnée, des *échecs* (erreurs + omissions). Dans la littérature, la lettre p est souvent utilisée pour désigner pc et q pour désigner pe[1]. Nous revenons ici à cette notation.

Lorsque la facilité d'une question est égale à 50 %, p et q valent chacun 0,5. Le produit entre p et q est alors maximal : pq = 0,25.

Rappelons que p est en fait la *moyenne* des résultats (0 et 1) des étudiants à la question considérée. Dans notre exemple, cinquante scores de 1 et cinquante scores de 0 donnent une moyenne totale de 0,5.

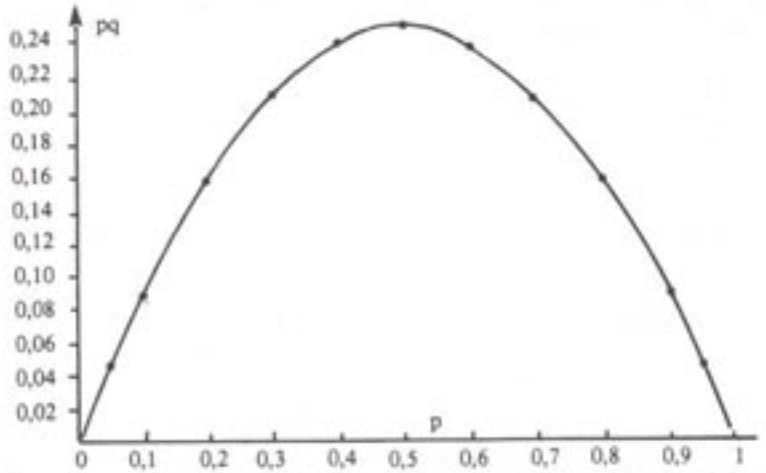
c) Calcul de la variance et de l'écart type

La variance des 100 scores vaut pq[1] et l'écart-type $\sqrt{pq}$.

Représentation graphique

Le graphique ci-dessous montre que le pouvoir séparateur d'une question est fonction de p :

Si p = 0,50 alors	pq = 0,50 × 0,50 = 0,25	et $\sqrt{pq}$ = 0,500
p = 0,40 alors	pq = 0,40 × 0,60 = 0,24	et $\sqrt{pq}$ = 0,489
p = 0,30 alors	pq = 0,30 × 0,70 = 0,21	et $\sqrt{pq}$ = 0,458
p = 0,20 alors	pq = 0,20 × 0,80 = 0,16	et $\sqrt{pq}$ = 0,400
p = 0,10 alors	pq = 0,10 × 0,90 = 0,09	et $\sqrt{pq}$ = 0,300
p = 0,05 alors	pq = 0,05 × 0,95 = 0,0475	et $\sqrt{pq}$ = 0,217



[1] Rappelons que la variance des scores vaut la somme des carrés des différences de chaque score (0 ou 1) à la moyenne p, cette somme étant divisée par N (le nombre de scores). Prenons par exemple, une question réussie par 3 étudiants sur 10 (N = 10).
- On observera trois fois le score 1 et sept fois le score 0.
- La moyenne p vaudra 0,3 et q vaudra 0,7.
- La différence entre chaque score et la moyenne...
  ...vaudra 0,3 (c'est-à-dire p) dans 7 cas (c'est-à-dire dans Nq cas).
- ... vaudra 0,7 (c'est-à-dire q) dans 3 cas (c'est-à-dire dans Nq cas).
- La variance vaudra $[N.q.(p)^2 + N.p.(q)^2]/N$ c'est-à-dire la moyenne des carrés des différences à la moyenne.
- Si l'on simplifie la formule par N, la variance devient $VAR = q(p)^2 + p(q)^2$.
- Si l'on met qp en évidence, VAR = (q.p)(p.q). Or p + q = 1 ; donc VAR = qp.
- Puisque la variance est la carré de l'écart type (σ), ce dernier a pour formule $\sigma = \sqrt{pq}$. La variance vaut pq.

d) Utilisation

Pendant bon nombre d'années, on n'a accordé d'importance qu'au seul pouvoir séparateur. Dans les épreuves de psychologie différentielle, on éliminait les questions dont la facilité (p) était inférieure à 20 ou 25 % et supérieure à 75 ou 80 %.

Or, le but principal d'une épreuve pédagogique n'est pas de mettre en évidence des différences entre étudiants, mais bien de savoir qui a appris et qui ne l'a pas fait ou l'a fait insuffisamment. Dès lors, même si une question est réussie à 0 % ou à 100 % il peut être intéressant de l'inclure dans une épreuve.

C'est la pertinence de l'objectif, jugée à la lumière d'une analyse théorique, qui doit être le critère de présence de la question dans un test pédagogique.

2. La concordance entre les scores (0 ou 1) à une question et les scores obtenus dans d'autres épreuves

Un autre problème est de savoir si la coupure (faible ou forte, peu importe) concorde avec d'autres coupures ou classements réputés plus valides qu'une question unique.

a) Le principe

Imaginons qu'un professeur de mathématique construise une épreuve scolaire de cinq questions sur la matière du programme et la présente à huit élèves auxquels il donne cours. Or, un mois auparavant, ces élèves avaient subi un test standardisé de mathématique comportant cent questions. Ce test standardisé est réputé très *valide* : il couvre toute la portion de matière considérée, nécessite quatre heures de passation, est corrigé de façon objective et le classement des étudiants qui en résulte peut être considéré comme *un critère de référence fiable.* Chacun des huit élèves a donc obtenu à ce test standardisé un score qui servira de critère extérieur et qui est repris dans un classement.

Imaginons que le professeur dispose pour chaque élève, d'une part, du score total au test standardisé (le critère extérieur) et, d'autre part, du score à chacune des cinq questions de son épreuve scolaire.

Il peut vérifier si la coupure faite par une question (répartition des élèves en deux groupes : ceux qui ont réussi et ceux qui ont échoué) concorde avec le classement obtenu au test standardisé.

Ainsi, si l'une des cinq questions est réussie par un seul étudiant et si celui-ci obtient précisément le meilleur résultat au test standardisé (critère extérieur) alors la concordance entre les deux scores est parfaite.

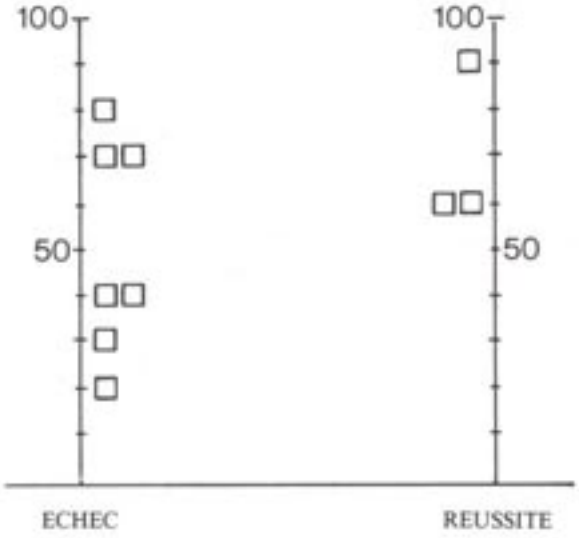
Si ceux qui réussissent la question sont des élèves plus compétents (au critère extérieur) que ceux qui y échouent, alors la question est concordante avec ce test pris comme critère (extérieur) de mesure de la compétence.

b) La représentation graphique

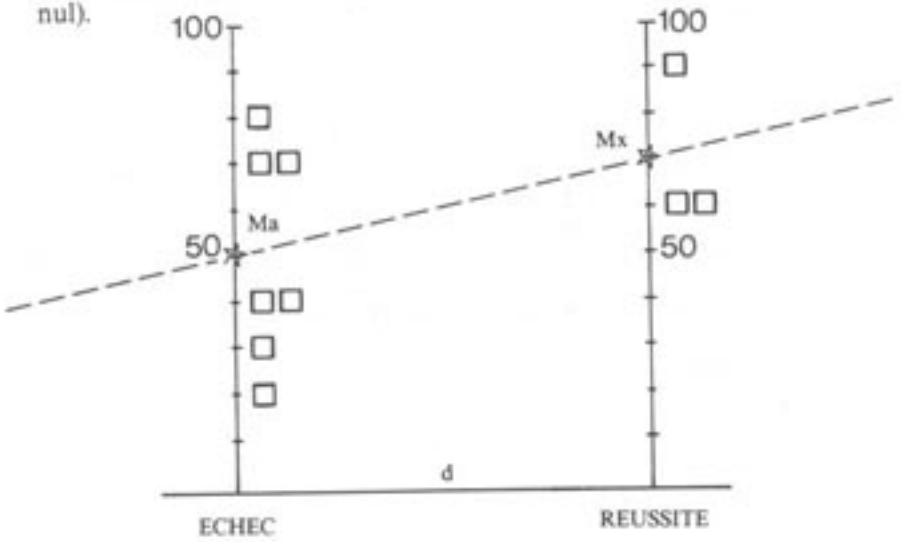
Cette relation de concordance peut s'exprimer de façon graphique, en appliquant la procédure suivante pour *chaque question.*

1. A partir d'une même ligne de base horizontale, élever deux échelles perpendiculaires de même hauteur. Cette hauteur correspond au score maximum qu'il est possible d'obtenir au *test critère* (ici 100 points). Ces deux échelles sont graduées.
2. Considérer la question x posée par le professeur. Sur l'échelle de gauche, on porte les scores au test critère des étudiants *qui n'ont pas réussi* à la question x considérée et sur l'échelle de droite les scores (au test critère) des étudiants *qui ont réussi* à cette question x.

Voici à titre d'exemple, un tel graphique dressé pour une question réussie par trois étudiants et ratée par sept autres ($p = 0{,}3$ et $q = 0{,}7$). Pour les trois premiers étudiants, les scores au test standardisé étaient 60,60 et 90 ; ces trois valeurs sont portées sur l'échelle de droite. Pour les sept étudiants qui ont échoué à la question, les scores au test étaient 20, 30, 40, 40, 70, 70 et 80 ; ces sept valeurs sont portées sur l'échelle de gauche.



3. On calcule alors la moyenne Mx des scores de l'échelle de droite et la moyenne Ma des scores de l'échelle de gauche.
4. On porte ces deux moyennes sur le graphique (Mx sur l'échelle de droite et Ma sur l'échelle de gauche).
5. On joint par une droite ces deux points (Ma et Mx). Si la droite obtenue monte de la gauche vers la droite, alors il y a concordance entre la question et le critère ; on dit aussi que la liaison est *positive*, car le cœfficient de pente de cette droite est positif. Si la droite obtenue descend de la gauche vers la droite, alors il y a discordance entre la question et le critère ; la liaison est *négative*, le cœfficient de pente de cette droite étant négatif. Si la droite est horizontale, il n'y a ni concordance ni discordance, on dit aussi que la liaison est nulle (cœfficient de pente nul).



Dans notre exemple, la moyenne (Mx) des trois étudiants qui ont réussi est 70 alors que la moyenne (Ma) des sept autres est 50. Ces deux valeurs sont représentées par des croix sur la graphique. Si l'on joint Ma à Mx, on obtient une droite *ascendante*, de pente positive (dessinée en pointillés). Il y a donc concordance entre les mesures obtenues par la question et les mesures obtenues par les test critère.

> Le signe de la pente de la droite, ou de la liaison, est donné par le signe de l'expression mathématique Mx – Ma.

c) Interprétation du graphique

Dans l'exemple précédent, la différence est positive, ce qui signifie que ceux qui ont réussi la question ont *une moyenne* (au test critère) plus élevée que les autres. Ce résultat est conforme aux attentes : il est logique que la question de l'épreuve scolaire construite par le professeur donne des résultats concordant avec ceux du test standardisé.

Dans cet exemple, les scores sont très dispersés : l'écart type (σ) est grand. Quand l'écart type est grand, il est possible d'observer entre Mx et Ma une différence plus importante que si les résultats avaient été moins dispersés. C'est pourquoi la valeur numérique de l'indice de discrimination tient compte de l'écart type, *en exprimant la différence Mx – Ma par rapport à* σ, dans l'expression $\frac{Mx - Ma}{\sigma}$.

Par ailleurs, la pente de la droite (en traits pointillés) dépend non seulement de l'ampleur de la différence Mx – Ma, mais aussi de la distance d qui sépare les deux verticales. L'indice de corrélation bisériale de point prend aussi cet aspect en considération, comme nous le verrons ci-après, lors de l'exposé du principe de calcul de la distance d.

B. La correlation bisériale de point

Ce cœfficient de corrélation est la combinaison en une seule formule des aspects quantitatif (pouvoir séparateur = $\sqrt{pq}$) et qualitatif ($\frac{Mx - Ma}{\sigma}$) du caractère discriminatif d'une question.

Le cœfficient de corrélation bisériale de point est arbitrairement symbolisé, comme la plupart des cœfficients de corrélation, par la lettre r d'abord.

Le terme *bisérial* exprime le fait que l'une des deux variables (la réponse à la question) répartit les étudiants en deux séries.

Le point symbolise le fait que la corrélation bisériale s'applique aux dichotomies vraies ou situations *dichotomiques*, par exemple la réussite ou l'échec à une question, le choix (ou non) d'une solution dans une QCM, le sexe (m ou f) d'un étudiant, etc.

On a recours à un autre cœfficient de corrélation pour les dichotomies fausses ou situations dichotomisées, par exemple deux catégories d'âge (au-delà de 25 ans et en deçà de 25 ans). Dans cet exemple, on a rendu dichotomique une variable continue (l'âge). Cette situation concerne aussi des variables telles que la taille, le poids, le quotient intellectuel, les résultats scolaires, etc. On utilise alors le rbis ou cœfficient de corrélation bisériale. On note souvent la corrélation bisériale *de point* par la série de symboles « rpbis » ou « r point bis », ou « r.bis » mais pas r bis, qui est le symbole de la *corrélation bisériale* (voir plus loin).

1. Les formules

a) Deux formules équivalentes

Les deux formules de base sont les suivantes :

$$rpbis = \frac{Mx\text{-}Ma}{\sigma}\sqrt{pq} \qquad \text{(Formule 1)}$$

$$rpbis = \frac{Mx\text{-}Mt}{\sigma}\sqrt{\frac{p}{q}} \qquad \text{(Formule 2)}$$

avec Mt = moyenne de tous les scores au test critère.

La deuxième formule est plus rapide à calculer. En effet, Mt (dans la formule 2) a la même valeur, quelle que soit la question envisagée alors que Ma (dans la formule 1) doit être recalculée pour chaque question. Cette différence est sans importance quand les calculs sont faits par ordinateur, ce qui est de plus en plus souvent le cas.

La formule du rpbis est un cas particulier de la corrélation de Bravais-Pearson.

b) La corrélation bisériale

$$rbis = \frac{Mx\text{-}Ma}{\sigma} \cdot \frac{p \cdot q}{y} = \frac{Mx\text{-}Mt}{\sigma} \cdot \frac{p}{y}$$

où y = l'ordonnée de la courbe normale obtenue par consultation de la table de la loi normale à la valeur p (les y sont aussi appelées « ordonnées réduites »).

Donc $rbis = rpbis \frac{\sqrt{p.q}}{y}$ et $rpbis = rbis \frac{y}{\sqrt{p.q}}$

c) Tables de valeurs pour les calculs

On trouvera ci-après une table reprenant des valeurs intervenant dans le calcul des cœfficients rpbis et rbis.

Table de valeurs intervenant dans le calcul de rpbis et de rbis

p	q	pq	$\sqrt{pq}$	y	y²	$\sqrt{\frac{p}{q}}$	$\sqrt{\frac{q}{p}}$	p	q	pq	$\sqrt{pq}$	y²	$\frac{pq}{y^2}$	$\sqrt{\frac{p}{q}}$	$\sqrt{\frac{q}{p}}$
,01	0,99	0,0099	0,0994	0,0266	,3721	,100	9,94	0,26	0,74	0,1924	0,4386	,3244	0,5930	,591	1,68
,02	0,98	0,0196	0,1400	0,0484	,4069	,141	7,00	0,27	0,73	0,1971	0,4439	,3306	0,5961	,608	1,64
,03	0,97	0,0291	0,1705	0,0680	,4279	,173	5,68	0,28	0,72	0,2016	0,4489	,3366	0,5989	,624	1,60
,04	0,96	0,0384	0,1959	0,0862	,4434	,200	4,89	0,29	0,71	0,2059	0,4537	,3423	0,6015	,640	1,56
,05	0,95	0,0475	0,2179	0,1031	,4607	,223	4,36	0,30	0,70	0,2100	0,4582	,3477	0,6039	,655	1,52
,06	0,94	0,0564	0,2374	0,1191	,4735	,244	3,95	0,31	0,69	0,2139	0,4624	,3528	0,6062	,670	1,49
,07	0,93	0,0651	0,2551	0,1343	,4847	,282	3,64	0,32	0,68	0,2176	0,4666	,3576	0,6085	,685	1,45
,08	0,92	0,0736	0,2712	0,1487	,4951	,300	3,39	0,33	0,67	0,2211	0,4702	,3621	0,6106	,700	1,42
,09	0,91	0,0819	0,2861	0,1624	,5043	,316	3,17	0,34	0,66	0,2244	0,4737	,3664	0,6124	,721	1,39
,10	0,90	0,0900	0,3000	0,1755	,5128	,331	3,00	0,35	0,65	0,2275	0,4768	,3706	0,6142	,734	1,36
,11	0,89	0,0979	0,3128	0,1880	,5207	,346	2,84	0,36	0,64	0,2304	0,4800	,3741	0,6158	,748	1,33
,12	0,88	0,1056	0,3249	0,2000	,5280	,379	2,71	0,37	0,63	0,2331	0,4829	,3776	0,6173	,764	1,30
,13	0,87	0,1131	0,3363	0,2115	,5347	,387	2,58	0,38	0,62	0,2356	0,4853	,3808	0,6186	,781	1,27
,14	0,86	0,1204	0,3469	0,2226	,5408	,400	2,47	0,39	0,61	0,2359	0,4877	,3837	0,6200	,800	1,24
,15	0,85	0,1275	0,3570	0,2332	,5467	,424	2,38	0,40	0,60	0,2400	0,4898	,3863	0,6212	,818	1,22
,16	0,84	0,1344	0,3666	0,2433	,5524	,435	2,29	0,41	0,59	0,2419	0,4918	,3887	0,6223	,830	1,20
,17	0,83	0,1411	0,3756	0,2531	,5534	,447	2,20	0,42	0,58	0,2436	0,4935	,3909	0,6231	,848	1,17
,18	0,82	0,1476	0,3841	0,2624	,5625	,469	2,13	0,43	0,57	0,2451	0,4950	,3928	0,6238	,866	1,15
,19	0,81	0,1539	0,3923	0,2714	,5670	,479	2,06	0,44	0,56	0,2464	0,4963	,3944	0,6243	,888	1,12
,20	0,80	0,1600	0,4000	0,2800	,5714	,500	2,00	0,45	0,55	0,2475	0,4974	,3958	0,6253	,905	1,10
,21	0,79	0,1659	0,4073	0,2882	,5756	,519	1,93	0,46	0,54	0,2484	0,4983	,3969	0,6258	,921	1,08
,22	0,78	0,1716	0,4142	0,2961	,5795	,529	1,88	0,47	0,53	0,2491	0,4990	,3978	0,6261	,943	1,06
,23	0,77	0,1771	0,4208	0,3036	,5833	,547	1,83	0,48	0,52	0,2495	0,4995	,3986	0,6260	,959	1,03
,24	0,76	0,1824	0,4270	0,3108	,5866	,565	1,78	0,49	0,51	0,2499	0,4998	,3988	0,6266	,979	1,01
,25	0,75	0,1875	0,4330	0,3178	,5899	,574	1,73	0,50	0,50	0,2500	0,5000	,3989	0,6267	1,	1,

n.b. $\sqrt{\frac{p}{q}} = \sqrt{\frac{pq}{pq}}$

2. Valeur numérique et interprétation graphique du rpbis

Imaginons l'exemple suivant où dix étudiants ont subi un test critère dont les scores peuvent varier de 0 à 20. Voici leurs résultats (1 = correct, 0 = échec) à une question x d'un autre test.

Etudiants	1	2	3	4	5	6	7	8	9	10
Score à la question x	1	1	1	0	0	0	0	0	0	0
Score au test critère	12	14	19	4	6	9	9	12	14	16

p = 0,3 et q = 0,7 Mt = 115/10 = 11,5

L'écart type (σ) des scores au test critère vaut $\sqrt{\frac{184,5}{10}} = 4,295$

Mx = (12+14+19)/3 = 45/3 = 15
Ma = (4+6+9+9+12+14+16)/7 = 70/7 = 10

a) Valeur numérique

Par la formule 1, rpbis

$$= \frac{15-10}{4,295}\sqrt{0,3.0,7} = \frac{5}{4,295}\sqrt{0,21} = 1,1641.0,4582 = 0,533$$

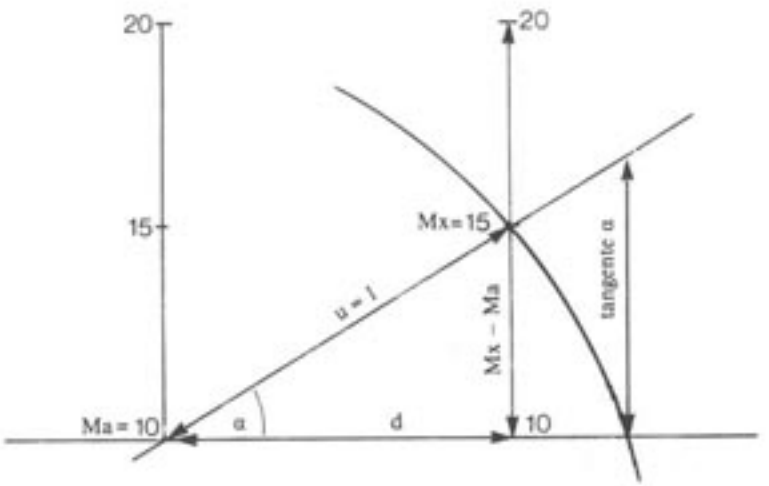
Par la formule 2, rpbis

$$= \frac{15-11,5}{4,295}\sqrt{\frac{0,3}{0,7}} = \frac{3,5}{4,295}\sqrt{0,4285} = 0,8149.0,6546 = 0,533$$

b) Représentation graphique

La valeur du rpbis peut aussi être visualisée sur le graphique constitué des deux échelles verticales parallèles. Ci-dessous, elles ont été dessinées à une distance précise (d) l'une de l'autre. Nous expliquerons ci-après comment cette valeur d a été calculée.

La distance (verticale) Mx – Ma apparaît comme le sinus de l'angle α formé par l'oblique et l'horizontale.



La distance (horizontale) d apparaît comme le cosinus de l'angle α

Rappelons que $\frac{\text{sinus } \alpha}{\text{cosinus } \alpha}$ = tangente α

On a donc $\frac{Mx - Ma}{d} = \frac{\text{sinus } \alpha}{\text{cosinus } \alpha}$ = tangente α

Il suffit que l'on fixe d égal à $\frac{\alpha}{\sqrt{pq}}$ pour avoir

$\frac{Mx - Ma}{\sigma}\sqrt{p.q}$ (donc *rpbis*) = tangente α.

Dans notre exemple, $d = \sigma/\sqrt{p.q} = 4,295/0,4582 = 9,373$.

On peut vérifier que $rpbis = (Mx - Ma)/d = \frac{5}{9,373} = 0,533$

C. *Préparation de l'analyse d'une épreuve*

1. Quatre remarques préalables

Avant d'envisager l'analyse d'une épreuve faite de QCM, quatre remarques s'imposent.

a) Le total du test comme critère

Dans la plupart des cas, le professeur ne dispose pas des scores des étudiants à un test-*critère* valide. On utilise alors la même formule, mais on prend comme *critère* le *score total* au test dont la question fait partie. La concordance entre la question et le total de l'épreuve devient dès lors un indice de cohérence interne entre la question et le test. La somme (pondérée) des rpbis peut d'ailleurs servir à calculer un cœfficient de consistance interne ou d'*homogénéité* du test dans son ensemble.

b) La corrélation automatique

Chaque question d'un test contribue au score total, pour une part égale à 1/NQ quand toutes les NQ questions sont d'égale facilité. On dit qu'il y a « recouvrement » (ou inclusion) entre le score de la question et le score total du test.

Dans le cas, purement théorique, où les questions d'un test n'auraient aucune corrélation entre elles (ex : $\varphi_{ij} = 0$), la *corrélation* entre une question et le total du test vaut automatiquement $1/\sqrt{NQ}$. Ainsi, pour NQ = 9, rpbis vaut 1/3 ou 0,33. C'est ce que l'on appelle la *valeur-repère* du rpbis d'une question. En fait, on peut calculer une telle valeur-repère pour chaque solution de chaque question (voir J-L. Hardy, 1982).

Par conséquent, tout rpbis de réponse correcte qui n'atteint pas *au moins* cette valeur-repère ne peut être considéré comme l'indice d'une « bonne » corrélation de la question avec l'épreuve entière. Toutes les corrélations question-total sont donc « gonflées » par cette liaison fondamentale.

Partant du même raisonnement, Henrysson (1963) propose de corriger (vers le bas) la corrélation « surfaite » à l'aide de la formule suivante (voir Guilford et Fruchter, 19, p. 466).

$$C\ rpbis = \left(\sqrt{\frac{n}{n-1}}\right)\left(\frac{rpbis_i \quad \sigma - \sqrt{p \quad q}}{\sqrt{\sigma^2 \quad - \quad p \quad q}}\right)$$

avec crpbis = r point bis de la question corrigé pour « recouvrement » *(overlap)*.
σ = écart type des résultats au total de l'épreuve.
On trouvera dans Hardy (1981 et 1983) une critique approfondie de cette formule.

Voici ses recommandations :
- ne pas calculer une valeur « corrigée » (comme le font Henrysson et d'autres) pour le rpbis ;
- calculer la valeur repère qui servira à interpréter le rpbis ;
- la valeur repère, pour la solution correcte, peut être calculée par la formule $1/\sqrt{NQ}$ quand on ne dispose pas de formule plus précise ;
- l'idéal est de calculer la valeur repère pour chaque solution de chaque question.

c) Les rpbis des réponses correctes et des distracteurs

Dans une QCM à k solutions, on peut calculer un indice de discrimination pour chacun des k+1 comportements possibles :

(omissions, choix de la solution 1, 2... k).

Le rpbis permet de comparer les étudiants qui ont adopté un de ces comportements (disons x) aux *autres* étudiants qui ont adopté les *autres* comportements (que nous désignerons par la lettre a).

> Dans un test bien construit, le rpbis de la solution correcte est normalement positif et les rpbis des solutions incorrectes (distracteurs) sont normalement négatifs.

Une QCM dont le rpbis de *la solution correcte* est *positif* est une question qui participe à l'homogénéité de l'épreuve, car elle mesure la même chose que les autres questions, et ce d'autant plus que cette corrélation (bisériale de point) est élevée.

Les distracteurs, eux, présentent normalement une corrélation (rpbis) *négative*.

L'omission présente généralement une corrélation comprise entre celles des distracteurs et celle de la réponse correcte. Qu'elle soit positive ou négative, la corrélation (rpbis) de l'omission est souvent proche de 0.

d) Explications de la valeur du rpbis

Un rpbis positif pour une solution signifie que ce sont les meilleurs étudiants (au total du test) qui ont choisi cette solution. Il serait donc paradoxal que le rpbis d'un distracteur soit positif. Si c'est le cas, il importe de se demander ce qui a pu produire un tel résultat.

Le résultat peut être du à

- un défaut fondamental :
  a. *Manque de validité :* la question ne mesure pas ce que l'on croyait.
  b. *Manque d'homogénéité* entre l'épreuve et cette question, qui mesure autre chose.
- un défaut formel :
  rédaction : mauvaise formulation, distracteurs non pertinents, ambigus, etc.
- un accident :
  erreur de codage : on a indiqué comme réponse correcte une solution incorrecte.

C'est le *sens clinique* du professeur qui lui permettra de faire le diagnostic correct et de procéder au remaniement adéquat de la question.

Une valeur paradoxale de l'indice de discrimination n'est qu'un *signe* de dysfonctionnement ; il n'en donne pas l'explication.

e) Le programme d'ordinateur

Une analyse des rpbis de toutes les solutions proposées à toutes les questions est aujourd'hui inconcevable sans l'aide de l'ordinateur. Le programme le plus rigoureux dont nous disposions est celui de Hardy (1981) intitulé ANIT (pour Analyse d'items) et écrit en langage FORTRAN. Les indices qui sont présentés ci-après ont été calculés par ce programme.

2. La préparation des données

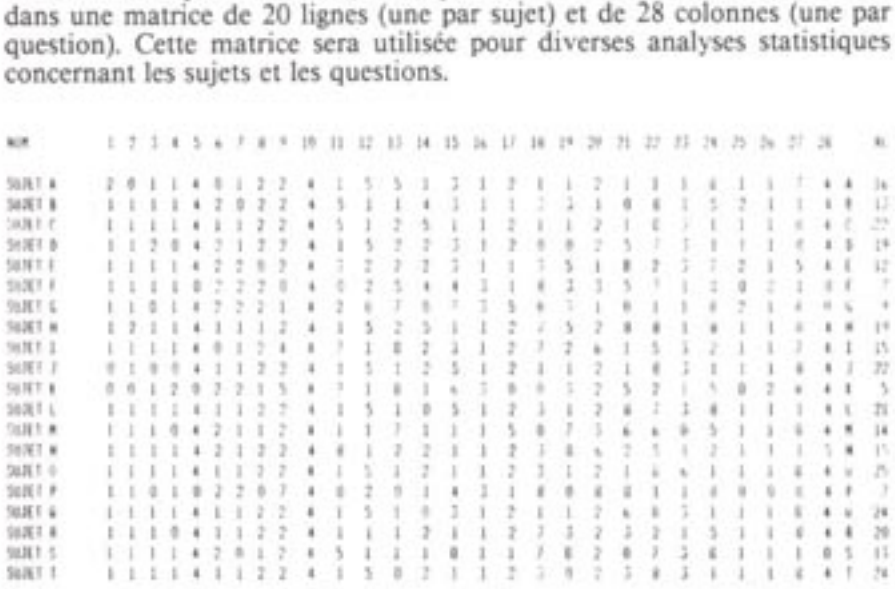
A titre d'exemple, voici les réponses d'étudiants aux 28 questions (dont 22 QCM et 6 questions ouvertes) d'un *module d'auto-évaluation* sur les petites annonces[1].

Afin de faciliter des calculs et les raisonnements qui en découlent, nous avons limité les données aux réponses de 20 étudiants seulement. Les indices mathématiques calculés sur un nombre si peu élevé d'étudiants ne peuvent être qu'indicatifs. En fait, on ne calcule pas les rpbis des questions quand on dispose de si peu de réponses. Le minimum, en pratique, est de 30 individus.

[1] La description détaillée de cette épreuve apparaîtra en même temps que la présentation des résultats, c.-à-d. les réponses des élèves et le calcul des rpbis.

a) La matrice de base

Dans l'exemple ci-dessous, les réponses des étudiants sont présentées dans une matrice de 20 lignes (une par sujet) et de 28 colonnes (une par question). Cette matrice sera utilisée pour diverses analyses statistiques concernant les sujets et les questions.



b) Les traitements préalables au calcul des rpbis

Ce tableau brut est l'objet de traitements de base préalables aux analyses proprement dites. Ces traitements et analyses peuvent être effectués à l'aide d'une machine à calculer ou d'un ordinateur.

Opération 1 : Calculer le score total (T) étudiant (ou nombre de réponses correctes).

Opération 2 : Calculer la moyenne (MT) de ces 20 scores. Ici, MT = 16,1 (sur 28, soit une réussite moyenne de 57,5 %).

Opération 3 : Calculer l'écart type de ces 20 scores à l'aide de la formule

$$\sigma = \sqrt{\frac{\Sigma\,(X\text{-}X)^2}{N}} = 5{,}97$$

Opération 4 : Calculer la popularité de chaque solution (p) pour les diverses QCM. On exprime souvent cette popularité en pourcentage (ex. : 84 %), plus agréable à lire, mais les calculs sont effectués sur les fréquences (ex. : 0,84).

D. *L'analyse d'une épreuve à l'aide d'indices de discrimination*

L'épreuve de 28 questions mentionnée ci-dessus est constituée de trois sous-épreuves :
- une sous-épreuve de connaissance (quatre questions)
- une sous-épreuve de compréhension (quatre questions)
- une sous-épreuve d'application (vingt questions).

Cette épreuve a été soumise à 20 étudiants. Les résultats seront présentés par sous-épreuve. Pour chaque solution proposée, le pourcentage de réponses (p x 100) est indiqué, le rpbis (positif ou négatif selon les cas) apparait juste sous ce pourcentage. Les valeurs-repères des rpbis apparaissent entre parenthèses ; elles sont calculées, pour chaque solution par le programme ANIT de Hardy.

1. Les quatre questions de connaissance

Énoncés des questions

Omission	Réponses correctes	Réponses incorrectes	Dans le vocabulaire technique des sciences économiques, on désigne :
10 % - 0,15	**85 %*** **0,12** **(0,60)**	5 % - 0,00	*Question 1 :* Une proposition de vente par le terme : .....
10 % - 0,31	**85 %*** **0,19** **(0,06)**	5 % 0,11	*Question 2 :* Une proposition d'achat par le terme : .....
15 % - 0,24	**80 %*** **0,16** **(0,07)**	5 % 0,11	*Question 3 :* Un objet acheté ou vendu (ou un bien matériel) par le terme : ....
20 % 0,22	**75 %*** **0,01** **(0,07)**	5 % - 0,43	*Question 4 :* Un bien non matériel acheté ou vendu par le terme : ....

Les réponses correctes (marquées d'un astérisque) ont des rpbis faibles (0,12 ; 0,19 ; 0,16 ; 0,01), proches de leur valeur-repère, ce qui indique que ces quatre questions *ne mesurent* sans doute *pas le même chose* que le reste de l'épreuve. Ceci correspond à la volonté du constructeur de dissocier les quatre questions de connaissances des quatre questions de compréhension et des vingt questions d'application.

Deux conclusions doivent être tirées :

1. L'hypothèse du constructeur est confirmée : les questions (de mémoire) mesurent autre chose que l'ensemble du test ; on peut être fort en connaissance de « mémoire » et obtenir néanmoins un score total faible, et vice versa.
2. Il serait intéressant de calculer un score séparé pour ces quatre questions et de ne pas les faire intervenir dans le calcul du score total au test.

2. Les quatre questions de compréhension

Ces quatre questions avaient pour amorce commune le texte encadré ci-dessous :

> Voici maintenant une annonce trouvée dans un quotidien belge :
> Offrons 32.000 F net par mois à
> COUPLE DOMESTIQUE
> 30-45 ans, référ. 1er ordre
> exig., sans enf., très propre,
> travailleur, homme permis
> conduire. Tél. 02-600.

Dans les tableaux qui suivent, la cellule correspondant à la réponse est encadrée en traits épais.

Omissions	Solution 1	Solution 2	Sol. 3	Sol. 4	
15 % - 0,69	0 %	0 %	0 %	**85 %*** **0,69** **(0,06)**	*Question 5 :* Qui a payé pour faire imprimer cette annonce ? 1. Un couple de domestiques 2. Le journal. 3. L'homme (qui a son permis de conduire) dont on parle dans l'annonce. 4. Quelqu'un d'autre.
10 % - 0,03	**40 %*** **0,82** **(0,09)**	50 % - 0,79			*Question 6 :* Cette annonce offre un salaire. 1. Vrai 2. Faux
10 % - 0,17	**65 %*** **0,82** **(0,09)**	25 % - 0,78			*Question 7 :* Cette annonce demande des services. 1. Vrai 2. Faux
10 % - 0,37	20 % - 0,28	**70 %*** **0,49** **(0,08)**			*Question 8 :* Cette annonce demande un emploi. 1. Vrai 2. Faux

On constate que :

1. Quand, dans une question, deux solutions seulement ont été utilisées comme réponse, les rpbis de ces deux solutions sont symétriques et opposés. Ainsi, dans la question 5, le rpbis de l'omission (-0,69) est l'opposé du rpbis de la solution correcte (0,69).
2. Tous ces rpbis sont positifs et élevés pour les solutions correctes (0,69 ; 0,82 ; 0,82 ; 0,49). Ces quatre questions mesurent donc bien *la même chose que le reste de l'épreuve.* La nuance que le constructeur avait cru introduire entre « compréhension » et « application » n'apparaît donc pas.

3. Les vingt questions d'application

Toutes les questions d'application (sauf les questions 25 et 26) ont les mêmes solutions proposées. Il s'agit en effet de dire, pour chaque annonce, quel est *son objet fondamental,* et les huit solutions proposées sont :

1. Une offre de service.
2. Une offre de produit.
3. Une demande de service.
4. Une demande de produit.
5. 1 et vraisemblablement 2.
6. 3 et vraisemblablement 4.
7. Il est impossible de savoir de quels biens économiques (services ou produits) il s'agit.
8. Il ne s'agit pas de biens économiques.

a) Procédure d'interprétation des indices

L'interprétation des résultats, dépend essentiellement des réponses que l'évaluateur apporte à *trois questions techniques* qu'il faut se poser systématiquement. Nous ne les avons présentées que pour la question 9.

Question 9

1°– Le rpbis de la solution correcte est-il supérieur à la valeur repère ?
Réponse : Oui (ce sont donc de bons élèves qui ont choisi la solution correcte).

2°– Les rpbis des distracteurs choisis sont-ils tous négatifs ?
Réponse : Oui.

3°– Les rpbis des distracteurs choisis sont-ils tous inférieurs à leur valeur-repère ?
Réponse : Oui (nous n'avons d'ailleurs pas indiqué ces valeurs repères).

Conclusion : La question 9 se comporte donc conformément à toutes les attentes.

Pour les questions suivantes, nous fournissons directement les réponses sur un tableau. Une analyse détaillée n'est fournie que pour certaines questions.

Questions	Omiss.	1 Une offre de service.	2 Une offre de produit.	3 Une demande de service.	4 Une demande de produit.	5 1 et vraisemblablement 2.	6 3 et vraisemblablement 4.	7 Impossible de savoir si c'est service ou produit.	8 Il ne s'agit pas de biens économiques.
Question 9 : [illegible]	5 % -0,35	5 % -0,27	75 % 0,73 (0,08)		5 % -0,04	5 % -0,43		5 % -0,35	
Question 10 : Suis acheteur 100.000 F pce [illegible]					100 %				
Question 11 : ESPAGNOL INTENSIF Costa del Sol - Petits groupes.	10 % -0,51 (-0,03)	50 % 0,72 (0,08)	5 % -0,27	5 % -0,16 (-0,02)		15 % -0,01 (-0,04)		10 % -0,34 (-0,03)	5 % -0,04 (-0,02)
Question 12 : PEINTURE [illegible]		40 % -0,20	15 % -0,52			40 % 0,73 (0,09)			5 % -0,27
Question 13 : SI VOUS AVEZ DES PROBLEMES DE STOCKAGE nous vous proposons : ENTREPOTS MODERNES Libres ou sous douane à partir de 50 m².	10 % -0,03	35 % 0,44 (0,08)	25 % 0,13 (0,03)			10 % -0,26		10 % -0,26	10 % -0,34

Questions	rpbis DE LA SOLUTION CORRECTE EST-IL SUPERIEUR A LA VALEUR REPERE ? Hypothèse explicative.	LES rpbis DES DISTRACTEURS SONT-ILS TOUS NEGATIFS ? Hypothèse explicative.	LES rpbis DES DISTRACTEURS SONT-ILS TOUS INFERIEURS A LA VALEUR REPERE ? Hypothèse explicative.	CONCLUSION
Question 9	OUI	OUI	OUI	PARFAIT
Question 10	—	—	—	—
Question 11	OUI	OUI	OUI, sauf 5. Certains étudiants auraient pensé « matériel scolaire »	INTERVIEWER LES ETUDIANTS
Question 12	OUI	OUI	OUI	PARFAIT
Question 13	OUI	OUI SAUF 2 (pour 5 et.) Cf analyse détaillée	—	SUPPRIMER OU MODIFIER L'ANNONCE

Question	0	1 Une offre de service	2 Une offre de produit	3 Une demande de service	4 Une demande de produit	5 1 et vraisemblablement 2	6 3 et vraisemblablement 4	7 Impossible de savoir si c'est services ou produits	8 Il ne s'agit pas de biens économiques	rpbis sol. corr. val. repère ? Hyp. explicative	RPBIS DISTRACT. TOUS NEGATIFS Hypothèse explicative	RPBIS DISTRACT. VALEUR REPERE ? Hypothèse explicative	CONCLUSIONS
Question 14 : CITROEN GS	15 % -0,13 (-0,03)	25 % -0,49	**40 % 0,40 (0,06)**		10 % -0,34	10 % -0,25 (-0,02)				OUI	NON pour SOL.5 le distract. 5 est ambigu		SUPPRIMER LE DISTRACTEUR 5 OU TEST.
Question 15 : GENS DE MAISON Demandes de travail à la journée — Femme BELGE	5 % -0,12	**35 % 0,46 (0,06)**		30 % 0,04 (-0,04)	10 % -0,51	10 % 0,30 (-0,02)	5 % -0,43	5 % -0,27		OUI	NON pour SOL.3 et SURTOUT SOL.5 (0,30) Distract. 5 ambigu.		SUPPRIMER LE DISTRACTEUR 5 OU TEST.
Question 16 : engage : bibliothéc. documental.		**80 % 0,76 (0,07)**		20 % -0,76						OUI	OUI	OUI	PARFAIT
Question 17 : ARDENNES MONT-XHOFFRAIX (Malmédy) MAGNIFIQUE CHALET	5 % -0,43	25 % -0,55	**60 % 0,63 (0,09)**			10 % -0,26				OUI	OUI	OUI	PARFAIT
Question 18 : Mr [redacted] Jean, [redacted], rue [redacted] à [redacted] déclare que sa femme, [redacted] a quitté le domicile conjugal depuis le 29 novembre. Il ne reconnaît plus à partir de ce jour, les dettes que sa femme aurait fait ou pourrait faire.	15 % -0,24	20 % 0,41 (-0,01)		30 % 0,25 (-0,02)				20 % 0,05 (-0,01)	**15 % -0,59 (0,06)**	NON	NON — Voir analyse détaillée.		SUPPRIMER LA QUESTION DE CE TEST

Question	Omissions	1 Une offre de service	2 Une offre de produit	3 Une demande de service	4 Une demande de produit	5 1 et vraisemblablement 2	6 3 et vraisemblablement 4	7 Impossible de savoir si c'est services ou produits	8 Il ne s'agit pas de biens économiques	rpbis sol. corr. val. repère ? Hyp. explicative	RPBIS DISTRACT. TOUS NEGATIFS Hypothèse explicative	RPBIS DISTRACT. VALEUR REPERE ? Hypothèse explicative	CONCLUSIONS
Question 19 : OUVRIERS Demandes	15 % 0,04 (-0,02)	**30 % 0,61 (0,08)**	5 % -0,04	25 % -0,31		10 % -0,03 (-0,02)		5 % -0,08	10 % -0,12	OUI	OUI SAUF POUR L'OMISSION (cf analyse détaillée)	OUI	OK
Question 20 : EXISTENCE ASSUREE		15 % -0,34	**60 % 0,63 (0,09)**	10 % -0,31			10 % -0,06		5 % -0,35	OUI	OUI	OUI	PARFAIT
Question 21 :	15 % -0,21 (0,00)	25 % 0,38 (0,00)	5 % -0,04 (0,00)	10 % 0,33 (0,00)		15 % -0,41 (0,00)	10 % 0,16 (0,00)	** **	20 % -0,11 (0,00)	JAMAIS CHOISIE piège : cas non classique	NON pour SOLUTIONS 1 et 6 piège : cas non classique	—	SUPPRIMER LA QUESTION DE CE TEST
Question 22 : MARIAGES		15 % -0,38	15 % -0,27			10 % -0,06	5 % -0,06	20 % -0,09	**35 % 0,64 (0,08)**	OUI	OUI	OUI	PARFAIT
Question 23 : dactylo télexiste	5 % -0,08	45 % -0,37		**45 % 0,46 (0,09)**			5 % 0,34 (-0,02)			OUI	NON POUR DISTRACT.6 1 seul individu (hasard ?)	—	ATTENDRE CONFIRMATION AVEC PLUS DE SUJETS

	Omissions	1. Une offre de service.	2. Une offre de produit.	3. Une demande de service.	4. Une demande de produit.	5. 1 et vraisemblablement 2.	6. 3 et vraisemblablement 4.	7. Impossible de savoir si c'est services ou produits.	8. Il ne s'agit pas de biens économiques	rpbis sol. corr. val. repère ? Hyp. explicative	RPBIS DISTRACT. TOUS NEGATIFS Hypothèse explicative	RPBIS DISTRACT. VALEUR REPERE ? Hypothèse explicative	CONCLUSIONS
Question 24 : Un mariage est un jour de fête. Pour que cette joie persiste, voir, choix et rire clairvoyant : Notre sérieux et notre discrétion vous aideront. l'Agence Marrim.		**30 % 0,72 (0,08)**	10 % -0,06			20 % -0,26		5 % -0,16	35 % -0,36	OUI	OUI	OUI	PARFAIT
Question 25 : Complétez l'annonce ci-dessous pour en faire une OFFRE de SERVICES. [illegible]	15 % -0,69	**70 % 0,80 (0,06)**	15 % -0,34							OUI	OUI	OUI	PARFAIT
Question 26 : Complétez la même annonce pour en faire une DEMANDE DE SERVICE.	5 % -0,35	**85 % 0,89 (0,06)**	10 % -0,56							OUI	OUI	OUI	PARFAIT
Question 27 : Alcooliques ou Toxicomanes gagnez votre abstinence en rejoignant tous les mardis en réunion à partir de 19 h. 30 pour le relèvement de votre personnalité. Le Groupe des Alcooliques Anonymes, chez les Soeurs		-0,22				5 % -0,16		10 % -0,03 (-0,03)	**60 % 0,29 (0,09)**	OUI	OUI	EGAL A LA VALEUR REPERE POUR SOLUTION 7 bénéfice du doute (2 étudiants seulement)	OK
Question 28 : PAIE CHER TABLEAUX ANTIQUITES	15 % -0,45				**80 % 0,43 (0,07)**	5 % -0,04				OUI	OUI	OUI	PARFAIT

b) *Analyse détaillée de certaines questions*

Question 13

Le distracteur 2 n'a pas un rpbis négatif. Interviewés, ces cinq « bons » étudiants ont déclaré qu'ils pensaient que l'annonce proposait la *vente* d'entrepôts à construire, et non la location.

Interrogés à leur tour, les professeurs de sciences économiques ont refusé cette interprétation à cause de la mention « libre ou sous douane ». Il apparaît donc que la question est bien libellée, mais qu'elle fait appel à des connaissances techniques étrangères à l'objectif de l'épreuve.

Il faut dès lors :

- soit supprimer ou modifier la solution 2, ce qui est impossible, car elle est commune aux questions 9 à 28 ;
- soit supprimer ou modifier l'annonce elle-même pour la rendre moins ambigüe. C'est cette dernière solution que nous adopterions.

Question 18

La configuration des rpbis est très inattendue ! Interviewés, les étudiants ont d'abord contesté la solution correcte : pour certains, dans la mesure où l'annonce parle de biens économiques (les dettes et leur paiement), la solution 8 ne peut être correcte ; les partisans de la solution 1 ont considéré que l'annonceur signalait la fin *d'un service* (ils la classaient donc à offre) ; les partisans de la solution 3 ont considéré que l'annonceur *demandait* qu'on lui rende le service de ne plus l'importuner (ils le classaient donc à demande).

Les professeurs de sciences économiques que nous avons consultés ont suggéré de supprimer cette annonce parce que les solutions, communes aux autres questions, ne lui étaient pas bien adaptées. Autrement dit, cette annonce pourrait faire l'objet d'une QCM ayant ses propres solutions proposées.

Question 19

Seule l'omission a un rpbis différent des attentes, mais cette différence légère est représentative des questions difficiles auxquelles un certain nombre d'élèves « moyens » préfèrent ne pas répondre.

Question 21

Le rpbis de la solution correcte n'a pas été calculé car elle n'a jamais été choisie. Des rpbis ont cependant été calculées pour les solutions choisies et pour l'omission, ce qui différencie cette question de la question 10, pour laquelle aucun rpbis n'a évidemment pu être calculé. Après discussion avec des enseignants, il est appru que cette question 21 devrait être supprimée, car elle constitue un « piège » et non un cas classique.

Question 23

Le distracteur 6 (3 et vraisemblablement 4) est très nettement confusionnel car son rpbis est positif. Néanmoins, il n'est calculé que sur *un seul individu.* D'autres observations seraient nécessaires pour confirmer cette mise en cause.

4. Bilan de l'analyse des rpbis

Au terme de ces analyses, on constate que :

- Pour une question (n° 10), les calculs n'ont pas été faits.
- Pour dix-neuf questions, tous les rpbis répondent aux attentes.
- Pour trois questions (2, 3, 23), un seul rpbis est inattendu et dû à un seul étudiant. Le hasard pouvant, dès lors, avoir joué grandement, le bénéfice du doute est laissé à ces questions.
- Pour quatre questions (11, 13, 14 et 15) une solution déterminée (la solution 5) a été mise en cause et une explication satisfaisante a pu être trouvée.
- Pour trois questions (4, 18, 21), on obtient des rpbis inattendus pour les divers distracteurs, et des explications ont pu être avancées.

Si l'on envisageait de supprimer la solution 5 (commune à vingt questions) parce qu'elle pose problème aux questions 11, 14 et 15, il faudrait, hélas ! supprimer la question 12, pourtant bien réussie. On le voit, le remaniement de l'épreuve reste un travail assez lourd ; mais les rpbis ont largement contribué à indiquer dans quel sens il devait avoir lieu.

5. Calcul des rpbis à partir de sous-ensembles du test

Nous avons procédé au calcul des rpbis des questions après suppression dans le test des questions 1, 2, 3, 4, 10, 18 et 21. La moyenne des étudiants à ce sous-test de 21 questions devient 11,7 (soit une réussite moyenne de 55,7 %) et l'écart type 6,25.

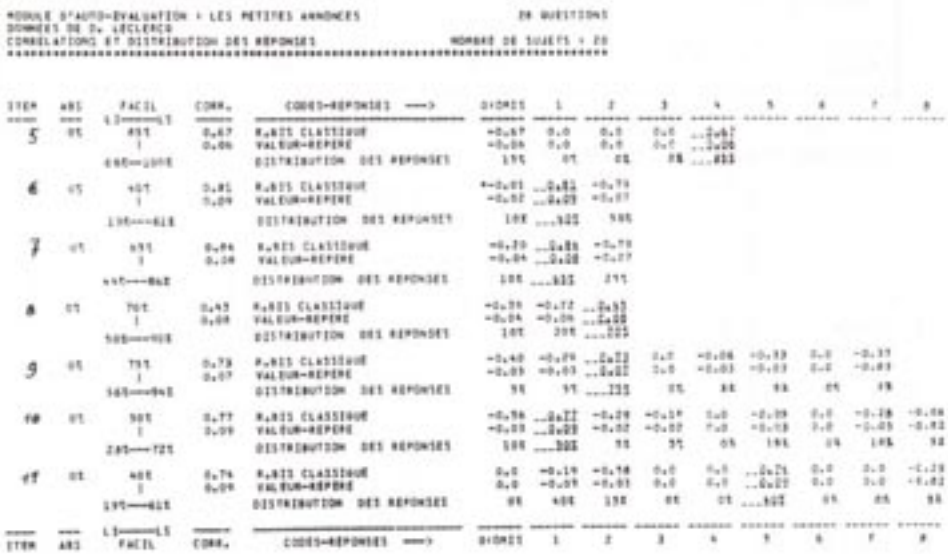
Il en résulte une beaucoup plus grande homogénéité de rpbis et une valeur moyenne plus élevée (0,46 pour les 28 questions et 0,63 pour les 21 questions).

La fidélité du test de 28 questions vaut 0,886 (estimation par la formule KR20) ou 0,885 (estimation par la formule basée sur les rpbis). Le test de 21 questions a une fidélité de 0,924 (par KR20) ou 0,934 (par les rpbis). Si on ajoutait *7 questions présentant des caractéristiques statistiques équivalentes* aux 21 questions gardées, on obtiendrait une épreuve de longueur comparable à celle du départ (28 questions) mais de meilleure qualité.

A la suite de cet allongement du test par un cœfficient n (valant 28/21 = 1,333), la fidélité résultante serait de 0,941 ; ce résultat indique que l'épreuve ainsi modifiée deviendrait un meilleur instrument de classement des individus.

6. Extrait du programme ANIT

Le tableau ci-après est fourni par le programme ANIT de J.L. Hardy.



En plus de l'indice de facilité (p) figurent sa limite inférieure (LI) et sa limite supérieure (LS). Ces limites se situent à – 1,96 erreur standard de p (pour LI) et à + 1,96 erreur standard de p (pour LS). Rappelons que l'erreur standard de p se calcule pour la racine carrée du produit de pq divisé par N (le nombre de sujets).

E. *Méthodes rapides de calcul d'indices de discrimination*

1. La méthode des quatre groupes

G. De Landsheere (1976, p. 86 à 89) décrit une méthode où l'on constitue quatre groupes d'élèves en fonction d'un critère (par exemple le résultat scolaire global). Le quart des élèves les plus faibles sont classés dans le groupe D, le quart suivant dans le groupe C, le quart encore supérieur dans le groupe B et le quart des meilleurs dans le groupe A. On calcule alors pour chacun des quatre groupes (de D, les plus faibles à A, les plus forts) le taux de réussite observé pour cette question. Si les taux se présentent en progression de D vers A, on considère que la question discrimine positivement.

Dans notre exemple précédent (28 questions présentées à 20 étudiants), ces quatre groupes sont constitués comme suit :

Groupe D : K(5), F(7), P(7), G(9), et E(12)
Groupe C : B(13), S(13), N(15), I(15) et M(14)
Groupe B : A(16), D(19), H(19), R(20) et L(21)
Groupe A : J(22), C(22), Q(27), T(24) et O(25)

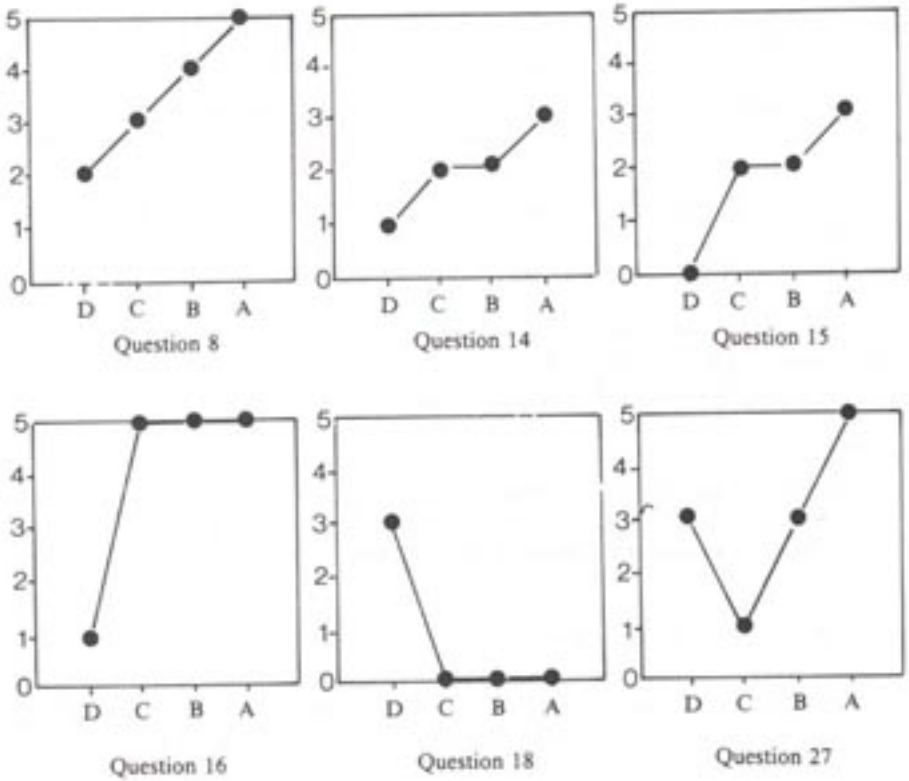
Les nombres de succès par question se répartissent de la manière suivante dans chacun des quatre groupes :

	QUESTIONS													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Groupe D	4	4	3	4	2	0	0	2	1	5	0	0	0	1
Groupe C	5	5	5	4	5	0	3	3	4	5	1	0	2	2
Groupe B	4	3	4	3	5	3	5	4	5	5	5	4	2	2
Groupe A	4	5	4	5	5	5	5	5	5	5	4	4	3	3
	17	17	16	15	17	8	13	14	15	20	10	8	7	8

	QUESTIONS													
	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Groupe D	0	1	0	3	0	1	0	0	1	0	0	2	3	3
Groupe C	2	5	2	0	0	1	0	1	2	0	4	5	1	3
Groupe B	2	5	5	0	2	5	0	1	2	1	5	5	3	5
Groupe A	3	5	5	0	4	5	0	5	4	5	5	5	5	5
	7	16	12	3	6	12	0	7	9	6	14	17	12	16

A titre d'illustration, les nombres de succès sont représentés graphiquement pour six questions. Pour les questions 8, 14, 15 et 16, on observe que les courbes correspondent aux attentes.

Par contre, les questions 18 et 27 présentent des graphiques paradoxaux.



Si de telles analyses étaient pratiquées à partir d'un nombre suffisant de sujets (30 dans chaque groupe), elles permettraient de repérer comme le font les rpbis, les questions ambiguës ou mesurant une autre variable que l'ensemble des autres questions.

2. La méthode des deux groupes extrêmes (l'indice D net)

a) Le principe

Une autre procédure consiste à ne constituer que deux groupes extrêmes d'étudiants : les plus faibles (groupe inférieur ou G.I.) et les plus forts (groupe supérieur ou G.S.). Selon les variantes, on constitue chacun de ces deux groupes extrêmes avec 27 % ou 30 % ou encore 33 % d'élèves. D'Agostino et Cureton (1975) recommandent 21 %.

Nous utiliserons la valeur 25 % dans notre exemple de 28 questions présentées à 20 étudiants (5 élèves sur 20 dans le groupe inférieur et 5 élèves sur 20 dans le groupe supérieur).

Le taux de réussite à la question doit être supérieur dans le groupe des forts.

Johnson (1951) a défini un indice D de discrimination appelé U-L *(upper-lower)*.

La formule est $D = \frac{U - L}{N}$ ou $D = \frac{GS - GI}{N}$

où

U = nombre de réponses correctes dans le groupe supérieur (ou GS).
L = nombre de réponses correctes dans le groupe inférieur (ou GI).
N = nombre d'étudiants dans le groupe supérieur (ou dans le groupe inférieur : c'est le même nombre).

Cet indice varie de -1 à +1.

Cet indice, encore appelé le D net (Discriminatif net, de l'anglais *Net Discrimination Index*) repose sur les travaux de Kelley (1939) et Flanagan (1939). Findley (1956) a aussi contribué à ce courant de recherche.

b) Un exemple

Dans notre exemple, le groupe supérieur (GS) est constitué par les cinq étudiants du groupe A défini ci-devant, et le groupe inférieur (GI) est constitué par les cinq étudiants du groupe D défini dans la section 1 ci-devant.

Dans le tableau ci-après, on trouvera, pour chaque question, les valeurs de GS et de GI, préparatoires au calcul du D net, ainsi que les valeurs des rpbis (pour rappel).

Question	1	2	3	4	5	6	7	8	9	10	11	12	13	14
GS (RC dans le groupe A)	4	5	4	4	5	5	5	5	5	5	4	3	3	3
GI (RC dans le groupe D)	4	4	3	4	2	0	0	2	1	5	0	0	0	1
D Net	0,00	0,20	0,20	0,00	0,60	1,00	1,00	0,60	0,80	0,00	0,80	0,80	0,60	0,40
rpbis	0,12	0,19	0,16	0,01	0,69	0,82	0,82	0,49	0,73	0,00	0,72	0,70	0,44	0,40

Question	15	16	17	18	19	20	21	22	23	24	25	26	27	28
GS Groupe A	3	5	5	0	4	5	0	5	4	5	5	5	5	5
GI Groupe D	0	1	0	3	0	1	0	0	1	0	0	2	3	3
D Net	0,60	0,80	1,00	-0,60	0,80	0,80	0,00	1,00	0,60	1,00	1,00	0,60	0,40	0,40
rpbis	0,46	0,76	0,83	-0,59	0,61	0,63	0,00	0,64	0,46	0,72	0,80	0,69	0,29	0,43

Notons que l'indice *D net* calculé sur peu de sujets constitue une mesure grossière de la discrimination. Quand le nombre de sujets augmente, la valeur de cet indice approche celle du rpbis. Celui-ci et le D net s'interprètent de la même façon.

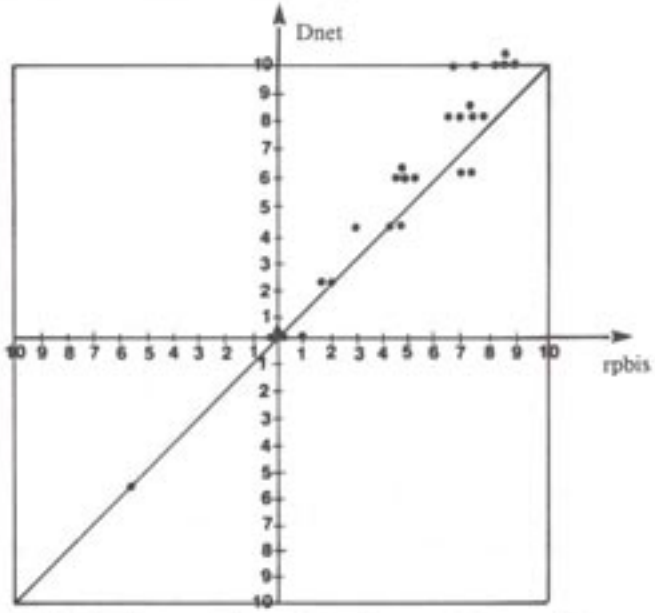
Les valeurs *positives* du D net indiquent une répartition des réponses conforme aux attentes. Les valeurs *négatives* du *D net* (ici la question 18) indiquent un résultat paradoxal : dans ce cas, les meilleurs réussissent moins bien que les faibles.

La valeur du *D net* indique la *puissance de discrimination* de la question. Ainsi, la question 11 (*D net* = 0,80) est plus discriminative que la question 27 (*D net* = 0,40).

F. Comparaison des divers indices et discussion

1) D net et rpbis

Le nuage de points ci-après montre la forte relation entre les D net et les rpbis (la corrélation entre eux vaut 0,96).



On le voit, le *D net* permet de repérer les questions qui font problème et de les classer selon leur pouvoir discriminatif. Nuttal et Skurnik (1969) fournissent des tables facilitant encore le calcul du D net.

Cet indice, facile à calculer, présente cependant des faiblesses : cinquante pour cent des données (le groupe des étudiants « moyens » sont ignorés – donc perdus – par l'analyse. On est alors amené à faire le calcul sur de petits nombres : le hasard peut jouer et on aboutit à des résultats grossiers.

C'est pourquoi on préférera des indices qui utilisent toutes les données disponibles : les cœfficients de corrélation bisériale (rbis) et bisériale de point (rpbis), même s'ils sont assez longs à calculer lorsqu'on ne possède pas d'ordinateur.

Hales (1972) a montré que le *D net* concorde avec l'indice de corrélation bisériale, même dans les cas de distributions non normales, pourvu que les calculs soient effectués sur un nombre suffisant de réponses.

2) rpbis et rbis

Nous comparerons les rpbis et les rbis sur une question précise, la question 11 du module d'auto-évaluation sur les petits annonces.

Rappelons que la question 11 avait un *D net* égal à 0,8. Les valeurs obtenues par les rpbis et rbis sont plus précises. En effet on dispose d'un indice de discrimination pour chaque solution, sauf pour celles qui n'ont pas été choisies (ici les solutions 4 et 6). Dans ces cas, il est impossible de calculer les rpbis et rbis.

La configuration des indices de discrimination pour la *question 11* est la suivante :

Solution	Omis	1*	2	3	4	5	6	7	8
rpbis	- 0,51	0,72	- 0,27	- 0,16	-	- 0,01	-	- 0,34	- 0,04
rbis	- 0,86	0,93	- 0,57	- 0,32	-	- 0,04	-	- 0,62	- 0,16

A chaque solution, le rbis est de même signe et de même ordre de grandeur (quoique plus élevé en valeur absolue que le rpbis). Il peut arriver (mais rarement) que les rbis dépassent la valeur 1 ou –1, ce qui n'est jamais le cas du rpbis.

Dans le cas présent, les rpbis seront préférés aux rbis puisque la variable est dichotomique et non « dichotomisée ».

3) Autres indices de discrimination (pour mémoire)

Glass (1966) donne la formule de calcul d'un autre indice de discrimination : le cœfficient de corrélation bisérial de rangs. Cet indice est indépendant de la difficulté de la question. Il n'est cependant pas recommandé pour les groupes nombreux ($N > 50$).

Certains (Ivens, 1971 ; Hoffmann, 1975) ont proposé de combiner l'indice de facilité de la question (p ou pc) et l'indice de discrimination de la réponse correcte.

D'autres ont tenté de combiner les différents indices de discrimination des distracteurs et de la réponse correcte. Ainsi, Das Gupta (1960) a proposé un *cœfficient multisérial de point* (généralisation du rpbis) où chaque solution est simplement considérée comme différente des autres (échelle nominale).

Jaspers (1965) a proposé un *cœfficient polysérial* (autre généralisation du rbis), qui exige que les distracteurs soient ordonnés quant à leur degré d'inexactitude (ou selon la gravité de l'erreur). Wood (1977, p. 247) remarque que ce type de mise en ordre peut rarement être réalisée de manière satisfaisante.

Préférant l'indice multisérial de point, Wood considère que le cœfficient polysérial n'apporte, en pratique, aucune information que ne porte déjà la configuration des k+1 rpbis d'une QCM.

Il y a même une perte d'information puisque les cœfficient polysérial et multisérial sont des indices globaux n'apportant pas de précision sur chacune des solutions en particulier. Ainsi, Wilmot (1975b) observe que des questions présentant le même indice multisérial de point pourraient présenter des configurations de rpbis différentes.

G. *Les données absentes*

Il peut arriver que certains étudiants n'aient pas le temps de répondre aux dernières questions d'un test. Il peut arriver aussi que le test soit peu commode à appliquer parce que le matériel est difficile à transporter, que l'on en ait perdu une partie ou qu'il tombe facilement en panne, etc.), et qu'alors certains étudiants n'aient *pas eu l'occasion de répondre aux questions correspondantes.* Faut-il éliminer de l'analyse les résultats d'un étudiant à qui, par exemple, il ne manque que quatre questions sur 100, et ainsi perdre 96 informations ? Dans les deux cas, la matrice des réponses révèle l'absence des données.

Il est possible de tenir compte de ces résultats, mais à plusieurs conditions :

1. L'absence de réponse par manque de temps ou pour une autre raison ne doit pas être confondue avec l'omission délibérée.
2. On ne peut pas calculer les rpbis par rapport aux scores totaux, mais on doit le faire à partir des pourcentages de réussite calculés sur le nombre de questions réellement abordées par le sujet (nombre variable pour chaque individu).
3. Les taux de réussite des questions (p) doivent eux aussi être calculés à partir du nombre (variable pour chaque question) d'étudiants qui les ont abordées.
4. Pour accepter les résultats d'un étudiant dans l'analyse, celui-ci doit avoir répondu à un nombre minimal de questions, que l'expérimentateur définit.
5. On n'accepte de procéder à l'analyse d'une question que si un minimum d'étudiants (à déterminer par l'expérimentateur) y ont répondu.

H. Conclusions

Grâce aux indices de discrimination, l'expérimentateur est mieux armé pour savoir si une question donnée mesure la même chose que les autres questions du test ou si elle est exempte de défauts techniques. Ces indices sont tout particulièrement précieux pour les QCM car, par ordinateur, on peut en calculer un pour chaque solution proposée sur laquelle des choix d'élèves se sont portés. Les valeurs obtenues doivent être confrontées à des valeurs repères.

Des recherches ont été menées sur les méthodes de sélection des distracteurs les plus appropriés.

R. Owens, G. Hanna et F. Coppedge (1970) ont étudié trois tests dont les distracteurs avaient été choisis par trois méthodes différentes :
- jugement d'experts
- popularité des solutions
- indices de discrimination.

Ils n'ont pas trouvé de différence significative dans la corrélation des scores à ces tests avec les scores à un test critère constitué de réponses ouvertes.

En fait, la sélection des distracteurs relève d'une *démarche clinique*, fondée, bien sûr, sur des indices mathématiques. La décision de garder, de modifier ou de supprimer un distracteur dépassera toujours les seules considérations numériques.

CHAPITRE IV

La signification des scores

A. Les quatre types d'échelles

La notation se fait sur deux types d'échelles et la cotation sur deux autres, soit quatre types de catégories au total, dont il importe de bien comprendre les propriétés mathématiques afin de ne pas se laisser abuser par des *interprétations erronées* de notes (appréciations qui peuvent ne pas être numériques) ou de scores (valeurs numériques).

Nous utiliserons le terme NOTATION pour désigner une appréciation dans le sens plus général, qu'elle soit numérique ou non. Nous réserverons au terme COTATION le sens plus restreint de note se traduisant par un nombre (dès lors appelée *cote*).

1. Les catégories d'une échelle nominale

Ces catégories diffèrent entre elles par des aspects uniquement qualitatifs.

Exemple non scolaire : Rangez le spectacle que vous venez de voir dans une des catégories suivantes :
1. comédie musicale
2. opéra
3. opérette
4. théâtre
5. café concert
6. conférence

etc.

Exemple scolaire : L'étudiant a défendu son examen (de maturité, par exemple) :
1. en français
2. en néerlandais
3. en anglais
4. en espagnol
5. en italien
6. en allemand

etc.

Les *codes* (de 1 à 6) n'ont qu'une valeur pratique, mais n'ont absolument *pas* l'ensemble des *propriétés mathématiques* des *nombres* entiers : le théâtre (code 4) ne vaut pas deux fois plus que l'opéra (code 2).

Les seuls signes mathématiques que l'on puisse utiliser avec ces codes sont les signes d'égalité (=), d'inégalité ($\neq$) ou d'égalité approximative ($\cong$). Dans l'exemple non scolaire, on peut écrire

$1 \neq 2 \neq 4 \neq 5$,
$2 = 2$, $4 = 4$, etc.
et $2 \cong 3$

Mais les signes de supériorité ($>$) et d'infériorité ($<$) sont exclus. On ne peut *pas* écrire ici $4 < 5$ ou $2 < 3$, bien que l'on utilise les chiffres (digits) qui, d'habitude servent à constituer les nombres. Ici, il s'agit bien de nombres, mais qui ne servent qu'à *énumérer*, à *désigner*, à *distinguer*, à *nommer* des catégories différentes les unes des autres et *pas du tout* à les ordonner ou les quantifier. Dans les échelles suivantes, les catégories seront non seulement différentes, mais elles auront, entre elles, des propriétés supplémentaires.

2. Les catégories d'une échelle ordinale

Ces catégories diffèrent entre elles par une relation de supériorité/infériorité, bref par une relation d'ordre.

Exemple non scolaire : Attribuez au tremblement de terre qui vient d'être observé une note sur l'échelle (ordinale) de Richter : 1, 2, 3, 4, 5, 6, 7.

Exemple scolaire : Quelle est la note que mérite le travail de cet étudiant ?
1. Insatisfaisant.
2. Faible.
3. Moyen.
4. Bon.
5. Très bon.
6. Excellent.
7. Parfait.

Dans ces deux exemples, on peut écrire $1 < 2 < 3 < 4 < 5 < 6 < 7$ mais on ne peut toujours pas utiliser les opérateurs d'addition et de soustraction. Ainsi, on ne peut pas écrire que $2 + 3 = 5$, car un tremblement de terre de force 2 et un tremblement de terre de force 3 ne font pas, à eux deux combinés, un tremblement de terre de force 5. De même, un travail « faible » (2) et un travail « moyen » (3) ne font pas à eux deux un travail « très bon » (5).

De même, ces notes ne peuvent être ni divisées, ni multipliées, car un travail « excellent » (code 6) n'est pas « deux fois meilleur qu'un travail moyen » (code 3).

Il eut été préférable d'utiliser des lettres pour *coder* (ex. : A, B, C, D, E, F) car on est moins tenté de se livrer à des opérations mathématiques abusives (bien que certains n'hésitent pas à traduire des lettres en chiffres pour s'adonner à de tels calculs).

Nous avons voulu garder les nombres pour insister sur leur utilisation en tant que *codes* avec des propriétés différentes selon les échelles où ils sont utilisés.

3. Les catégories d'une échelle métrique d'intervalles égaux

Ces catégories possèdent une propriété que n'avaient pas les précédentes : des *intervalles égaux* entre elles. On ne connaissait pas la longueur de l'intervalle entre les catégories 3 et 4 (de l'échelle de Richter ou de l'échelle de notation scolaire), c'est-à-dire qu'on ne pouvait pas dire s'il était égal à l'intervalle entre les catégories 4 et 5.

Désormais, dans les échelles métriques, cet intervalle est le même entre tous les échelons.

Exemple non scolaire : Mesure de la température en degrés centigrades.
$-3, -2, -1, 0, +1, +2, +3, +4, +5, +6, +7$, etc.
Un même intervalle connu (ici un degré centigrade) sépare ces différentes catégories.

Exemple scolaire : Notes possibles de l'étudiant à l'épreuve d'histoire :
0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.
Un même intervalle connu (ici un point) sépare ces différentes catégories.

Dans ces deux exemples, on peut écrire non seulement
$1 \neq 3$ (comme dans les catégories nominales).
$1 < 3$ (comme dans les catégories ordinales).

mais $4 + 2 = 6$
et $8 - 2 = 6$

(Les deux points qui séparent 4 de 6 sont la même distance qu'entre 6 et 8) ; il y a le même nombre de degrés centigrades entre 6 et 8 qu'entre 2 et 4.

Cela a dès lors un sens de parler de températures *moyennes* sur une certaine période, et de cotes (ou de scores) *moyens*.

Et c'est bien ce type d'information que l'on communiquera aux élèves et aux parents, même à travers une échelle métrique, car certaines valeurs-repères ont une signification bien connue.

Ainsi, 10/10 ou 20/20 ou 100 % signifient d'habitude « performance parfaite, n'ayant pour égale que celle du professeur dans les mêmes circonstances ». On se prend à rêver à la façon dont on noterait une performance d'étudiant meilleure que celle du professeur. On devrait lui donner plus de 20/20, ce qui ne devrait poser aucun problème (ces notes n'étant que des échelles métriques d'intervalles, il n'y a ni maximum ni minimum absolu). Et pourtant, on sait que les professeurs n'écriront jamais dans un bulletin 22/20, parce que, pour les parents, 20/20 signifie en outre (*abusivement*) « performance maximale, indépassable ».

De même, 0/20 signifie « performance nulle », équivalente à quelqu'un qui ignore la matière (qui s'abstiendrait de répondre à toutes les questions). Est-il possible d'imaginer performance pire ? Evidemment, et cela s'observe tous les jours : les personnes qui ne s'abstiennent pas, mais répondent par de monumentales bourdes et qui, croyant savoir, ne remettent pas leurs « connaissances » en question. Et pourtant, on sait que les professeurs n'écriront jamais dans le bulletin – 4/20, parce que pour les parents 0/20 signifie en outre (*abusivement*) « performance plancher, sous laquelle il est impossible de descendre ».

Enfin 10/20 ou 50 % (ou des valeurs s'en approchant) signifient « limite de la satisfaction », c'est-à-dire valeur de la performance au-delà de laquelle on peut « passer dans la classe supérieure ». On frémit d'ailleurs à la pensée que « passent de classe » des étudiants n'ayant maîtrisé qu'un mot sur deux, un concept sur deux, un principe sur deux de tous ceux qui leur ont été enseignés.

4. Repères sociaux et repères techniques

Les enseignants ont raison de n'écrire ni score négatif ni score supérieur au maximum dans la bulletin, car le bulletin étant un organe de communication, il importe de respecter le langage du récepteur. Même si de tels scores sont tout à fait pertinents, ils le sont techniquement et non socialement ; or dans le bulletin, on doit parler social.

Ceci est aussi vrai pour les *mots* de l'évaluation que pour les *nombres*. Ainsi, si un professeur apprécie qu'un étudiant maintienne son taux de progression, il se gardera bien d'écrire « Bravo pour votre inertie ». Bien que la définition que les physiciens donnent de l'inertie (tendance à garder le mouvement acquis) s'applique parfaitement à la situation, le professeur devra tenir compte du sens courant, populaire, qui ne manquera pas d'être donné à la phrase par la plupart des parents (non physiciens).

Mais les enseignants n'ont pas à communiquer qu'avec les parents. Ils doivent pouvoir communiquer entre eux, entre spécialistes, et tout d'abord avec eux-mêmes. Et les repères techniques sont différents des repères sociaux.

Par exemple, les scores négatifs indiquent une performance inférieure à la performance d'un sujet ignorant. C'est le cas, par exemple, des enfants de 13 ans à qui l'on fait passer un prétest sur les notions de poids et de masse. Il a été démontré (D'HAINAUT, 1971) combien ils confondent les deux, mais sont sûrs d'eux-mêmes et répondent massivement ce qui est considéré comme des incongruités pour le physicien (exemples de telles incongruités : la balance sert à mesurer des poids, on met des poids « marqués » sur un des deux plateaux : le poids s'exprime en kg, etc.).

Il vaudrait beaucoup mieux que les étudiants n'aient aucune idée du concept plutôt que ces idées fausses, car le professeur devra passer autant de temps à extirper les conceptions erronées qu'à installer les correctes. Il en va de même du professeur de langue qui doit corriger une prononciation fausse déjà bien ancrée ou un professeur de natation qui doit faire modifier un mouvement déjà fortement automatisé.

Sur ces points de matière déjà (mé)connus, ces élèves ne partent pas de 0/20, mais de – 5, voire de – 10/20 parfois. Faire progresser les élèves de – 10/20 à + 10/20, c'est donc réaliser non pas la moitié du chemin vers la perfection, mais les deux tiers (20 points gagnés sur les 30 possibles).

Ce genre de considération est cruciale lorsque des professeurs – des professionnels – parlent entre eux des niveaux atteints par les élèves. Il est aussi crucial qu'ils possèdent des codes (numériques ou autres) de communication précis et non ambigus.

5. Estimation de la compétence et modèle de RASCH

Nous avons déjà eu l'occasion d'exposer le modèle de RASCH (au chapitre 1) à propos de l'estimation de la difficulté des questions. Nous avons vu qu'il consistait à attribuer à la question une valeur numérique permettant de calculer la probabilité de succès à cette question d'un étudiant dont on a par ailleurs aussi estimé la compétence sur la même échelle. On a vu que la valeur numérique n'avait d'importance que pour ce qu'on en faisait, et ici il ne s'agissait pas de communiquer.

La façon dont la compétence de l'étudiant est estimée dans le modèle de RASCH est révélatrice de la signification intrinsèque des scores 0 % et 100 %. En effet, par des procédés mathématiques qu'il serait trop long d'expliquer ici, on estime une compétence de l'étudiant sur l'échelle arbi-

Principe a : jugement et non mesure

Par l'attribution de la note (ou du grade) finale, le jury ne prétend pas *mesurer* l'étudiant. Il s'est basé sur *une série de mesures*, qu'il a ensuite pondérées, puis appréciées à la lumière de données plus qualitatives. Il s'agit donc d'une *combinaison* entre des données (et des formules) quantitatives et des données qualitatives aboutissant à l'attribution d'une *rétribution*, d'une *récompense*, plus qu'une mesure, et ce selon des critères de *justice distributive*.

On pourrait s'étonner qu'il n'existe pas de formule (même complexe) d'*agrégation* des diverses mesures. Il est symptomatique que les évaluations complexes (par exemple le salaire ou le montant du prix du transfert d'un joueur de football professionnel), ne fassent pas, eux non plus, l'objet d'un calcul rigoureux (par exemple, le prix du transfert n'est pas simplement le nombre de buts divisés par l'âge, multiplié par le nombre de matches, etc.).

Il nous paraîtrait intéressant *d'étudier la possibilité* de telles formules pour les évaluations pédagogiques, mais nous ne sommes sûrs ni de leur faisabilité, ni de leur degré de fécondité (notamment de l'absence d'effets pervers), ni de leur coût (pour se procurer les données).

Principe b : note et jury sont inséparables

Le jury ne prétend pas non plus que son échelle d'évaluation soit universel, et que le même étudiant eut obtenu le même « grade » dans une autre université. C'est d'ailleurs une source (regrettable) de confusion et d'injustice lorsque des étudiants issus d'universités différentes sont en compétition (en concours), et qu'ils sont classés essentiellement sur la base de leurs grades universitaires. Les jurys exigeants pénalisent ainsi leurs étudiants, et certaines universités ont bien compris cet enjeu.

Ce relativisme des grades culmine aux Etats-Unis où il est beaucoup plus important de savoir où (dans quelle université) vous avez obtenu votre « degré » que le grade que vous avez obtenu.

2. Le cas de l'enseignement secondaire

Nous traitons le cas de l'enseignement secondaire après celui de l'université parce que les conventions de *communication*, notamment entre parents et enseignants sont beaucoup moins claires. En effet, communiquer la décision d'un jury sous la forme d'un nombre (ex : 68 %), ouvre la porte à des ambiguités communicationnelles déjà dénoncées.

Devant une telle précision (68 % n'est pas 63 % !), les parents s'attendent à ce que ce score puisse être justifié par des sous-scores tout aussi numériques. Et c'est possible, direz-vous : il suffit que chaque professeur fournisse un score et que ces scores soient pondérés pour obtenir le score final. La transparence est donc sauve. Mais où sont alors introduits les éléments qualitatifs qui, autant dans le secondaire qu'à l'Université, doivent être pris en compte ? Dans chacune des cotes fournies par les professeurs (avant le calcul de la moyenne pondérée) ! Mais là aussi, les parents voudront obtenir du professeur la justification chiffrée de la cote. Et là aussi, c'est possible, sur base de la cote de chaque mois, elle-même fonction transparente des cotes de chaque évaluation (interrogation ou examen).

Les parents et les élèves sont attentifs aux règles d'attribution des points – et ils ont raison – et apprécient que ces règles soient transparentes. Ainsi, le retrait d'un point par exercice erroné (ou par faute d'orthographe) apparaissent comme des principes « justes » dont on ne conteste souvent que la pondération (par exemple retirer « un point » est jugé trop sévère, on préférerait un demi-point seulement).

Devant une telle mécanisation de la cotation, le professeur est parfois bien en peine de justifier d'autres appréciations (travail en classe, initiative, participation, méthode de travail, etc.) parce qu'il ne dispose pas de données quantifiées pour les justifier ni parfois de la formulation théorique (ex. : on ne dispose pas d'un modèle *complet* de la capacité à apprendre, ni de l'autonomie, ni de la convivialité, etc. et on n'en disposera probablement jamais) ni donc des instruments de mesure qui s'y rapportent.

A côté des « mesures » dures, doivent donc intervenir des appréciations « molles » ou « subjectives ». Mais quel est le statut exact de ces appréciations ?

3. Les repères de la communication

Bien que réclamant la transparence, peu de parents acceptent que le professeur donne une note mensuelle (en mathématique par exemple) dont la seule signification serait « le maximum moins le nombre d'erreurs commises ». Ils veulent – et ils ont raison – que, par sa notation, le professeur pondère l'importance des diverses erreurs (elles ne sont pas toutes de la même gravité) et leur signification (le raisonnement est-il correct ? etc.) pour la discipline.

En fait, ils veulent que la note leur dise si la compétence de leur enfant (jugée à travers des performances) est insuffisante, satisfaisante, bonne ou très bonne par rapport aux critères de l'enseignant (qui est sensé se référer à des critères nationaux ou légaux).

Et c'est bien ce type d'information que l'on communiquera aux élèves et aux parents, même à travers une échelle métrique, car certaines valeurs-repères ont une signification bien connue.

Ainsi, 10/10 ou 20/20 ou 100 % signifient d'habitude « performance parfaite, n'ayant pour égale que celle du professeur dans les mêmes circonstances ». On se prend à rêver à la façon dont on noterait une performance d'étudiant meilleure que celle du professeur. On devrait lui donner plus de 20/20, ce qui ne devrait poser aucun problème (ces notes n'étant que des échelles métriques d'intervalles, il n'y a ni maximum ni minimum absolu). Et pourtant, on sait que les professeurs n'écriront jamais dans un bulletin 22/20, parce que, pour les parents, 20/20 signifie en outre (*abusivement*) « performance maximale, indépassable ».

De même, 0/20 signifie « performance nulle », équivalente à quelqu'un qui ignore la matière (qui s'abstiendrait de répondre à toutes les questions). Est-il possible d'imaginer performance pire ? Evidemment, et cela s'observe tous les jours : les personnes qui ne s'abstiennent pas, mais répondent par de monumentales bourdes et qui, croyant savoir, ne remettent pas leurs « connaissances » en question. Et pourtant, on sait que les professeurs n'écriront jamais dans le bulletin – 4/20, parce que pour les parents 0/20 signifie en outre (*abusivement*) « performance plancher, sous laquelle il est impossible de descendre ».

Enfin 10/20 ou 50 % (ou des valeurs s'en approchant) signifient « limite de la satisfaction », c'est-à-dire valeur de la performance au-delà de laquelle on peut « passer dans la classe supérieure ». On frémit d'ailleurs à la pensée que « passent de classe » des étudiants n'ayant maîtrisé qu'un mot sur deux, un concept sur deux, un principe sur deux de tous ceux qui leur ont été enseignés.

4. Repères sociaux et repères techniques

Les enseignants ont raison de n'écrire ni score négatif ni score supérieur au maximum dans la bulletin, car le bulletin étant un organe de communication, il importe de respecter le langage du récepteur. Même si de tels scores sont tout à fait pertinents, ils le sont techniquement et non socialement ; or dans le bulletin, on doit parler social.

Ceci est aussi vrai pour les *mots* de l'évaluation que pour les *nombres*. Ainsi, si un professeur apprécie qu'un étudiant maintienne son taux de progression, il se gardera bien d'écrire « Bravo pour votre inertie ». Bien que la définition que les physiciens donnent de l'inertie (tendance à garder le mouvement acquis) s'applique parfaitement à la situation, le professeur devra tenir compte du sens courant, populaire, qui ne manquera pas d'être donné à la phrase par la plupart des parents (non physiciens).

Mais les enseignants n'ont pas à communiquer qu'avec les parents. Ils doivent pouvoir communiquer entre eux, entre spécialistes, et tout d'abord avec eux-mêmes. Et les repères techniques sont différents des repères sociaux.

Par exemple, les scores négatifs indiquent une performance inférieure à la performance d'un sujet ignorant. C'est le cas, par exemple, des enfants de 13 ans à qui l'on fait passer un prétest sur les notions de poids et de masse. Il a été démontré (D'HAINAUT, 1971) combien ils confondent les deux, mais sont sûrs d'eux-mêmes et répondent massivement ce qui est considéré comme des incongruités pour le physicien (exemples de telles incongruités : la balance sert à mesurer des poids, on met des poids « marqués » sur un des deux plateaux : le poids s'exprime en kg, etc.).

Il vaudrait beaucoup mieux que les étudiants n'aient aucune idée du concept plutôt que ces idées fausses, car le professeur devra passer autant de temps à extirper les conceptions erronées qu'à installer les correctes. Il en va de même du professeur de langue qui doit corriger une prononciation fausse déjà bien ancrée ou un professeur de natation qui doit faire modifier un mouvement déjà fortement automatisé.

Sur ces points de matière déjà (mé)connus, ces élèves ne partent pas de 0/20, mais de – 5, voire de – 10/20 parfois. Faire progresser les élèves de – 10/20 à + 10/20, c'est donc réaliser non pas la moitié du chemin vers la perfection, mais les deux tiers (20 points gagnés sur les 30 possibles).

Ce genre de considération est cruciale lorsque des professeurs – des professionnels – parlent entre eux des niveaux atteints par les élèves. Il est aussi crucial qu'ils possèdent des codes (numériques ou autres) de communication précis et non ambigus.

5. Estimation de la compétence et modèle de RASCH

Nous avons déjà eu l'occasion d'exposer le modèle de RASCH (au chapitre 1) à propos de l'estimation de la difficulté des questions. Nous avons vu qu'il consistait à attribuer à la question une valeur numérique permettant de calculer la probabilité de succès à cette question d'un étudiant dont on a par ailleurs aussi estimé la compétence sur la même échelle. On a vu que la valeur numérique n'avait d'importance que pour ce qu'on en faisait, et ici il ne s'agissait pas de communiquer.

La façon dont la compétence de l'étudiant est estimée dans le modèle de RASCH est révélatrice de la signification intrinsèque des scores 0 % et 100 %. En effet, par des procédés mathématiques qu'il serait trop long d'expliquer ici, on estime une compétence de l'étudiant sur l'échelle arbi-

traire choisie à partir du nombre de réponses correctes de cet étudiant à l'épreuve (par exemple 25 questions). Il est cependant deux situations qui ne permettent *pas* d'estimer cette compétence, c'est l'absence totale de réponse correcte (0/25) et un ensemble de réponses entièrement correctes (25/25).

Les raisons de cette incapacité d'estimer sont simples. Avec un score de 5/25 (20 % de réussite), on peut calculer l'erreur standard du pourcentage, par la fameuse formule $\sqrt{pq/N}$. Ici, cette erreur standard vaut $\sqrt{(0{,}2 \times 0{,}8)/25} = \sqrt{0{,}16/25} = \sqrt{0{,}08} \cong 0{,}09$ (9 %).

Or quand le pourcentage (p) vaut 0 ou 100, il est *impossible* de calculer ce genre d'erreur de mesure.

CHAPITRE V

L'ajustement des scores en fonction des réponses devinées

INTRODUCTION

Lorsque des étudiants ont répondu à des QCM, on recourt fréquemment à des formules mathématiques, dites « de correction » pour calculer les scores. Le but est de rendre ces derniers *comparables* aux scores obtenus à des questions ouvertes. Les tarifs prévus en cas de réponse incorrecte (TI) sont résumés par la formule célèbre – 1/k-1 (k étant le nombre de solutions proposées).

Ces procédures sont en vigueur depuis quelque soixante dix ans ; elles font périodiquement l'objet de contestations. Elles ont donné lieu à une somme énorme de travaux (théoriques et expérimentaux) et de publications (surtout dans la littérature anglo-saxonne). Nous en présenterons ici une synthèse en français.

Il n'est possible de juger tous ces travaux qu'à partir d'un cadre théorique rigoureux ; c'est pourquoi les premières sections de ce chapitre apporteront les précisions techniques nécessaires... elles le sont d'autant plus que nous voulons dénoncer des raisonnements techniques et procédures dépassés, mais dont le prestige reste grand. Par bonheur, les raisonnements, techniques et procédures de remplacement existent.

A. Quelques précisions sur certaines probabilités

1. Probabilité d'émision et probabilité d'exactitude

Dans le présent chapitre, il sera beaucoup question de la probabilité que certains événements se produisent.

Ainsi, un professeur peut essayer d'estimer la probabilité que l'élève E fournisse la réponse correcte à une question donnée. C'est ce que nous appellerons *probabilité d'émission de la réponse* correcte et que nous symboliserons par *p.ém.*

Par ailleurs, un élève qui a déjà formulé une réponse (mentalement ou par écrit) peut essayer d'estimer la probabilité que son professeur la juge correcte. C'est ce que nous appellerons *probabilité d'exactitude de la réponse fournie* et que nous symboliserons par *p.ex.*

L'étudiant ne se pose pas le problème de la p.em. (il sait ce qu'il écrit ou ce qu'il pense, c'est le professeur qui ne le sait pas). Le professeur, lui, ne se pose pas le problème de la p.ex. (il sait ce qui est correct et ce qui ne l'est pas ; c'est l'étudiant qui ne le sait pas).

La p.em. concerne un acte qui sera produit par l'étudiant.
La p.ex. concerne un (autre) acte qui sera produit par le professeur.

Donc, une expression telle que « la probabilité que la réponse soit correcte » est ambiguë si l'on ne précise pas par qui (et dans quelles circonstances) elle est estimée. C'est cependant, hélas ! la formule la plus employée ; elle laisse supposer que p.em. = p.ex., ce qui est loin d'être démontré.

Comme la plupart des autres, nous utiliserons la lettre p (majuscule ou minuscule) pour désigner la probabilité quand on ne précise pas lequel des deux événements ci-dessus est envisagé, ou que les relations considérées sont valables pour les deux événements à la fois. Cette probabilité est évidemment comprise entre 0 et 1.

$$0 \leq p \leq 1 \qquad \text{formule 1a}$$

On désigne souvent par q la probabilité complémentaire d'échec (non-réussite).

$$p + q = 1 \qquad \text{formule 1b1}$$

donc $p = 1 - q$ formule 1b2

et $q = 1 - p$ formule 1b3

Les trois formules ci-dessus sont évidemment équivalentes. Nous leur avons donné des numéros distincts pour nous y référer facilement par la suite.

2. Probabilité subjective et probabilité calculée selon un algorithme strict

Qu'il s'agisse d'estimer la probabilité d'émission (p.em.) quand on est professeur, ou la probabilité d'exactitude (p.ex.) quand on est élève, on peut s'y prendre de deux façons très différentes.

Tout d'abord, on peut *se fier à un modèle* (souvent mathématique), en suivre les principes, et appliquer (selon un algorithme strict) les formules qui en découlent. C'est souvent ce que font les professeurs, en utilisant le modèle le plus courant, à savoir « en cas de réponse au hasard, chaque solution d'une QCM a une chance égale d'être choisie ».

Il en découle que la probabilité de succès est d'une chance sur k si la QCM propose k solutions. Tous ceux qui utilisent ce modèle (et ils sont majoritaires) aboutissent à la même estimation de probabilité. C'est pour cette raison que l'on a tendance à considérer cette estimation comme « objective ». En fait, elle n'est objective que pour ceux qui ont au préalable *opté* pour le modèle considéré.

Il existe une autre façon, bien plus fréquente dans la vie, d'estimer les probabilités : se fier à son intuition. C'est le seul comportement valable dans les cas (très nombreux) où aucun modèle mathématique pertinent n'est disponible. C'est ce qu'est amené à faire un étudiant, constamment placé dans des situations « uniques au monde ». En effet, bien que la question posée soit commune à tous les étudiants, sa situation cognitive lui est strictement personnelle. Or, la probabilité d'exactitude dépend de cette situation, qui n'est pas « modélisée ». En combinant informations et raisonnements divers, l'étudiant arrivera à une (intime) « conviction » sur ses chances de succès. Mais il ne peut pas, la plupart du temps, justifier cette *probabilité subjective* par l'application d'un modèle ou d'un algorithme.

Quand nous voudrons préciser qu'il s'agit de probabilité subjective, nous utiliserons le symbole ps (en minuscules ou en majuscules). Sauf mention contraire, il s'agira alors de la probabilité subjective estimée par l'étudiant qui évalue l'exactitude de sa réponse.

3. Le nombre de succès le plus probable sur plusieurs questions

On peut essayer d'estimer, non plus la probabilité p qu'*une* réponse à une question soit correcte, mais le nombre de réussites (réponses correctes) à un test comportant NQ questions.

C'est un peu (mais en plus complexe) comme si on essayait de prédire combien de fois un joueur obtiendra la face « cinq » en lançant N fois le dé (par exemple 100 fois). Dans le cas du dé, nous disposons d'un modèle simple qui dit à peu ceci : « Ce dé est un dé normal, non biaisé. Or, dans les dés normaux, chacune des six faces a une chance égale d'être produite à chaque lancer et le résultat d'un lancer n'est influencé par aucun autre lancer ». A partir de ce modèle, nous calculons que, à chaque lancer, la face « cinq » a (comme les autres faces d'ailleurs) une probabilité de 1/6 ou 0,166666 de « sortir ». On calculera donc que *le nombre le plus probable* de résultats « cinq » sur cent jets de dé est de 16,666.

Toutefois, si l'on jette vraiment 100 fois le dé, ce nombre ne sera jamais observé. En effet, les 101 nombres *possibles* (de 0 à 100) sont tous entiers. Cependant, plus on répétera l'expérience des 100 jets de dés, plus *la moyenne* des nombres de « cinq » observés se rapprochera de 16,666 et plus ces nombres observés se répartiront autour de cette moyenne sous la forme d'une « courbe de Gauss ». Ainsi, les nombres 16 et 17 seront le plus souvent observés, tandis que des nombres tels que 2 et 30 le seront très rarement.

Si, après répétition de l'expérience un grand nombre de fois, la moyenne était nettement différente de 16,6666, alors on serait amené à *remettre le modèle en question*, notamment la première phrase : « Ce dé est un dé normal » ; on se demanderait si ce dé n'est pas pipé.

De la même façon, le nombre « attendu » ou « espéré » de réussites aux NQ questions d'un test est une valeur théorique, résultant d'un calcul ; il peut donc être décimal et par conséquent ne jamais être observé.

4. Le score le plus probable

Le score qu'un étudiant peut espérer obtenir par sa réponse à une question dépend de deux choses.

La première est la probabilité de succès de la réponse fournie par l'étudiant. Nous savons que le professeur comme l'étudiant peuvent estimer cette probabilité de succès, chacun selon ses méthodes propres.

La seconde est le barème (1) de points que le professeur a fixés. Il a en effet prévu au moins trois tarifs (1) : un en cas de réponse correcte, un autre en cas de réponse incorrecte et un troisième en cas d'omission.

Ces deux aspects sont indépendants l'un de l'autre. Le premier dépend totalement de l'étudiant, tandis que le second dépend entièrement du professeur.

Nous verrons que le comportement de l'étudiant est influencé par ces deux aspects, et même par leur combinaison. *Un étudiant choisit une solution non seulement selon ce qu'il sait, mais aussi en fonction de ce qu'il risque de gagner ou de perdre,* de l'ampleur de ce risque, et même de l'ampleur de la différence entre le gain et la perte éventuels.

Ne pas envisager *simultanément* ces deux facteurs, c'est s'enfermer dans de faux modèles, c'est renoncer à comprendre un mécanisme complexe, en le traitant comme s'il était simple. Les choix devinés ont été « compensés » pendant plus d'un demi-siècle (2) par une formule simpliste ne faisant intervenir qu'un des deux aspects ci-dessus (le barème de tarifs). Nous expliquerons d'abord dans le présent chapitre cette formule, ses justifications et ses faiblesses. Nous tâcherons ensuite de montrer pourquoi elle doit être remplacée.

(1) Voir définition précises dans la section C ci-après.
(2) Cf W. Mc Call (1920) et P. West (1923).

B. Trois situations contrastées de réponse

1. Situation 1 : Connaissance parfaite et réponse assurée

Un étudiant qui *connaît parfaitement* sa matière fournit sa réponse avec une *certitude totale.* Il n'existe, quant à l'événement qui suivra, aucun doute dans son esprit : sa réponse sera jugée exacte par le correcteur. Cet étudiant estime que l'événement « réponse jugée correcte » a 100 % de chances de se produire, que cet événement est totalement dû à ses connaissances et que le hasard n'y est pour rien.

S'il avait à exprimer sa certitude sous forme de *probabilité subjective* (ps) de succès, celle-ci vaudrait 1 (conviction absolue).

$$ps = 1$$

Si on appelle qs la probabilité (subjective) de se tromper, on voit que qs = 0 puisque qs = 1 - ps.

2. Situation 2 : Connaissance nulle et réponse au hasard

S'il *ignore totalement* la matière, l'étudiant peut soit omettre de répondre, soit répondre au hasard. Dans ce dernier cas, il *ne peut compter que sur la chance* pour espérer une issue heureuse. Si la consigne d'une QCM précise qu'une (et une seule) des k solutions est correcte, la probabilité de fournir la réponse correcte grâce à un choix entièrement *aléatoire* est égale à 1/k. L. D'Hainaut (1973) appelle cette valeur 1/k la « probabilité stochastique ». Nous la désignerons par le symbole *ph* (p comme *probabilité* et h comme *hasard*).

Pour une QCM à k solutions avec une seule correcte,

$$ph = 1/k \qquad \text{formule 1C1}$$

donc $$qh = 1 - (1/k) \qquad \text{formule 1C2}$$

ou $$qh = (k - 1)/k \qquad \text{formule 1C3}$$

On constatera que ps et qs ne font pas l'objet de formules, car chaque étudiant utilise ps et qs à sa guise. Pour chacun d'entre eux, il faudrait vérifier par exemple si ps = 1 - qs. En fait, les seules formules que l'on puisse écrire sont du type :

ps *devrait valoir* 1 - qs

De même, un étudiant ignorant, mais réaliste (qui ne surestime ni ne sous-estime ses chances de succès) devrait fixer sa probabilité subjective (ps) à la probabilité due au hasard.

En cas d'ignorance totale, ps devrait valoir ph, soit $\frac{1}{k}$

3. Situation 3 : Connaissance partielle et réponse devinée

La connaissance parfaite et l'ignorance totale constituent des *situation limites.* Souvent, l'état des connaissances de l'étudiant se situe entre ces deux extrêmes.

Coombs, Milholland et Womer (1956) parlent de *partial information.*

Nous avons défini la *connaissance partielle* comme « une situation cognitive où la connaissance de l'individu sur une question n'est ni totale ni nulle, mais intermédiaire. Dans ce cas, l'individu est souvent capable d'avancer des solutions à titre d'hypothèses, de les hiérarchiser entre elles (c'est-à-dire de distinguer celles qui lui paraissent les plus probables). Cependant, l'étudiant ne peut pas s'engager totalement à propos de l'une de ces solutions » (Leclercq, 1977). L'issue heureuse (réponse jugée exacte) résulte dès lors non seulement des connaissances de l'étudiant, mais aussi de la chance (ou sort favorable). Plus sa connaissance se rapproche de la perfection (notamment plus ps se rapproche de 1), plus le rôle du hasard est réduit (plus 1 - ps diminue).

Devant une question à choix multiple, la probabilité subjective (ps) accompagnant une solution choisie est au moins égale à 1/k. Mais, en fait, elle est presque toujours supérieure. En effet, l'étudiant peut tirer parti de sa connaissance partielle pour considérer certaines solutions comme plus probables (pour lesquelles ps $\geqq$ ph) que d'autres (pour lesquelles ps $\leqq$ ph).

L'étudiant choisira, bien entendu, la solution la plus probable.

Donc, dans ce cas $ph \leqslant ps \leqslant 1$

On parle alors de *réponse devinée,* ou, en anglais, de *guessing* (du verbe *to guess,* deviner).

Dès 1931, Granich définissait le *guessing* comme : « la tendance à répondre à des questions dont la réponse est inconnue, *soit en tout, soit en partie,* alors que l'étudiant ne peut déduire *avec certitude* une réponse de l'information dont il dispose ».

Nous avons souligné, dans cette définition, les expressions qui montrent combien il serait erroné de traduire *guessing* par *choix au hasard* ou *choix aléatoire.* Ces traductions ne conviendraient qu'à la situation limite d'ignorance totale, cas où les Anglo-Saxons parlent de choix aveugle *(blind guessing).*

L. D'Hainaut (1973) réserve d'ailleurs son expression « choix heureux par ignorance » aux réponses d'étudiants qu'il qualifie de « naïfs » (dont il a vérifié l'ignorance totale sur la matière testée).

Pour traduire le terme *guessing,* nous suggérons d'adopter soit « devinette » (NOIZET et FABRE, 1975), soit « divination » (DE LANDSHEERE, 1978), soit encore « choix deviné ».

FABRE (1977) écrit : « La traduction du terme anglais guessing par « devinette » plutôt que par l'habituel « réponse au hasard » renvoie à des connaissances bien établies par la psychologie expérimentale (LAZARUS et Mc CLEARY, 1951 ; BRICKER et CHAPNIS, 1953) : le choix aveugle *d'une réponse est un cas limite, rare,* dans la mesure où les sujets s'efforcent dans les situations difficiles de faire des hypothèses et de les étayer par des informations, aussi fragmentaires soient-elles. La conception habituelle de la « réponse au hasard » témoigne d'une méconnaissance de ce fait, ce qui se traduit par des pratiques de « correction de l'effet du hasard sur les scores » dont la validité est discutable ».

C. *Barèmes et tarifs*

1. Les conséquences d'une réponse

Nous ne considérerons ci-après que les situations de questionnement où les réponses sont suivies de conséquences administrées à l'étudiant soit par le correcteur (ou le professeur), soit par la réalité.

En classe, les conséquences sont souvent administrées par le professeur : un hochement de tête, une approbation ou une désapprobation verbale, des félicitations ou des railleries, autant de conséquences non numériques, non convenues à l'avance.

Nous appellerons tarif une *conséquence convenue à l'avance pour une réponse précise.* Dans nos exemples, ces tarifs seront numériques.

Lors d'une question, *trois tarifs au moins* doivent être précisés :
- en cas de réponse correcte,
- en cas de réponse incorrecte,
- en cas de d'omission.

Nous désignerons ces trois tarifs par TC, TI et TO. Nous appellerons barème un ensemble de tarifs.

Larousse (1981) définit un tarif comme « un tableau indiquant le coût..., le montant..., le prix... », et un barème comme un « répertoire de tarifs ».

Trop souvent, les barèmes sont improvisés. Aucune règle précise ne guide le choix des tarifs. Nous verrons que ces options irréfléchies ont un impact non négligeable sur les réponses des élèves.

2. Les barèmes les plus courants

Le barème le plus fréquemment utilisé par les enseignants est le suivant : le tarif en cas de réponse correcte (TC) vaut + 1point, le tarif en cas de réponse incorrecte (TI) vaut 0, de même que le tarif de l'omission (TO). Ce barème sera désormais appelé le *barème simple* (BS).

On rencontre parfois un barème où TC est le symétrique de TI, et où TO vaut 0. Par exemple, TC vaut + 1 et TI, - 1. Ce barème sera appelé BE (E comme « équilibre » ou « égalité » entre TC et TI).

Il arrive aussi que l'on attribue deux fois plus de points positifs à TC que de points négatifs à TI ; par exemple, TC vaut 2 et TI - 1. Ce barème sera appelé BD (D comme « double »).

BS : TC = + 1 TI = 0 TO = 0	BE : TC = + 1 TI = − 1 TO = 0	BD : TC = + 2 TI = − 1 TO = 0

Pour analyser une situation de questionnement et les réponses qui en découlent, *on doit disposer du barème en vigueur.* C'est pourquoi on s'habituera à le mentionner systématiquement quand on décrit une procédure de *testing*, de même qu'on l'a mentionné par écrit dans la consigne destinée aux étudiants.

D. Les scores attendus

1. Les scores attendus pour une question (SAQ)

Si un étudiant ne répond pas, il sait que le tarif TO (de l'omission) lui sera appliqué. Dès que le professeur apprend l'omission, lui aussi sait que le score à la question sera TO.

Nous verrons, dans ce chapitre, que la réponse de l'étudiant (ou l'omission) sont influencés non pas par le score observé (qui n'est pas encore connu de l'étudiant qui répond) mais par les scores attendus qui peuvent être calculés (ou, plus souvent, estimés) avant de fournir la réponse. L'étudiant choisit la réponse qui lui apporte le score attendu le plus favorable.

Si l'étudiant répond, il est possible de calculer le *score attendu* à la question, à partir des tarifs possibles et de la probabilité que chacun de ces tarifs soit appliqué.

La réponse de l'étudiant peut lui rapporter TC points (positifs) avec une probabilité (de succès) égale à p. ou TI points (négatifs) avec une probabilité (d'échec) égale à q.

Le *score attendu* est la somme de ces deux tarifs pondérés par leurs probabilités :

$$SAQ = (p.TC) + (q.TI)$$ formule 2a

où SAQ = score attendu à une question

Le *score observé* sera l'un des scores possibles et sera désigné par SQ (score à la question).

Exemple :

A une question, un étudiant fournit une réponse et indique que, pour lui, elle a 80 % de chances d'être correcte. Que cette probabilité de 0,8 ait été estimée « subjectivement » par l'étudiant, ou calculée « objectivement » selon un modèle théorique, on a donc p = 0,8 et q = 0,2.
Le barème D (ou barème double) est d'application : (TC = +2, TI = -1).
Dans ce cas, SADQ (c'est à dire SAQ avec le barème D) vaut
$(0,8 \times 2) + (0,2 \times -1) = 1,6 - 0,2 = 1,4$.

Avec le même barème (BD), une réponse ayant une probabilité de
0,7 donnerait SADQ = (+2 x 0,7) + (-1 x 0,3) = 1,4 - 0,3 = 1,1
0,5 donnerait SADQ = (+2 x 0,5) + (-1 x 0,5) = 1 - 0,5 = 0,5
0,3 donnerait SADQ = (+2 x 0,3) + (-1 x 0,7) = 0,6 - 0,7 = -0,1
0,1 donnerait SADQ = (+2 x 0,1) + (-1 x 0,9) = 0,2 - 0,9 = -0,7

Le score observé (avec le barème D) à cette question (SDQ) ne peut valoir que +2 ou -1 (les valeurs de TC et TI).

En cas de réponse au hasard, le score attendu à une question est appelé SAQH et vaut :

$$SAQH = (ph.\ TC) + (qh.\ TI)$$ formule 2b

En fonction des formules 1c, si cette question est une QCM à k solutions proposées,

$$SAQH = \left(\frac{1}{k} \cdot TC\right) + \left(\frac{k-1}{k} \cdot TI\right)$$ formule 2c

Si le barème simple (BS) est d'application, le score attendu à une question pour le barème simple en cas de réponse au hasard (c'est-à-dire SASQH) peut être calculé par la formule suivante :

$$SASQH = (ph.\ 1) + (qh.\ 0) = ph = \frac{1}{k}$$

donc

$$SASQH = 1/k > 0$$ formule 2d

Si la QCM comporte quatre solutions proposées (k = 4),

$$SASQH = \left(\frac{1}{4} \cdot 1\right) + \left(\frac{3}{4} \cdot 0\right) = \frac{1}{4}$$

Nous verrons que la formule de correction pour divination (retrait de points en cas de réponse incorrecte) la plus célèbre est basée sur le calcul de SASQH.

2. Les scores attendus pour un test (SAT)

Prenons le cas d'une épreuve où toutes les questions ont le même poids et où un seul barème est d'application pour toutes les questions.

Pour un étudiant donné, le score total (ST) à l'ensemble d'une épreuve est la somme (symbole Σ) de ses scores aux diverses questions (SQ), ce qui s'écrit :

$$ST = \sum_{1}^{NQ} SQ$$ formule 3a

Cette formule est évidemment d'application pour des scores attendus. Donc, en cas de réponse au hasard, on peut écrire que

$$SATH = \sum_{1}^{NQ} SAQH$$

formule 3b

Dans une épreuve, composée de QCM ayant chacune le même nombre k de solutions proposées (épreuves « uniforme pour k »), SAQH est le même pour toutes les questions.

S'il y a NQ questions, le score attendu pour un test auquel on répond au hasard (SATH) se calcule de la façon suivante :

Dans une épreuve homogène, SATH = NQ. SAQH — formule 3c

Si le barème simple BS est d'application, en utilisant la formule 2d, on transforme la formule 3c comme suit :

Dans une épreuve « uniforme pour k », SASTH vaut NQ/k — formule 3d

Exemple :

Pour BS et dix QCM à cinq solutions, SASTH = NQ/k = 10/5 = 2. Tout score (score au test avec barème simple) à peine supérieur ou inférieur à 2 a des chances d'être en partie le résultat de réponses fournies au hasard. Plus un score est élevé (et distant de 2), moins il est susceptible d'avoir été obtenu grâce au hasard.

E. La correction pour divination standard ou barème G

1. Des scores non comparables

Imaginons qu'un étudiant réponde à deux épreuves A et B.
L'épreuve A est constituée de dix questions ouvertes. L'épreuve B est constituée de dix QCM à *deux solutions proposées* (dont une seule est correcte, ce dont l'étudiant est averti).

Le même barème simple (BS) est d'application pour les deux épreuves. Notre étudiant a fourni cinq réponses correctes au test A et cinq réponses correctes au test B. Or, pour le test B, *cinq points* est le score attendu (le plus probable) d'un étudiant ignorant qui répondrait *au hasard* (SASTH).

Si l'on attribuait cinq points à l'élève pour chacun des deux tests, ces scores ne seraient pas comparables car les deux tests n'ont pas la même *puissance.*

TVERSKY (1964) définit la *puissance* d'un test comme *« 1 moins la probabilité d'atteindre la performance parfaite par chance »* et considère (p. 387) que l'on doit chercher à créer des tests dont la puissance soit la plus grande possible. Il est évident que, dans notre exemple, la probabilité d'atteindre cinq points par chance n'est pas la même pour le test A et pour le test B.

On comprend que, dans *un souci de rendre comparables* des scores à des épreuves différentes (A et B par exemple), les utilisateurs de tests à choix multiple aient cherché à rectifier les scores qui résultent de ce mode de questionnement.

2. Le barème G ou la correction pour divination standard

La rectification la plus connue consiste à retirer des points en cas de réponse incorrecte [TI négatif). Dans ce barème G, TC est fixé à +1, TO à 0, et TI est calculé de *manière qu'en cas de réponse* au hasard, le *score attendu à la question* (SAGQH) *soit nul,* comme c'est le cas pour une question ouverte. Ce score attendu (SA) sera désigné par SAGQH. Les trois dernières lettres (G, Q et H) de ce code indiquent :
- que le barème G est en vigueur
- que c'est un score attendu pour une question (Q)
- que ce score est attendu en cas de réponse au hasard (H).

Que doit valoir TI dans le barème BG ? De la formule 2b, nous savons que SAGQH = (ph. TC) + (qh. TI). Puisque SASQH doit être nul, il faut que ph. TC contrebalance exactement qh. TI.

Or, TC = 1 ; donc ph doit valoir -qhTI.

et

$$\text{TI doit valoir } \frac{-ph}{qh} \quad \text{(dans le barème G)}$$

formule 4a

Voici quelques exemples :

Si une QCM a 5 solutions (k = 5), SAGQH vaudra (0,20 x 1) + (0,80 x TI) ; donc TI doit valoir -1/4 ;
Si une QCM a 4 solutions (k = 4), SAGQH vaudra (0,25 x 1) + (0,75 x TI) ; donc TI doit valoir -1/3 ;
Si une QCM a 3 solutions (k = 3), SAGQH vaudra (0,33 x 1) + (0,66 x TI) ; dont TI doit valoir -1/2 ;
Si une QCM a 2 solutions (k = 2), SAGQH vaudra (0,50 x 1) + (0,50 x TI) ; dont TI doit valoir -1.

On voit [1] que

$$\text{TI doit valoir } \frac{-1}{k-1}$$

formule 4b

Ce barème sera appelé BG car il est connu sous le nom de *correction for guessing* ou *correction pour divination* (G. De Landsheere, 1979) standard.

Cette pratique est ancienne ; on la trouve déjà chez W. Mc Call (1920)... et elle est déjà critiquée par P. West (1923).

Si TC est différent de +1, la formule de calcul de TI est modifiée en conséquence dans ce nouveau barème (BG') :

$$\text{Si TC} = b \quad \text{TI doit valoir} - \frac{b}{k-1} \text{ dans BG}$$

formule 4c

[1] On peut facilement passer de la formule 4a à la formule 4b en modifiant dans la première le numérateur selon la formule 1C1 et le dénominateur selon la formule 1C2 (ces formules ont été présentées ci-avant).
On a alors TI = 1/k 1-(1/k), et il suffit de multiplier par k le numérateur et le dénominateur pour obtenir la formule 4b.

F. Le barème G en cas d'épreuve uniforme (k est constant)

1. Test où l'omission est autorisée

Soit :

NQ = le nombre de questions au test.
NR = le nombre de réponses émises.
NRC = le nombre de *réponse correctes* observées à ce test *pour un étudiant* particulier.
NRI = le nombre de *réponses incorrectes* observées pour le même étudiant.
NO = le nombre d'*omissions* observées, pour cet étudiant.
NRE = NRI + NO = le nombre d'*échecs* observés, pour cet étudiant.
NRH = le nombre de *réponses fournies au hasard* par cet étudiant. Ce nombre n'est pas observé ; il est inconnu du correcteur.

Bien entendu,

$$\text{NRC} + \text{NRI} + \text{NO} = \text{NQ}$$

formule 5a

a) Le principe de correction

Lorsqu'aucune correction n'est appliquée, c'est-à-dire avec le barème simple (BS), le score simple total (SST) est égal au nombre de réponses correctes (NRC).

Quand l'épreuve est homogène (k est identique pour toutes les questions), il est possible d'obtenir un score corrigé « pour *guessing* » (SGT) à partir du nombre de réponses correctes (NRC) et du nombre de réponses incorrectes (NRI) de l'élève à cette épreuve. Les omissions éventuelles (NO) de l'élève ne sont *pas utiles pour le calcul* de SGT : (score au test, corrigé pour divination selon la formule classique).

$$\text{Dans le cas où l'omission est autorisée, SGT} = \text{NRC} - \frac{\text{NRI}}{k-1}$$

formule 5b

(voir démonstration dans les sections ci-après).

Exemple :

Pour une épreuve homogène de 20 QCM où k = 5,

Un étudiant a fourni 13 réponses correctes et 6 incorrectes (et donc 1 omission).

$$SGT = 13 - \frac{6}{4} = 11{,}5$$

Un autre étudiant a fourni 13 RC et 2 RI (et donc 4 omissions).

$$SGT = 13 - \frac{2}{4} = 12{,}5$$

On voit que les omissions jouent un rôle important dans le calcul de SGT, bien qu'elles n'apparaissent pas dans la formule 5b ci-dessus.

Cette formule 5b a des effets équilavents (le score SGT résultant est le même) à l'application du barème BG pour chacune des questions.

b) Justification de la formule 5b

Cette formule découle du modèle 1 de l'activité mentale décrit par B. Choppin (1971) pour une question et que l'on peut exprimer comme suit en ce qui concerne un test :

> Confronté aux NQ questions d'un test, l'étudiant
> - omet un certain nombre NO d'entre elles,
> - fournit un certain nombre de réponses correctes par *connaissance* (ou compétence) : NRCC
> - fournit un certain nombre de réponses au hasard (NRH) pour le reste des questions dont il ignore la réponse.

Selon ce modèle 1, NO + NRCC + NRH = NQ — formule 5c

Une fraction $\frac{1}{k}$ de ces NRH réponses au hasard constituent des réponses correctes par hasard (NRCH) ; ce nombre est inconnu lui aussi :

Selon le modèle 1, NRC = NRCC + NRCH — formule 5d

L'autre fraction $\frac{k-1}{k}$ de ces NRH réponses au hasard est faite de réponses incorrectes par hasard (NRIH).

Selon le modèle 1 , $NRIH = \frac{k-1}{k} NRH$ — formule 5e

Notons, que selon le modèle 1 , NRIH est la seule source d'erreurs (NRI) :

Selon le modèle 1 , NRI = NRIH — formule 5f

Si l'on combine les formules 5c, 5d et 5e, on obtient :

Selon le modèle 1 , $\overbrace{NO + \overbrace{NRCC + NRCH}^{NRC} + NRIH}^{NR} = NQ$ — formule 5g

(Cf section 4c)

La correction traditionnelle (barème BG appliqué au total d'un test à k constant) vise à attribuer à l'étudiant un score au test égal à NRCC (inconnu) et non pas égal à NRC (observé).

Remarque : Pour calculer NRCC, il suffirait de connaître NRCH (et de soustraire NRCH de NRC dans la formule 5d).

Or, de la formule 5g, il découle que $NRCH = NRH - NRI$ — formule 5h

Des formules 5e et 5f, on sait que $NRI = NRIH = \frac{k-1}{k} NRH$.

Si l'on isole NRH, on a $NRH = \frac{k}{k-1} NRI$ — formule 5i

Introduisons cette valeur de NRH dans la formule 5h.

On a : $NRCH = \frac{k}{k-1} NRI - NRI = \frac{k.NRI}{k-1} - \frac{(k-1)\,NRI}{k-1}$

$$= \frac{(k.NRI) - (k.NRI) + NRI}{k-1}$$

$$NRCH = \frac{NRI}{k-1}$$ formule 5j

2. Test où l'omission est interdite

a) Le principe général

Lorsque l'omission n'est pas autorisée, le modèle 1a sert de référence. Ce modèle est en tous points conforme au modèle 1, décrit ci-dessus, sauf que l'étudiant ne peut omettre.

Dans le modèle 1a, NO = 0
NQ = NRC = NRE
NRE = NRI

formule 6a

Lorsque l'omission est interdite et que l'épreuve est homogène, selon le modèle 1a,

$$SGT = NRC - \frac{NQ - NRC}{k-1}$$

ou $$SGT = NRC - \frac{NRE}{k-1}$$

formule 6b

ou encore selon le modèle 1a,

$$SGT = \frac{K.NRC - NQ}{k-1}$$ formule 6c

En effet, la formule 6c découle de la formule 6b :

$$SGT = NRC - \frac{NR}{k-1} = \frac{(k-1)\,NRC}{k-1} - \frac{NRE}{k-1} = \frac{k.NRC - NRC - NRE}{k-1} = \frac{k.NRC - (NRC + NRG)}{k-1}$$

b) Un cas particulier : les questions VRAI-FAUX

La consigne VRAI-FAUX standard autorise l'omission. Il existe cependant des situations d'alternative où l'omission n'est pas possible, comme le montre l'exemple suivant ; il s'agit d'une épreuve d'orthographe destinée à des correcteurs d'épreuves d'imprimerie.

Biffez les mots soulignés incorrectement orthographiés.

Le rinocéros courrait dans la savane entre les heuphorbes.
1 2 3 4

Tout mot biffé est codé FAUX et tous les autres sont codés VRAI : il n'y a donc aucune place pour l'omission ni, bien sûr, pour une « valorisation pour omission » (voir section G ci-après).

Puisque $$SGT = NRC - \frac{NQ - NRC}{k-1}$$ formule 6b

et que, dans ce cas-ci, k valant 2, k-1 vaut 1, la formule devient :

$$SGT = NRC - NQ + NRC$$

donc dans le modèle 1a (voir chapitre 4, section c) pour des questions VRAI-FAUX,

$$SGT = 2NRC - NQ$$ formule 6d

NQ est une constante ; SGT est donc une fonction linéaire de NRC, comme SST l'est. Le classement des étudiants entre eux est donc équivalent, dans ce cas-ci, avec SST ou avec SGT (ce dernier est, en général, inférieur à SST et ne peut lui être supérieur).

Exemple : Pour une épreuve de dix questions (NQ = 10), les scores NRC peuvent aller de 0 à 10 ; les scores SGT peuvent aller, eux, de -10 à +10.

3. Pour le taux de réussite d'un élève

La formule 5h peut s'écrire sous une autre forme si l'on part de proportions pour un étudiant (d'où la lettre E)
- de ses réponses correctes (Epc)
- de ses réponses incorrectes (Epi)
- de ses échecs (Epe)
- de ses omissions (Epo).

Bien entendu

$$Epc + Epo + Epi = 1$$
$$Epe = Epo + Epi$$

formule 7a

Selon le modèle 1 , $SGT = NQ\left(Epc - \frac{Epi}{k-1}\right)$ (section 4c)

formule 7b

Selon le modèle 1a, $SGT = NQ\left(Epc - \frac{Epe}{k-1}\right)$ (section 4c)

formule 7c

4. Pour le taux de réussite à une question

On peut, de même exprimer le score moyen (de tous les élèves d'une classe) à une *question donnée*, avec le barème G (SGQ).

Nous désignerons par

pc la proportion de réponses correctes pour la question.
pi la proportion de réponses incorrectes pour la question.
pe la proportion d'échecs pour la question (pe = pi + po)
po la proportion d'omissions pour la question.

Bien entendu

$$pc + po + pi = 1$$
$$pe = po + pi$$

formule 8a

Selon le modèle 1 , $SGQ = pc - \frac{pi}{k-1}$ (section 4c)

formule 8b

Selon le modèle 1a, $SGQ = pc - \frac{pe}{k-1}$ (section 4c)

formule 8c

Dans la littérature, on trouve souvent le symbole p pour notre notation pc et le symbole q pour notre notation pe.

La formule 8c apparait alors sous la forme : $SGQ = p - \frac{q}{k-1}$

G. *La valorisation de l'omission ou barème V*

Le barème G a, deux fonctions : d'une part, rendre comparables les scores obtenus à des QCM et les scores obtenus à des questions ouvertes, et d'autre part, décourager les choix aveugles en encourageant l'omission. Pour atteindre ce second but, le barème *BG pénalise les réponses incorrectes* (TI négatif).

On peut aussi, et c'est ce que fait le barème V, *valoriser les omissions :* le tarif de l'omission (TO) est alors positif. Bien entendu, les scores (SVQ et SVT) ainsi obtenus ne sont pas comparables aux scores (SSQ et SST) obtenus à des réponses ouvertes avec le barème simple (BS), où le tarif de l'omission (TO) est nul.

Dans le barème V (V comme *valorisation*), ou BV, le tarif de l'omission (TO) a une valeur égale au score attendu à une QCM en cas de réponse au hasard avec le barème simple (BS).

Dans le barème V,
$TO = SASQH = 1/k$

formule 6e

Ainsi, pour une QCM à quatre solutions, chaque omission rapporte à l'étudiant un quart de point.

Le tableau ci-après présente les différents tarifs (TC, TI et TO) des barèmes BS, BV et BG, les scores attendus à une question en cas de réponse au hasard (SAQH) et indique si les scores obtenus en appliquant ces barèmes sont comparables aux scores obtenus pour des questions ouvertes quand on applique le barème simple (BS).

Barème	Tarifs			Score attendu pour une réponse au hasard	En cas d'ignorance, l'étudiant a intérêt à :	Les scores sont-ils comparables avec des scores pour questions ouvertes ? (BS en vigueur) :
	TC	TO	TI			
BS	+1	0	0	SASQH = 1/k	répondre au hasard	non
BG	+1	0	-1/k-1	SAGQH = 0	omettre ou répondre	
BV	+1	+1/k	0	SAVQH = 1/k	au hasard omettre ou répondre au hasard	oui / non

Le barème corrigé (BG) et le barème valorisant l'omission (BV) aboutissent au même classement des sujets entre eux. Le second est au début plus favorablement accueilli par les étudiants ; à la longue, ils s'habituent à associer la connaissance nulle à 1/k point (cela peut prendre du temps, tant est tenace l'association entre l'ignorance et le score nul). Ils comprennent alors que les deux scores SGQ et SVQ sont équivalents, à un déplacement près sur l'échelle des scores. La section k du présent chapitre en apporte une preuve expérimentale.

La valorisation de l'absence de réponse présente cependant un avantage psychologique : celui de déculpabiliser l'omission, alors qu'elle est trop souvent présentée aux étudiants comme pire que toutes les erreurs. Or, la compréhension de ses propres incertitudes incite l'étudiant à vérifier (au moyen d'un dictionnaire, par exemple, pour l'orthographe) avant de se prononcer. Avouer ses doutes ou s'abstenir de répondre est souvent bien plus souhaitable socialement que des déclarations péremptoires, mais incorrectes, qui orientent autrui sur des pistes fausses, parfois dangereuses. Dans cette perspective, le doute et même l'abstention doivent être valorisés. Nous avons approfondi cette question ailleurs (D. Leclercq, 1978, 1983).

H. La correction dans le cas de plusieurs réponses autorisées (NRA > 1)

La consigne peut préciser que le *nombre* de *réponses* autorisées (NRA) est compris entre 0 et k. Nous examinerons les trois cas qui se présentent le plus fréquemment.

CAS 1 : Le nombre de solutions correctes (NSC) vaut 1 (NSC = 1). On applique alors la *correction for guessing* standard à chaque réponse fournie, la somme donnant le *score SGT* de l'étudiant. Le maximum est obtenu quand l'étudiant fournit une seule réponse (NR = 1) et que cette réponse est correcte. Les sujets qui choisissent toutes les solutions obtiennent le même score (nul) que ceux qui n'en choisissent aucune.

Exemple : Considérons une QCM à k solutions, le barème G est appliqué. Imaginons qu'un étudiant choisisse toutes les k solutions : il gagnera une fois un point, à k-1 reprises il perdra 1/(k-1) points.

Dans le cas où $0 \leq NRA \leq k$ et $NSC = 1$, si $NR = k$,
alors $SGQ = 1 - [\,(k-1)\,\frac{1}{k-1}\,] = 0$ formule 9a

où

SGQ, avec NR = k signifie « score à la question avec le barème G pour le choix de toutes les solutions proposées ».
NRA = nombre de réponses autorisées.
NSC = nombre de solutions correctes.
NR = nombre de réponses fournies.

CAS 2 : Une ou plusieurs solutions peuvent être correctes ($1 \leq NSC \leq k$). Soit une QCM qui présente k solutions dont NSC sont correctes, que doit valoir TI pour que SAGQH d'une réponse vaille 0 ?

Dans le cas où $0 \leq NRA \leq k$ et $0 \leq NSC \leq k$, si $NR = 1$
TI doit valoir $-\frac{NSC}{k-NSC}$ pour que $SAGQH = 0$ formule 9b

où

SAGQH = score attendu pour des réponses au hasard, avec le barème G.
NSC = nombre de solutions correctes.

De nouveau, celui qui rejette toutes les solutions obtient un score total corrigé nul, c'est-à-dire le même score SGQ que celui qui n'en rejette aucune.

Dans le cas où $0 \leq NRA \leq k$ et $0 \leq NSC \leq k$,
si NR = k,
$SGQ = NSC - [\,(k\text{-}NSC)\,\frac{NSC}{k\text{-}NSC}\,] = 0$

formule 9c

Ainsi, pour k = 6 et NSC = 2, on accordera 1 point par réponse correcte (SGQ peut donc valoir 2 au maximum) et on retirera 2/4 (soit 0,5) point par réponse incorrecte. Si l'on préfère travailler avec des valeurs entières, les points gagnés et perdus pourront être multipliés par $\frac{k\text{-}NSC}{NSC}$.

Il faut cependant être conscient que, dès lors, on change le poids de cette question par rapport aux autres questions de la même épreuve.

En fait, sauf si k et NSC sont constants d'une question à l'autre, le poids des questions varie. Pour avoir un poids constant, il faudrait attribuer $\frac{1}{NSC}$ pour chaque réponse correcte et $\frac{1}{k\text{-}NSC}$ pour chaque réponse pour chaque réponse incorrecte.

CAS 3 : La consigne permet (ou impose) de choisir ou de barrer chacune des solutions proposées.
On se trouve dans une situation bien connue : chaque solution constitue en fait une QCM à deux solutions (vrai-faux). On accorde alors 1 point par réponse correcte et -1 point par réponse incorrecte, ce qui correspond à la correction classique.

I. La correction standard surcorrige ou sous-corrige selon les cas

Pour estimer la capacité d'un étudiant, on se réfère d'habitude à son score obtenu à un test composé de questions ouvertes. Il semble dès lors évident qu'avec des QCM, le score corrigé SGT constitue une meilleure estimation de la capacité de l'étudiant que ne l'est le score simple SST.

Pourtant, cette correction est elle même criticable.

Certains chercheurs, notamment, Holzinger (1924) et Gulliksen (1960), ont pris position contre la correction pour divination standard pour les questions vrai-faux : ils avancent que *si* tous les étudiants répondent à toutes les questions, ils seront classés dans le même ordre avec le score total simple SST ou le score total corrigé SGT.

Mais cette condition est rarement réalisée, car la consigne « Ne devinez pas : une correction pour divination sera appliquée » a pour effet d'encourager les omissions et suscite chez l'étudiant ce que Stanley (1954) appelle une « correction psychologique ». D'autres auteurs ont montré que, selon les cas, les formules décrites ci-dessus aboutissent soit à des *surcorrections* soit à des *sous-corrections.*

1. Les cas où les formules surcorrigent

Rappelons que la formule 5b, $SGT = NRC - \frac{NRI}{K\text{-}1}$,
apparemment conçue pour le modèle 1 (qui distingue NRI et NRO) est, en fait, basée sur le modèle 1a qui assimile NRI (le nombre de réponses incorrectes) à NRIH (le nombre de réponses incorrectes dues au hasard), ce qu'exprime la formule 5f.

On a vu que le nombre de réponses correctes dues au hasard (NRCH) valait $\frac{NRI}{k\text{-}1}$.

Dès lors, dans le modèle 1a, on peut partir du nombre de réponses incorrectes observé (NRI) pour estimer le nombre de réponses au hasard (NRH) :

Dans le modèle 1a, $NRH = NRI + \frac{NRI}{k\text{-}1}$

formule 10a

Or, il n'est pas possible de faire confiance au modèle 1a sur ces deux points. Dès 1952, Davis attirait l'attention sur le fait qu'un certain nombre de réponses incorrectes (RI) étaient dues à des erreurs commises « en toute bonne foi », l'étudiant croit que sa réponse est exacte à la suite de conceptions fausses ou d'une méconnaissance de la matière.

Il devient dès lors nécessaire de distinguer un type supplémentaire de réponse incorrecte : RIM, celles qui sont commises par *méconnaissance*, dont le nombre (inconnu) est NRIM.

Dès lors, $NRI = NRIM + NRIH$

L'estimation de NRH (formule 10a) en est modifiée et devient :

$$NRH = NRIH + \frac{NRIH}{k\text{-}1}$$ formule 10c, puis

$$NRH = (NRI - NRIM) + \frac{NRI - NRIM}{k\text{-}1}$$ formule 10d

La formule 10a apparaît dès lors comme un cas particulier où NRIM = 0, ce qui est rare.

On voit donc que cette formule 10a *surcorrige*. Malheureusement, la formule 10c est inapplicable dans la mesure où NRIM est inconnu du professeur (comme NRIH d'ailleurs). Seul l'étudiant détient cette information, qui est la clé du problème de la correction.

Pour compenser cette surcorrection, divers auteurs ont proposé des *pénalisations* (valeurs de TI) *réduites*, mais ces valeurs sont empiriques et ne reposent sur aucun raisonnement précis.

Ainsi, Michael *et al.* (1963) ont proposé d'utiliser une valeur de TI réduite de moitié :

$$TI = -\frac{1}{2} / (k\text{-}1)$$ formule 10e

Ce qui est fondamentalement remis en cause ici, c'est l'opposition que font certains modèles entre « l'étudiant qui *sait* » et « l'étudiant qui *ne sait pas* », comme si seules ces deux situations existaient. Il est évident que toutes les nuances existent entre ces deux extrêmes. Souvent, l'élève sait, mais n'en est pas sûr ou croit savoir alors qu'il se trompe. C'est ce dernier cas qui est envisagé ici : à lui seul, il remet en cause ces modèles, qui se révèlent faux et dépassés.

Dans le prolongement de ce qui vient d'être discuté, il apparaît notamment que la formule 5g doit être amplifiée comme suit :

$$NRO + \underbrace{\underbrace{NRCC}_{NRC} + \overbrace{\underbrace{NRCH}_{NRC} + \underbrace{NRIH}_{NRI}}^{NRH} + \underbrace{NRIM}_{NRI}}_{NR} = NQ$$ formule 10f

2. Les cas où les formules sous-corrigent

Dans le même article de 1952, Davis signale que les formules de correction standard *sous-corrigent en cas de connaissance partielle.*

Dans ces cas, en effet, l'étudiant est capable d'éliminer x des k solutions proposées (modèle 2 selon CHOPPIN) avant de choisir au hasard parmi celles qui restent.

Ici, c'est le dénominateur k-1 des formules qui est contesté. Il doit, en effet, être remplacé par k-x-1 : par exemple, la formule 10c devient :

$$NRH = NRIH + \frac{NRIH}{k\text{-}x\text{-}1}$$ formule 10g

Cette formule ne peut déboucher sur des applications pratiques, puisque x est inconnu du professeur. Seul l'étudiant connaît ces valeurs. On voit à nouveau que c'est lui qui détient les informations indispensables pour fonder la correction.

Dans ce cas-ci, le choix au hasard ne s'effectue plus qu'entre k-x solutions, et produit statistiquement moins de réponses incorrectes (NRIH) qu'avec k solutions. Le nombre de réponses émises au hasard (NRH) est donc sous-estimé par la formule 10a. La formule classique (5a) sous-corrige.

C'est ce qui fait dire à Hammerton (1965), par exemple, que les étudiants ont intérêt à répondre au hasard quand la correction pour divination standard est d'application.

Ces réflexions mettent en cause la « réponse au hasard » telle que la conçoivent les modèles 1 et 2 : ils ignorent les nuances, négligent la situation de connaissance partielle ce qui donne lieu non pas à une réponse au hasard, mais à une réponse devinée, ce qui est très différent.

3. La sous-correction et la surcorrection se compensent-elles ?

On pourrait penser que les surcorrections (dues à des réponses incorrectes « de bonne foi ») et les sous-corrections (dues à des réponses en état de connaissance partielle) se compensent.

Rien ne permet d'étayer une telle hypothèse. Il se pourrait que l'on observe quelquefois une telle compensation sur un ensemble d'étudiants et de questions. Il n'en reste pas moins que certains étudiants seront favorisés et d'autres lésés, sans que le professeur puisse les distinguer. Les étudiants, eux, le savent. Ceux qui sont favorisés, d'habitude, ne le font pas remarquer. Ceux qui sont lésés font valoir, avec amertume, que la formule les considère comme des tricheurs et les pénalise aveuglément, et ils ont raison.

L'état des connaissances n'est pas seul à déterminer la réponse d'un étudiant, l'attitude de chacun vis-à-vis du risque joue aussi : certains étudiants sont très audacieux, d'autres très circonspects. Ce trait de personnalité peut être très général (pour toutes les activités abordées par l'individu) ou limité à certaines matières (par exemple, un élève sera audacieux en sciences, mais timoré en français) ou encore à certaines formes d'examens (par exemple, un élève sera audacieux uniquement en face de QCM).

Ce problème n'a pas échappé aux chercheurs. Il est traité dans la suite du présent chapitre.

On en revient toujours au même point : le professeur essaye d'appliquer des corrections « aveugles » à partir de paramètres qu'il ignore, mais dont les étudiants ont une assez bonne conscience. La solution est évidemment de partir du modèle 3 de CHOPPIN (selon lequel les étudiants envisagent PLUSIEURS solutions et en pondèrent mentalement la PLAUSIBILITE), de présenter une consigne et des barèmes qui y soient conformes, de recueillir de l'étudiant lui-même les informations nécessaires, bref d'utiliser les probabilités subjectives ou des degrés de certitude.

J. Personnalité et prise de risque

1. L'indice de Ziller

Afin de résoudre les problèmes qui viennent d'être évoqués, Ziller a imaginé un indice R' représentant le degré d'*acceptation du risque* d'un étudiant. Cet indice ferait intervenir d'une part le nombre de réponses fournies au hasard (NRH) et d'autre part, le nombre de questions pour lesquelles la réponse correcte est méconnue (NRM).

Ziller ne donne, hélas, pas de définition verbale de NRM. Dans les formules qu'il propose, on constate qu'il définit NRM comme la somme de NRH et de NRO.

La formule de Ziller est :

$$R' = \frac{NRH}{NRM}$$ formule 11a

Evidemment, cette formule est purement théorique, car NRH et NRM sont tous deux inconnus. Pour résoudre ce problème, Ziller remplace NRM par NRH + NO, puis NRH par sa valeur de la formule 5i.

Nous avons donné le nom RZ à l'indice que Ziller (1957) propose alors pour estimer R' :

$$RZ = \frac{\frac{k}{k-1} NRI}{(\frac{k}{k-1} NRI) + NRO}$$ formule 11b

Le calcul de cet indice est soumis à deux restrictions importantes : le test doit être homogène (k constant pour toutes les questions) et NRI + NRO doit être supérieur à 0.
RZ varie de 0 à 1. Les valeurs faibles indiquent des prises de risques faibles.

Par ailleurs, cet indice est basé sur un présupposé théorique pour nous inacceptable : toute réponse incorrecte (RI) est considérée comme une réponse aléatoire (RH). Autrement dit, cette formule ignore les réponses incorrectes dues à une méconnaissance (RIM), introduites dans les formules 10b, 10d et 10f.

D'autre part, cet indice postule aussi que la probabilité de succès d'un choix au hasard est de $\frac{1}{k}$.

Cela revient à ignorer le mécanisme du modèle 2. La probabilité complémentaire d'échec, c'est-à-dire $\frac{k-1}{k}$,

est donc surestimée. Par conséquent, cet indice RZ est passible des mêmes critiques que la correction pour divination standard : il sous-estime le phénomène de la divination.

2. L'indice de Ziller-Slakter

Slakter (1968) a proposé un indice inspiré de celui de Ziller que nous désignerons par RZS. Il l'a adapté à la consigne de Coombs, Milholland et Womer (1956), qui demande non pas de choisir la solution correcte, mais d'éliminer les solutions incorrectes.

Il propose $$RZS = \frac{k.NSIC}{(k.NSCI) + NSIO}$$ formule 11c

où

NSCI = nombre de solutions correctes incorrectement éliminées.
NSIC = nombre de solutions incorrectes correctement éliminées.
NSIO = nombre de solutions incorrectes omises (non éliminées).
k = nombre de solutions proposées (identique pour toutes les questions).

RZS varie de 0 à 1. Les valeurs faibles indiquent des prises de risques faibles, comme dans la formule 11b.

M. Slakter a utilisé cette formule dans cinq expériences où la consigne différait. Chaque expérience présentait 100 QCM à quatre solutions proposées. Le tableau ci-dessous présente les RZS moyens et les scores moyens calculés selon le barème G où l'étudiant reçoit un point par solution incorrecte éliminée et -(k-NSC) si une solution correcte est éliminée.

Groupe	Consigne	RZS moyen	Score moyen
1	Eliminez *une* solution incorrecte. (On devine peu).	0,17	53
2	Eliminez *deux* solutions incorrectes. (On devine moyennement).	0,45	100
3	Eliminez *trois* solutions incorrectes. (On devine fortement).	1,0	135*
4	Eliminez autant de solutions incorrectes que vous voulez (consigne classique de Coombs *et al.*).	0,74	132*
5	*Choisissez* la réponse correcte (consigne habituelle).	RZ = 0,73	

* = différence non significative.

3. L'utilisation de l'indice de Ziller dans une formule de correction

L'indice de Ziller remet évidemment en cause la formule 5b

$(SGT = NRC - \frac{NRI}{k-1})$ et l'estimation de NRCH par la formule 5h

(NRCH = NRH − NRI).

On va donc les corriger comme suit :

$$SGT = NRC - (NRCH \,.\, RZ)$$ formule 12a

Or, (formule 5j), $NRCH = \frac{NRI}{k-1}$

On aura donc $SGT = NRCH - (\frac{NRI}{k-1} \,.\, RZ)$

Si l'on remplace RZ par sa valeur dans la formule 10b, on a :

$$SGT = NRC - \frac{NRI}{k-1} \,.\, \frac{\frac{k}{k-1} \,.\, NRI}{[(\frac{k}{k-1}) NRI] + NO}$$ formule 12b

ou, après simplification,

$$SGT = NRC - \left(\frac{NRI}{k-1} \cdot \frac{kNRI}{kNRI + kNO - NO} \right)$$

formule 12c

ou encore

$$SGT = NRC - \frac{(NRI)^2}{(k-1)\,[NRI + NO - (NO/k)]}$$

formule 12d

Il s'agit là , à notre connaissance, de la formule la plus sophistiquée visant à « corriger le score pour compenser la divination » selon l'approche classique.

4. Autres approches de la tendance individuelle à deviner

a) Les questions fictives (sans contenu)

Certains chercheurs ont introduit, dans des tests, des *questions* à contenu fictif. Voici l'une de ces questions (Cross et Frary, 1977) :

> Laquelle des lois ci-dessous ne concorde pas avec la règle de Pauling sur les gaz polyatomiques ?
> 1. La loi de Boyle.
> 2. La loi de Charles.
> 3. La loi combinée des gaz.
> 4. La loi d'Avogadro.

La règle de Pauling concernant les gaz polyatomiques est fictive. Slakter (1967) a montré que la réponse à de telles questions sous la consigne « score corrigé » est un bon indicateur de la personnalité qui prend beaucoup de risque en situation d'examen.

b) L'autopondération

Dès 1938, Swineford liait la certitude exprimée par l'étudiant aux conséquences qui y sont associées. Sa consigne était la suivante :
« Choisissez les points que vous souhaitez obtenir en cas de réponse correcte : deux points, trois points ou quatre points. En cas d'erreur, vous perdrez le double des points demandés ».

La consigne de Swineford est appelée « autopondération » par Fabre (1975).

Swineford estimait « la tendance à répondre en prenant des risques » par la formule suivante :

$$G = \frac{\text{NRI marquées 4}}{NRI + \frac{NRO}{2}} \times 100$$

formule 12e

En 1941, il soumet l'indice G ci-dessus à une analyse plus approfondie et conclut que la capacité d'un étudiant dans la matière est indépendante de son comportement de « joueur ». Par contre, selon lui, les contenus ou méthodes qui constituent des innovations encourageraient à deviner ou à « jouer » au hasard.

K. Influence de la correction pour divination sur la validité et la fidélité des mesures

Sabers et Feldt (1968) ont abordé ce problème de façon expérimentale :

1. Ils ont proposé un *test d'aptitude* à l'algèbre à deux groupes d'étudiants (A et B) avec la consigne classique « Ne devinez pas aveuglément ; toute réponse incorrecte entrainera la perte d'une fraction de point » (sans autre précision).
2. Ils ont proposé le même test d'aptitude à deux *groupes de contrôle* (C et D) avec une consigne annonçant le barème simple (BS).
3. *Neuf mois plus tard*, ils ont présenté aux sujets des quatre groupes, un *test de rendement* (en mathématique moderne pour les groupes A et C, et en mathématique traditionnelle, pour les groupes B et D).
4. Ils ont invité les enseignants à évaluer, selon leur expérience quotidienne, la compétence en mathématique de chaque étudiant, au moyen d'une note.

Les nombres d'élèves dans chaque groupe étaient compris entre 380 et 613.

Le test d'aptitude était composé de QCM à quatre solutions, si bien que la correction pour divination classique aurait été $NRC - \frac{NRI}{3}$: le total des réponses correctes diminué du nombre de réponses incorrectes multiplié par 1/k-1 (formule 5b). Or dans l'étape 1 ci-dessus, le barème en vigueur n'a pas été précisé (tout spécialement la valeur de TI par laquelle multiplier NRI).

Les auteurs se sont demandé quel serait le cœfficient de NRI qui, dans une équation de régression multiple, permettrait de prédire le mieux les résultats aux épreuves de rendement neuf mois plus tard.

Rappelons qu'une équation de régression multiple (du premier degré) se présente sous la forme :

$Y' = C_1X_1 + C_2X_2 + C_3X_3 + \ldots C_nX_n + b$

où Y = valeur *observée* de la variable à prédire, c'est-à-dire la variable critère (ici le rendement neuf mois plus tard).
Y' = valeur *prédite* de la variable critère ou cible (Y).
$X_1, X_2 \ldots X_n$ = variables prédictives (ici X1 sera remplacé par NRC et X2 par NRI).

$C_1, C_2 \ldots C_n$ = cœfficient de régression (ou poids de ces diverses variables prédictives).
b = constante.

La corrélation entre les Y *observés* (ici les résultats aux épreuves de rendement neuf mois plus tard) et les Y' *calculés* (ici à partir de NRC et NRI au test d'aptitude) est appelée corrélation *multiple* (ou R multiple) par ce que Y' est calculé à partir de *plusieurs* (au moins deux) variables prédictives.

Ici, l'équation a la forme : $Y' = C_1\,NRC + C_2\,NRI + b$

Or, on a fixé $C_1 = +1$ (barème simple). Fixons b à une valeur de notre choix, par exemple à zéro. Rappelons que puisque l'on veut pénaliser les réponses incorrectes, le cœfficient C_2 doit être négatif.

L'équation se réduit donc à $Y' = NRC + (C2.NRI)$.

Il suffit dès lors de calculer la valeur de C_2 qui maximise la corrélation entre Y et Y' (R multiple le plus élevé possible). Cette formule de calcul du cœfficient « de régression » est proche d'une formule de calcul de la corrélation. Elle est en effet basée sur le calcul de la covariance entre X et Y et de la variance de Y.

On s'attend à ce qu'au terme de ces calculs, C_2 vaille $\frac{-1}{k-1}$, la valeur de TI du barème G, c'est-à-dire, dans ce cas ci $-\frac{1}{3}$.

Voici les résultats que Sabers et Feldt obtiennent.

	Valeurs Y à prédire	Equation de prédiction permettant de calculer Y'	R multiple entre Y et Y'	r simple entre Y et NRC	
				groupe expérimental	groupe de contrôle
Math. moderne	Test de rendement	$NRC - \frac{1}{8,3}NRI$	0,74	0,74	0,78
	Notes du professeur	$NRC - \frac{1}{3,4}NRI$	0,73	0,73	0,69
Math. traditionnelle	Test de rendement	$NRC - \frac{1}{1}NRI$	0,69	0,68	0,75
	Notes du professeur	$NRC - \frac{1}{1,2}NRI$	0,64	0,61	0,64

On peut faire les constatations suivantes :

1. Les pénalisations (valeur de TI) sont bien plus élevées pour le test de mathématique traditionnelle. Dans ce dernier, les réponses incorrectes seraient donc plus révélatrices de la compétence (plus exactement de l'incompétence) que les réponses incorrectes dans le test de mathématique moderne.
2. Les cœfficients de corrélation multiple ne sont cependant pas très différents de cœfficients de corrélation simple, basé sur les seuls nombres de réponses correctes. Les corrélations simples des deux groupes de contrôle sont d'un niveau équivalent à celles du groupe expérimental. Ces deux observations amènent à penser que la *validité prédictive n'est pas augmentée par la correction pour divination.*

L. Consignes et devinette

Dans les sections précédentes du présent chapitre, on a vu que les consignes sont conçues pour influencer le comportement de l'étudiant. Elles ont, en effet, un impact (voir par exemple l'expérience de WOOD 1976 dans la section K3C ci-après).

Nous regroupons dans cette section la description de recherches (et de leurs résultats) ayant pour but de répondre expérimentalement à une série de questions telles que :

- Les étudiants se comportent-ils réellement comme le modèle de la correction for guessing le prévoit ?
- L'annonce d'une correction pour solutions devinées a-t-elle l'effet dissuasif escompté ?
- Est-elle d'ailleurs bien comprise ?
- Tous les étudiants réagissent-ils de la même façon ?
- En particulier, le goût du risque, différant selon les individus, n'influence-t-il pas de façon importante leur façon de répondre ?

Dans tous les raisonnements qui suivent, les auteurs font deux hypothèses fondamentales :

1. L'étudiant cherche à obtenir le plus de points possible.
2. L'attrait des points est uniquement lié (fonction linéaire) à leur valeur numérique : recevoir six points est deux fois plus attractif que recevoir trois points.

Cette seconde hypothèse est, évidemment la plus faible des deux : on peut, par exemple, imaginer que plus l'étudiant est assuré de sa réussite, moins il est intéressé à gagner davantage de points. En d'autres termes, l'utilité (subjective) ne correspondrait pas forcément à la valeur (objective) des points. VAN NAERSSEN et BRUYNIS (1966) ont mené une recherche expérimentale pour tester cette seconde hypothèse. Leur résultat ne la remet pas en cause. Cette expérience est une des rares du genre.

Nous nous en tiendrons donc, jusqu'à preuve du contraire, au deux hypothèses énoncées ci-dessus.

1. La consigne habituelle du barème G

Au moment où il communique les consignes d'une épreuve aux étudiants, le professeur *doit* leur signaler qu'une correction pour réponses devinées sera appliquée.

C'est évidemment cette annonce préalable qui a pour fonction de dissuader les étudiants de devenir. Cette consigne est-elle efficace ? La suite de cette section apportera des informations sur ce point.

Nous pensons qu'il faut communiquer tout simplement le barème (l'ensemble des tarifs). Or, traditionnellement, on utilise une consigne verbeuse et, à notre avis, très criticable.

Lorsqu'ils recourent au barème G (ou BG), de nombreux auteurs anglo-saxons utilisent la consigne suivante, due à DAVIS (1967).

> Votre score sera calculé comme suit :
> le nombre de réponses correctes moins une fraction du nombre de vos réponses incorrectes.
>
> Il vous est recommandé de répondre aux questions, même quand vous n'êtes pas sûr de votre réponse, spécialement si vous êtes capable d'éliminer une ou plusieurs solutions incorrectes ou si vous avez une préférence pour une solution par rapport aux autres.
>
> Cependant, il est préférable d'omettre de répondre plutôt que de choisir de façon aveugle parmi les solutions proposées.

A notre avis, seule la seconde phrase de cette consigne est nuancée et exacte. Or, c'est souvent cette deuxième phrase qui est supprimée dans les versions condensées de la consigne.

La phrase 1 est, à notre avis, trop vague : la valeur exacte de TI devrait être communiquée aux étudiants, d'une part parce qu'ils y ont droit, d'autre part parce qu'ils ont besoin de cette précision pour prendre leurs décisions de réponse.

La troisième phrase contient une contre-vérité puisque, avec le barème BG, choisir aveuglément entraîne un score attendu égal à 0, c'est-à-dire *égal* à l'omission, *et NON plus défavorable* que l'omission.

Lors d'une première interrogation, la majorité des étudiants respectent scrupuleusement cette consigne et donc omettent au lieu de devenir aveuglément ou même raisonnablement. Ce faisant, il se nuisent à eux-mêmes.

Heureusement, l'expérience venant, les étudiants voient de plus en plus clair dans les enjeux. Ils comparent entre eux les modes de réponses, les résultats obtenus... et tirent la conclusion pratique qui s'impose. Bref, après trois ou quatre tests, la consigne (erronée) ci-dessus ne parvient plus à décourager la devinette, aveugle ou non.

2. La méthode des deux consignes successives

LORD (1975) a démontré que le score SGT est un meilleur estimateur de la compétence que le score SST. Or tout son raisonnement repose sur l'hypothèse que la consigne a eu un impact sur le comportement de l'étudiant. L'annonce du barème de correction pour guessing (BG) amènerait l'étudiant à omettre pour un certain nombre de questions (NO) auxquelles ils auraient fourni des réponses au hasard (NRH) si le barème simple (BS) avait été annoncé.

La méthode des deux consignes successives permet de tester ce genre d'hypothèse. Elle se pratique en trois étapes :

1. On demande aux étudiants de répondre *avec un crayon bleu* à une épreuve homogène de NQ QCM (k est le même pour toutes les questions), la consigne annonçant le barème BG.
2. On demande aux étudiants de répondre *avec un crayon rouge* aux questions omises précédemment, la consigne annonçant le barème BS pour ces réponses rouges (rappelons que le barème BS ne pénalise pas les erreurs et encourage donc les choix au hasard).
3. On calcule, *sur ces réponses rouges*, pour chaque étudiant un score SST (comme annoncé) et un score SGT (à des fins de recherche).

Si l'étudiant choisit ses réponses rouges au hasard (comme la théorie le suppose), le score rouge simple (barème BS) est égal au nombre de réponses rouges (NRR) divisé par le nombre de solutions possibles (k), soit NRR/k et le score rouge corrigé selon le barème BG vaut 0. En effet (1), de NRR/k, on retire $NRR - \frac{NRR}{k}$ fois la pénalisation $1/K - 1$, soit

$$(NRR - \frac{NRR}{k})\ (1/k - 1).$$

CROSS et FRARY (1977) ont appliqué cette procédure à une épreuve homogène de 40 QCM (avec $k = 4$) présentée à 241 étudiants. Lors de la première étape, ils ont utilisé la consigne classique de DAVIS (1967). Ils ont observé que les scores rouges s'éloignaient fort des scores attendus, c'est-à-dire NRR/k pour SST et 0 pour SGT.

Le score simple *rouge* (SST) fut, en moyenne, de 0,328, ce qui dépasse nettement la valeur attendue (0,25). Le *score corrigé rouge* (SGT) fut, en moyenne, 0,225, ce qui dépasse nettement la valeur attendue 0.

(1) On peut montrer que $NRR/k = (NRR - \frac{NRR}{k})(1/k - 1)$.

Le deuxième nombre vaut $\frac{NRR - NRR/k}{k-1}$. Multiplions numérateur et dénominateur par K. Nous avons $\frac{kNRR - NRR}{k(k-1)}$. Mettons NNR en évidence dans le numérateur ; nous avons $\frac{NRR(k-1)}{k(k-1)}$, soit NRR/k.

La valeur observée (0,328) du score simple rouge (SST) est proche de 0,33 (valeur attendue quand les QCM présentent trois solutions). Tous les étudiants qui étaient en mesure de rejeter une (ou deux) des quatre solutions auraient donc dû, dès le départ, fournir pour ces questions une réponse devinée plutôt que de s'abstenir. La consigne (de DAVIS) le leur recommandait d'ailleurs.

M. SLAKTER (1968) a lui aussi observé que les étudiants qui prennent peu de risque en respectant la consigne : « Ne répondez pas au hasard » sont désavantagés. C'est ce que SLAKTER a appelé « la pénalisation de ceux qui ne devinent pas » (*The penalty for not guessing*). Il a obtenu de 20 étudiants une seconde réponse rouge à des questions omises plusieurs jours au préalable. Les scores rouges étaient corrélés à .27 avec le degré de prise de risque (voir formule de ZILLER), mais seulement à .10 (non significative) avec l'aptitude.

Les observations de CROSS et FRARY (les réponses « forcées » ou « rouges » sont meilleures que sous l'effet du pur hasard) sont confirmées par des travaux antérieurs dus à MEAD et SMITH (1967), EBEL (1968), SHERRIFFS et BOOMER (1954), SLAKTER (1968), VOTAW (1936).

3. Le comportement des étudiants face à la consigne « Ne devinez pas »

Pour introduire une épreuve semblable à celle décrite ci-devant (deux consignes successives), CROSS et FRARY (1977) ont projeté le texte de la consigne de DAVIS sur grand écran et l'ont lue en même temps à haute voix. Après l'étape 1, ils ont posé une QCM... sur la consigne. Voici ces questions avec, entre parenthèses, le pourcentage de réponses (sur 407 sujets) :

Au début du test, la consigne vous a recommandé :

(7 %) 1. d'omettre de répondre si vous n'êtes pas certain de vous ;

(2 %) 2. d'omettre de répondre si vous ne pouvez éliminer qu'une seule solution incorrecte ;

(75 %) 3. d'omettre de répondre si vous devez devinez aveuglément ;

(15 %) 4. de n'omettre aucune réponse.

Près de 25 % des étudiants ne se souvenaient plus de la consigne !

Les mêmes auteurs ont aussi interrogé les étudiants sur leurs habitudes de réponse.

Sans considérer les consignes, qu'avez-vous l'habitude de faire lorsque vous n'êtes pas certain de la réponse à une QCM?

(6 %) 1. Je m'interdis de deviner ;

(60 %) 2. Je devine seulement si j'ai une préférence pour une solution ou si je peux en éliminer au moins une ;

(34 %) 3. Je fournis une réponse, aveuglément s'il le faut ;

(0 %) 4. Je suis toujours certain des réponses que je donne.

Enfin, une troisième question a porté sur le comportement des étudiants lors du test qu'ils venaient de passer.

Pendant ce test, j'ai décidé

(43 %) 1. de suivre les consignes (1) aussi strictement que possible ;

(15 %) 2. de deviner plus souvent que les consignes (1) le recommandaient, en espérant que la chance jouerait en ma faveur. Je crois cependant que j'aurais obtenu un score satisfaisant sans ces devinettes supplémentaires ;

(30 %) 3. de deviner plus souvent que les consignes (1) le recommandaient parce qu'autrement je n'aurais sans doute pas obtenu un score satisfaisant à ce test ;

(12 %) 4. de deviner moins souvent que les consignes (1) le recommandaient pour éviter de perdre des points pour des réponses incorrectes.

(1) Dans le texte original de la question, on précise « les consignes telles que je les ai comprises ».

Parmi tous les sujets qui ont correctement compris la consigne (ils ont choisi la solution 3 à la question 1 ci-dessus), CROSS et FRARY ont repéré des « saints » et des « pécheurs ».

Les « saints » sont définis comme ceux qui ont choisi la solution 2 à la question 2 (je devine seulement si...) et la solution 1 à la question 3 (je suis les consignes).

Les « pécheurs » sont ceux qui ont choisi la solution 3 à la question 2 (je réponds même de façon aveugle) et les solutions 2 ou 3 à la question 3 (je devine plus souvent...).

Sur 71 « saints » et 31 « pécheurs », ils ont calculé divers indices, rassemblés dans le tableau ci-dessous.

	Saints	*Pécheurs*	*Différence significative à P.05*
Score simple moyen SST *bleu*	10,00	8,89	non
Score corrigé moyen SGT *bleu*	6,48	5,09	non
Nombre d'omissions	2,69	2,94	non
Prise de risque (1)	2,20	2,90	oui
Score au questionnaire MMPI de personnalité (2)	10,21	14,03	oui
Moyenne des scores SGQ *« rouges »* (0 en principe)	0,268	0,167	non
Proportion moyenne des scores SSQ *« rouges »* (0,25 en principe)	0,318	0,285	non

Observons qu'ici le score rouge corrigé moyen SGQ est supérieur à O. Les *étudiants qui ont compris et ont essayé de suivre consciencieusement les consignes* (les saints) *sous-estiment l'information* partielle qu'ils possèdent et perdent plus de points (0,268 en moyenne) en ne devinant pas que les pécheurs (qui ne perdent en moyenne que 0,167 point par question omise).

En conclusion de leur étude, CROSS et FRARY mettent l'accent sur le problème éthique que pose l'utilisation d'une consigne adéquate, et recommandent d'entraîner systématiquement les étudiants à répondre efficacement aux tests.

(1) Voir la section K du présent chapitre.

(2) Le questionnaire était le *Minnesota Multiphasic Personnality Inventory (MMPI)*. Un score élevé est un signe d'anxieté, de sentiment d'inadaptation.

4. L'importance du libellé de la consigne

a) L'expérience de TAYLOR

TAYLOR (1966) a réalisé une expérience pour mesurer l'effet de trois consignes différentes :

Consigne A (destinée à encourager la devinette) :
« Vous devez obtenir le plus possible de réponses correctes. Ne craignez pas de deviner et de tenter votre chance à chaque question. »

Consigne B (neutre) :
« Faites du mieux que vous pouvez. »

Consigne C (destinée à minimiser la devinette)
« Ne fournissez pas de réponse avant d'avoir étudié soigneusement la question et ne répondez que si vous êtes certain de votre réponse. »

TAYLOR n'observe pas de différence significative de rendement entre les groupes qui ont reçu ces différentes consignes.

b) L'expérience de WOOD

Dix ans plus tard, WOOD (1976) met au point une expérience assez semblable par son principe général, mais présentant dans les *consignes une différence fondamentale*. WOOD insiste, en effet, sur les *conséquences* précises (gains et pertes) de trois types de situations :

A. Travaillez soigneusement et répondez au plus grand nombre de questions possible. Votre score sera égal au nombre total de réponses correctes que vous aurez données (barème BS).

B. Si vous êtes tenté de deviner, rappelez-vous qu'une réponse incorrecte ne vous rapportera aucun point, alors que chaque réponse omise vous rapportera 1/5 de point (barème BV).

C. Si vous êtes tenté de deviner, rappelez-vous qu'un choix incorrect vous fera perdre 1/4 de point, soit – 0,25, alors que chaque réponse entraîne simplement la note 0 (barème BG).

c) Comparaison des deux expériences

On voit que la consigne A de Wood est plus précise que celle de TAYLOR. De plus, la consigne B de TAYLOR a disparu et la consigne C est remplacée, chez WOOD, par deux consignes beaucoup plus précises.

WOOD a appliqué ces trois consignes à deux épreuves de mathématique comptant 45 QCM à cinq solutions chacune.

De l'analyse des données recueillies, WOOD conclut que les groupes A, B et C étaient bien équivalents au départ et que les consignes B et C ont bien eu un effet de dissuasion. Le groupe A omet moins souvent que les autres et le groupe C est celui qui commet le moins d'erreurs (les groupes B et C omettent à peu près également).

5. La devinette avouée

A la fin d'une épreuve comportant 45 QCM à cinq solutions chacune, WOOD (1976) a posé la question suivante :

« A peu près combien de fois avez-vous totalement deviné la réponse à une question ? »

1. Pas plus de cinq fois.
2. Pas plus de dix fois.
3. Plus de dix fois.

Voici les résultats observés pour deux groupes d'étudiants (« mathématiques inférieures » et « mathématiques supérieures » cumulés par nous.

Réponse au questionnaire : J'ai deviné	Groupe A barème simple (BS) TI = 0 devinette encouragée	Groupe B barème valorisé (BV) TO = + 0,20 devinette découragée	Groupe C barème guessing (BG) TI = - 0,25 devinette découragée
Moins de 5 fois	72	124	122
Entre 5 et 10 fois	61	46	38
Plus de 10 fois	39	11	12
Nombre d'étudiants	172	181	172

Le groupe A avoue avoir deviné beaucoup plus que les deux autres, qui se comportent en gros de la même façon.

Le mérite de cette expérience est de montrer la nette relation entre le comportement et la consigne (ici les barèmes).

On peut évidemment regretter que les déclarations des étudiants ne soient pas corroborées par une mesure objective. La procédure ci-après utilisée par EBEL répond à cette préoccupation.

6. Le contrôle objectif de la devinette avouée

R. EBEL (1968) a présenté à des étudiants une épreuve faite de questions VRAI-FAUX, corrigées selon *le barème V* (valorisation de l'omission).

Plus exactement, il a demandé aux étudiants de noter, *sur une feuille séparée*, leurs *réponses complètement aléatoires*. Pour chacune d'elles, il garantissait le tarif $\frac{1}{K}$ (ici 1/2).

Cette technique est intéressante. D'une part, en utilisant une feuille de réponses séparée, on valorise l'omission (car ces réponses n'apparaissent pas sur la feuille de base). D'autre part, en demandant quand même une réponse sur cette feuille séparée, on peut *juger de* son exactitude et donc de *la pertinence des décisions* de l'étudiant d'utiliser la feuille de base ou la feuille séparée (celle des omissions).

C'est ce qu'à fait EBEL en calculant les pourcentages de réponses correctes pour les quatre groupes d'étudiants qui ont participé à cette expérience. Comme il s'agit de questions VRAI-FAUX, il aurait dû observer, pour ces réponses *aléatoires*, un taux de réussite de 50 %.

Or, il a observé respectivement 56, 55, 54 et 52 % des réponses exactes dans les quatre groupes.

EBEL conclut que les étudiants ont donc (légèrement) intérêt à répondre à l'aveuglette lorsque la correction pour divination classique est d'application.

Pour les sujets, la corrélation entre le nombre de réponses correctes au test et le nombre de choix aveugles était de – 0,21. Par contre, entre le nombre de réponses correctes au test et la proportion de réussite aux choix aveugles, la corrélation n'était que de – 0,01 (c'est-à-dire nulle).

Pour les questions, la corrélation entre la facilité et le taux de choix aveugles est plus élevée (- 0,52), comme on pouvait s'y attendre.

M. Corrections non classiques pour divination

La correction pour réponses devinées est basée sur l'attirance *théorique* de chaque solution, attractivité obtenue en divisant 1 par le nombre de solutions proposées (k). Cette démarche serait légitime si le nombre ainsi obtenu était proche de l'*attirance réelle* de ces solutions, que l'on estime par leur « popularité ». Par *popularité d'une solution* pour une population déterminée, on entend le pourcentage de sujets (de cette population) qui ont choisi cette solution (avec la consigne « choisir ») ou l'ont éliminée (avec la consigne « éliminer »).

Il est bien connu que l'on ne rencontre jamais de QCM dont toutes les solutions (ou les seuls distracteurs) présentent des popularités équivalentes.

A la suite de Chernoff (1962), on a envisagé de corriger, non plus en fonction de l'attirance théorique (1/k), mais en fonction de l'attirance réelle mesurée par la « popularité observée » des diverses solutions proposées. La méthode la plus simple consiste à tenir compte de la popularité de la solution correcte, autrement dit, de la seule facilité de la QCM. Une méthode un peu plus élaborée consiste à tenir compte de la popularité de *chaque* solution.

1. La correction de Risse (1972) basée sur la facilité de la question

Nous avons désigné par p la fréquence des réponses correctes à une QCM, c'est-à-dire la popularité de la réponse correcte

q (= 1 – p) est la fréquence d'échecs à cette QCM.

Risse applique un barème (que nous appelerons R) avec les tarifs suivants.

Dans le barème R,
TC = p . q = p (1-p)
TE = – $(p)^2$

formule 13a

rappel : TC = tarif en cas de réponse correcte
TE = tarif en cas d'échec (omission ou réponse incorrecte).

Pour chaque question, le score moyen attendu (avec le barème R) est 0 (SARQ = 0).

En effet, un taux p d'étudiants, recevra p (1-q) et un taux 1 – p d'étudiants recevra – (p)

On aura donc : [p p. (1-p)] + [(1-p). – $(p)^2$] = SAQ
c'est-à-dire : [p^2 (1-p)] – [(1-p) p^2] = 0 = SAQ

Fabre (1975) apporte les précisions suivantes :

« On peut montrer facilement que cette procédure revient à pondérer chaque réussite par p et à attribuer 0 à tout échec. En effet, ajoutons, pour chaque question, p^2 à tout score.

Si la réponse est correcte, le gain est alors :
[p (1-p)] + p^2 = p

Si la réponse est incorrecte, la perte est égale à
-p^2 + p^2 = 0.

La moyenne des notes ainsi obtenues est égale à p^2 que l'on peut sans difficulté ramener à 0 ».

La caractéristique essentielle de cette procédure est donc que *l'écart entre le gain et la perte*, pour une question, est une fonction linéaire croissante de la facilité de cette question (voir tableau et figure ci-après). Ceci revient à accorder d'autant plus d'importance à un item qu'il a été mieux réussi et à pénaliser d'autant plus un échec qu'il est rare.

2. La correction basée sur l'écart type des scores observés à la question

Fabre (1975) développe le raisonnement qui suit.

Soit une série de scores 0 et 1 dont la moyenne est p (facilité de la question) et l'écart type $\sqrt{pq}$. On peut concevoir un barème T où

TC = $1/\sqrt{p}$
TE = 0

formule 13b

On voit que l'écart entre le gain et la perte vaut toujours

$$1/\quad p$$

Un autre barème (que nous appellerons BF, avec F pour Fabre) permet de garder le même écart : on attribue

$\sqrt{q/p}$ en cas de réponse correcte
– $\sqrt{p/q}$ en cas d'erreur

formule 13c

Le tableau ci-dessous (Fabre 1975, p. 55) fait apparaitre que les barèmes BF et BR pénalisent fortement les échecs rares. Seul BF valorise fortement les réussites rares. C'est un moyen d'atténuer les effets de la devinette.

P	Barème BR (de Risse) TC	TE	Ecarts	Barème BF (lié à l'écart-type) TC	TE	Ecarts
0,10	0,09	-0,01	0,10	3,00	-0,33	3,33
0,20	0,16	-0,04	0,20	2,00	-0,50	2,50
0,30	0,21	-0,09	0,30	1,53	-0,65	2,18
0,40	0,24	-0,16	0,40	1,22	-0,82	2,04
0,50	0,25	-0,25	0,50	1,00	-1,00	2,00
0,60	0,24	-0,36	0,60	0,82	-1,22	2,04
0,70	0,21	-0,49	0,70	0,65	-1,53	2,18
0,80	0,16	-0,64	0,80	0,50	-2,00	2,50
0,90	0,09	-0,81	0,90	0,33	-3,00	3,33

3. La correction de D'Hainaut basée sur les réponses de sujets naïfs

La popularité de chaque solution peut être mesurée non pas pour des élèves qui sont sensés connaître la matière, mais pour des élèves qui, au contraire, *ne savent rien* du contenu.

D'Hainaut (1975) a proposé un indice de correction qui tient compte des « choix heureux par ignorance » réalisés par des *« sujets de niveau nul »* (l'expression est de D'Hainaut).

Les *sujets de niveau nul* sont « des sujets qui, par rapport aux connaissances ou aptitudes que prétend mesurer une question, peuvent être considérés comme représentant le niveau nul d'aptitude ou d'acquisition. Par exemple, des élèves de cinquième primaire sont des « sujets nuls » pour une question de temps primitifs grec ou une question relative à la formule des stérols ».

A des élèves de première année de l'enseignement secondaire (douze ans), qui n'ont en principe aucune connaissance de la mécanique, D'Hainaut a présenté la question suivante :

Pourcentages des choix effectués par les sujets de niveau nul (popularités) :

(1) 26 % →
(2) 2 % →
(3) 35 % →
(4) 37 % →

Pour faire dévier un corps en mouvement en terrain plat, il faut lutter contre :
1. Son poids.
2. Sa masse.
3. Son poids et sa masse.
4. Ni l'un, ni l'autre.

RC = 2

Cet exemple montre de nouveau combien l'attirance réelle (popularité) d'une solution proposée peut être différente de l'attirance théorique (ici 25 %) et combien la *correction* pour divination classique serait inappropriée.

D'Hainaut (1974) s'intéresse à la notion de « choix heureux par ignorance » qu'il définit comme « le fait de donner une réponse juste à une question fermée en se laissant guider par d'autres facteurs que la connaissance de la réponse ou l'application correcte du processus mental qui y conduit » (p. 61).

Le *gain statistique par ignorance* dans une épreuve, désigné par G, est défini comme « le nombre moyen de points que les choix heureux par ignorance peuvent faire gagner à un sujet de niveau nul qui passe tout l'épreuve » (G = gain individuel probable d'un sujet de niveau nul).

La *perte statistique par ignorance* dans une épreuve, désignée par P, est définie comme « le nombre moyen de points que les choix malheureux d'un étudiant de niveau nul peuvent lui faire perdre ».

D'Hainaut propose un cœfficient de compensation, que nous appellerons DH semblable à celui de la « correction pour divination traditionnelle ». Le score obtenu à un item doit être multiplié par DH.

DH = 1-G, où G = moyenne des gains obtenus par les sujets de niveau nul à cette question.

Le score est donc atténué puisque $0 \leq DH \leq 1$.

Dans le barème DH (BDH),
TC = DH = 1 – G
où G = gain moyen obtenu par les sujets de niveau nul à la question.

formule 13d

G (et donc DH) est calculé sur une autre population que celle qui passe le test.

Le problème, avec ce genre de correction, est qu'on ne dispose pas de critères précis qui définissent ce qu'est le « niveau nul ».

4. Corrections basées sur les popularités des distracteurs

Parmi les barèmes possibles, le plus connu est le suivant (que nous appellerons BP).

Si, pour une question, on appelle pi la proportion totale des réponses incorrectes et pd(i) la popularité du distracteur i (proportion des réponses i), on peut attribuer pour le choix du distracteur i, un tarif TI(i) négatif calculé comme suit :

Dans le barème BP.

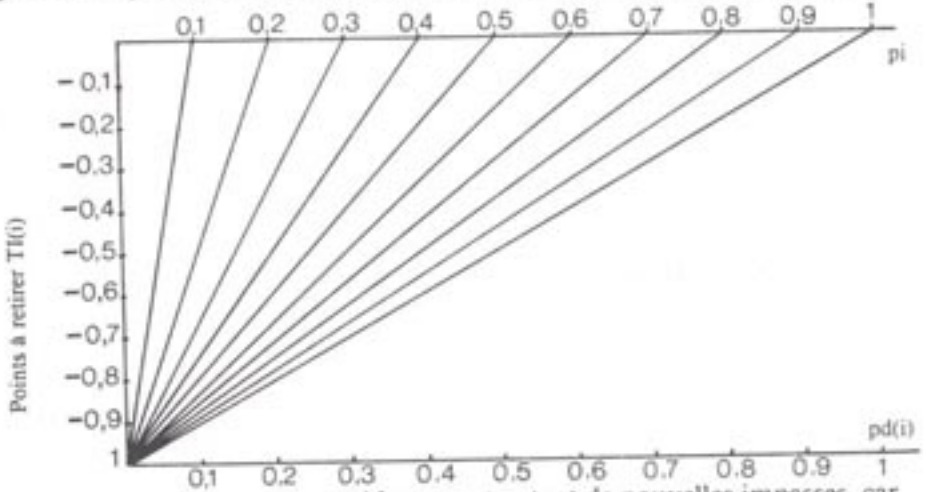
$$TI(i) = \frac{pi - pd(i)}{pd(i)}$$

formule 13c

Cette méthode pénalise le choix d'un distracteur en fonction inverse de la popularité dont il jouit par rapport aux autres distracteurs.

La pénalisation est nulle quand la popularité des autres distracteurs est nulle.

L'abaque ci-dessous permet de trouver les points à retirer (en verticale) lorsque l'on connaît pi (horizontale supérieure) et pd(i) (horizontale inférieure). On élève de pd(i) une verticale ; l'endroit où celle-ci coupe l'oblique partant de pi détermine le nombre de points à retirer (sur la verticale).



Nous pensons que ces procédures sont autant de nouvelles impasses, car elles se fondent sur des attirances statistiques *moyennes*, sans rapport avec l'attirance qu'une solution exerce *sur un étudiant particulier*. Néanmoins, le présent chapitre ne pouvait passer sous silence ces procédures concurrentes de la correction classique pour divination.

N. Conclusions

Au terme de ce long exposé technique, il ressort que :

1. Les procédures de *« correction for guessing »* habituelles, et tout spécialement leurs axiomes ne résistent pas à un examen théorique approfondi et ne paraissent pas constituer une base saine pour attribuer des scores aux élèves.
2. L'étudiant possède l'information (son degré de doute pour chaque question) qui permettrait un système de cotation sain et juste qui dépende *à la fois* de l'individu *et* de la question, et mieux encore, de l'*interaction des deux*.
3. Les procédures adaptées à cette approche existent, elles sont d'une utilisation facile tant par le professeur que par l'étudiant : elles portent les noms de « probabilités subjectives » ou « degrés de certitude ». Un exposé de ces procédures générales (non limitées aux QCM) et surtout de leurs exploitations possibles dépasserait le cadre de ce travail et a fait l'objet d'ouvrages (Leclercq 1975, 1983a et 1983b) divers, dont le plus récent, dans cette collection sous le titre « Auto-évaluation et connaissance partielle ».

CHAPITRE VI

La qualité des épreuves scolaires

A. La validité d'un test.
B. La fidélité d'un test.
C. Les formules d'estimation de la fidélité d'un test.
D. Le nombre optimal de questions et de solutions.
E. L'entraînement des étudiants.
F. La fraude et sa prévention.

Introduction

Ce que les étudiants craignent le plus, dans une épreuve scolaire, est qu'elle soit injuste. Ils redoutent que l'épreuve elle-même ne soit pas représentative de l'ensemble des connaissances et performances à évaluer (c'est le problème de la validité). Ils redoutent également que la note qui leur est accordée (la mesure) soit « fluctuante » et dépende de circonstances passagères, telles que le hasard du choix des questions, des variations de leur propre « forme » ou de celle de l'examinateur, etc. (C'est le problème de la fidélité).

Les QCM peuvent tantôt contribuer à la validité d'un test, tantôt la mettre en danger, selon le type de validité considéré.

Le manque de fidélité des mesures obtenues au moyen d'une épreuve scolaire peut avoir plusieurs sources. La première est l'instrument lui-même, et tout spécialement le nombre de questions. Les QCM apportant une amélioration sensible aux tests sur ce point, nous décrirons les formules d'amélioration de la fidélité avec le nombre de questions.

Deux autres sources de manque de fidélité sont le contexte d'administration et l'individu. C'est pourquoi nous examinerons les problèmes posés par la préparation des étudiants et par la fraude, sa prévention et sa détection.

A. La validité d'un test

Se poser le problème de la validité d'un test, c'est se demander si les *mesures* qu'il permet de recueillir *sont représentatives* de la capacité, du « trait psychologique », de la compétence que l'on veut évaluer.

Autrement dit, les mesures obtenues n'ont-elles pas été biaisées par d'autres variables que celle que l'on veut mesurer ? *Le test mesure-t-il cette variable, toute cette variable et rien que cette variable* ?

En s'inspirant de Thorndike et Hagen (1969) et de De Landsheere (1975 et 1977), on peut distinguer quatre moyens d'assurer ou d'éprouver la validité d'une épreuve, donnant lieu à quatre types de validité (ces types ne s'excluant pas les uns les autres) :
- la validité de contenu
- la validité de « construct »
- la validité concourante
- la validité prédictive.

1. La validité de contenu

Le test couvre-t-il la matière ? Les questions posées constituent-elles un échantillon représentatif de l'ensemble des questions possibles ?

Pour assurer la validité de contenu, Hively *et al.* (1968) et Schoemaker (1975) recommandent de définir « l'univers » des questions possibles, puis de choisir, dans cet univers, selon des critères précis.

En permettant de poser un grand nombre de questions, les QCM contribuent à ce type de validité.

2. La validité de « construct »

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*).

C'est de la validité de construct qu'il faut le plus se préoccuper lorsqu'on recourt aux QCM, car on risque souvent de s'éloigner du construct de départ.

Les deux autres types de validité suivants peuvent confirmer ou infirmer la validité de construct.

3. La validité concourante

L'épreuve aboutit-elle aux mêmes mesures qu'une autre épreuve réputée valide ? Une réponse nette à cette préoccupation est apportée par la corrélation entre les mesures obtenues par la nouvelle épreuve et celles obtenues par l'épreuve de « référence » (ou « critère ») sur les mêmes individus.

Cette approche présente une ambiguité grave. Si la corrélation est imparfaite (ce qui, en pratique, est toujours le cas), il se peut que les mesures de la nouvelle épreuve soient plus valides que celles de l'ancienne, et on est alors renvoyé aux autres types de validité.

Le problème de la validité concourante a tout son sens dans les cas où l'on veut aboutir aux mêmes mesures que celles que fournit une épreuve donnée, mais de façon plus économique : gain de temps, coût moindre, correction plus facile, etc. Les QCM apportent souvent une réponse à ces exigences.

4. La validité prédictive

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. La fidélité d'un test

Tout comme pour la validité, on ne s'interroge pas à proprement parler sur la fidélité d'un test, mais bien sur la *fidélité des mesures* obtenues à l'aide de ce test.

Le problème posé est ici « A quel point peut-on leur faire confiance ? Quelle est leur fiabilité ? ».

Autrement dit, ces mesures ne doivent-elles pas être affectées d'un cœfficient d'incertitude, d'erreur dû à l'instrument, aux procédés de mesure et à l'objet mesuré ?

1. Première source d'erreur : l'instrument

Imaginons que l'on constitue plusieurs tests parallèles sur une même matière et que l'on s'arrange pour leur donner une très grande validité de contenu et de construct. Ces tests devraient être équivalents et donner exactement les mêmes mesures sur les mêmes individus. En pratique, cela n'arrive jamais. De légères différences entre les questions amènent des différences de rendement.

La corrélation entre les mesures (sur les mêmes individus) obtenues à partir de tests parallèles apporte une réponse claire à la question de l'erreur de mesure due à l'instrument. Quand on ne dispose que d'un seul test, ce qui est le cas le plus fréquent, le problème est tout aussi présent, mais il est moins aisé de le résoudre. Diverses procédures de remplacement ont été imaginées : scinder le test en deux et considérer ces deux sous-tests comme des formes parallèles, ce qui permet de calculer la corrélation entre ces deux mesures, ou même considérer chaque question comme un test en soi et calculer les corrélations entre questions ou la corrélation entre la réussite à chaque question et le score au total du test.

2. Deuxième source d'erreur : le contexte d'administration

Un test présenté le lundi aurait-il donné les mêmes résultats s'il avait été présenté le vendredi ? Pour le savoir, il ne suffit pas de le répéter (retest) le vendredi, car lors de la première présentation, les élèves ont beaucoup appris ; ils ont pu aussi acquérir de nouvelles compétences durant l'intervalle de trois jours. En sciences humaines, il est donc quasi impossible de présenter deux fois un même test de connaissances aux mêmes individus dans les mêmes circonstances. C'est regrettable, car on aurait pu ainsi contrôler l'influence du local, de l'heure de la journée, de la sévérité des surveillants, éclairage, chaleur, bref toutes les variables qu'un retest dans

d'autres circonstances permettrait de neutraliser. Un tel retest permettrait de calculer les corrélations entre un test et lui-même. C'est pourquoi la fidélité est traditionnellement désignée par le symbole r_{tt} (r comme corrélation et t comme test).

La méthode de bipartition (découpage du test en deux parties) déjà évoquée ci-dessus à propos de la première source d'erreur ne constitue qu'un pis-aller permettant d'*estimer la fidélité.*

3. Troisième source d'erreur : l'individu

Même si les circonstances extérieures sont contrôlées, l'individu est sujet à des variations internes (excitation, fatigue, etc.). Ici également, on souhaiterait pouvoir répéter le test un autre jour, ce qui permettrait de mesurer une performance moyenne pour l'étudiant et de calculer (par corrélation) l'importance de l'erreur due aux variations des individus.

Pour les raisons déjà données, cette procédure est impossible et on sera le plus souvent amené à *estimer* la fidélité plutôt qu'à l'observer.

C. *Les formules d'estimation de la fidélité d'un test*

Rappelons que la fidélité d'un test est sa capacité de fournir des mesures (ici des classements) stables (reproductibles) si le test est appliqué de nouveau à un autre moment, ou dans d'autres circonstances.

Dans la section B du présent chapitre, nous avons signalé que la fidélité d'un test dépend
- de l'instrument (essentiellement les questions qu'il contient) ;
- du contexte d'administration du test (ou circonstances externes) ;
- de l'individu qui passe le test (ou variabilité interne du sujet).

Il s'agit là de trois *sources d'erreurs de mesure.* La fidélité serait parfaite (valeur 1) s'il n'y avait les erreurs de mesure !

Elle se définit mathématiquement par :

rtt = 1 – erreurs de mesure

L'indice de fidélité est symbolisé par rtt où r désigne la corrélation, t les scores obtenus au test et le second t les scores obtenus au *même test* administré une *autre fois,* mais *dans les mêmes circonstances.*

Les deux façons les plus courantes de mesurer la fidélité d'un test sont :
- la procédure test-retest ;
- les formes parallèles.

Aucune des deux ne correspond parfaitement à la définition « même test, mêmes circonstances ».

La procédure *test-retest* consiste à présenter le même test aux étudiants à deux moments différents. Pour obtenir l'indice rtt de fidélité du test, on calcule alors la corrélation (par la formule de Bravais-Pearson) entre les scores obtenus en ces deux occasions. Cette procédure ne remplit pas la condition « même circonstances » : la première administration a « enseigné » différentes choses aux élèves (sur la méthode de questionnement, sur les contenus) et l'intervalle entre les deux administrations a pu être mis à profit de façons diverses par les étudiants.

La procédure des *formes parallèles* consiste à créer deux tests équivalents comportant un même nombre de questions, chaque question ayant son équivalent (sa jumelle si possible) dans l'autre forme du test. Les deux formes du test sont administrées au même moment aux mêmes étudiants. Pour obtenir l'indice rtt de fidélité du test, on calcule alors la corrélation (par la formule de Bravais-Pearson) entre les scores obtenus à ces deux formes parallèles. La difficulté de créer des questions différentes – mais équivalentes – fait que cette procédure ne répond pas vraiment à l'exigence « mêmes tests ».

Faute de pouvoir mesurer la fidélité, on en est réduit à l'*estimer* par divers artifices, tous basés sur la *cohérence interne (internal consistency)* ou *homogénéité* de l'épreuve.

Ces méthodes exploitent trois formules de la corrélation de Bravais-Pearson.

La première méthode consiste à scinder le test en deux sous-tests (méthode *half split*) et à calculer la corrélation (de Bravais-Pearson) entre les mesures (X et Y) ainsi obtenues. Cette corrélation doit être « corrigée » (par la formule de Spearman-Brown).

La deuxième méthode se fonde sur les corrélations (notées rij) des questions entre elles et débouche sur les deux formules de Kuder et Richardson, appelées KR20 et KR21.

La troisième méthode exploite (dans la formule de Guilford) les corrélations (notées rit) entre chaque question et le test.

Toutes ces méthodes aboutissent à quantifier, en fait, non pas la fidélité, mais la *cohérence interne* du test.

Le tableau ci-après rappelle à quels problèmes les diverses procédures répondent (les hachures symbolisent l'absence de réponse).

Méthodes de calcul		Cette méthode tient-elle compte des SOURCES d'erreurs suivantes ?		
		Les questions du test	Variations dans l'administration	Variations chez l'individu
De la fidélité	Test-retest (corrélation de B.P.)		oui	oui
	Formes parallèles (corrélation de B.P.)	oui		
De la cohérence interne	*Half split* (corrélation de B.P. + correction de S.B.)	oui		
	KR20 ou KR21	oui		
	Formule de Guilford (corrélation rpbis)	oui		

Il ressort du tableau que le calcul de la fidélité, pour couvrir toutes les sources d'erreurs, doit se faire en combinant plusieurs méthodes. Par exemple, on peut construire trois formes parallèles (A, B et C) d'un même test, les formes A et B étant administrées le même jour et la forme C à un autre moment. La corrélation entre A et B mettra en évidence l'ampleur des erreurs dues au choix des questions ; la corrélation entre A et C l'ampleur des erreurs dues aux autres sources.

1. La correction de Spearman-Brown pour deux demi-tests

On considère la corrélation rxy entre les scores (X et Y) à deux sous-tests (A' et A") créés à partir du test original (A).

Afin d'assurer une longueur maximale et égale aux deux sous-tests, chacun d'entre eux contient souvent la moitié des questions. Cette méthode de coupure (*split*) d'un test en deux moitiés (*half*) est connue sous le nom anglo-saxon : *half split method.* C'est pourquoi on trouve souvent la notation rhh (h comme *half* au lieu de rXY) pour rappeler qu'il s'agit d'une *corrélation entre les deux moitiés* d'un test. Cette corrélation (rXY ou rhh) est calculée par la formule de Bravais-Pearson (voir ci-avant).

Souvent, on constitue les deux sous-tests (A' et A") en groupant les questions paires dans l'un et les questions impaires dans l'autre, d'où le nom anglais *odd-even* (pair-impair).

On suppose que si les parties du test (h et h) donnent des résultats semblables, (rhh = 1), le test total donne des résultats d'une grande stabilité.

Mais la corrélation rhh (c'est-à-dire rXY) ne peut que sous-estimer la fidélité, car le nombre de question est diminué de moitié dans chaque sous-test. On doit donc *rectifier* le cœfficient de corrélation (rhh) par la formule de correction de Spearman et Brown :

$$rtt_{(SB)} = \frac{2\ rhh}{1 + r\ hh}$$

Le symbole SB entre parenthèse rappelle les noms de Spearman et Brown.

Pour notre test A, (où rhh = 1), nous aurions rtt (SB) = $\frac{2\ .\ 1}{1 + 1} = \frac{2}{2} = 1$

2. Les formules de Kuder-Richardson

On considère les n x n cœfficients de corrélation rij (calculés par l'indice φ) entre les questions d'un test prises deux à deux. On peut utiliser leur moyenne mrij comme estimation de la fidélité.

Bien qu'il suffise de calculer $\frac{n\ (n\ -\ 1)}{2}$ corrélations pour connaître l'ensemble des n × n nécessaires, les calculs restent fastidieux (45 corrélations pour 10 questions).

Or, mrij peut être calculé d'une autre façon, proposée par Kuder et Richardson dans des formules célèbres.

a) La KR20

Leur formule 20 (connue sous le nom de KR20) requiert le calcul de
- la variance (σ^2) des scores au total de l'épreuve ;
- la fréquence de réponses correctes (p) pour chaque question ;
- la fréquence d'échecs (q) pour chaque question.

La formule est :

$$KR20 = \frac{NQ}{NQ\text{-}1}\ \frac{(\sigma^2 - \Sigma\ (p\ .\ q))}{\sigma^2}$$

Considérons un test A,

		1	2	3	4	5	6	Questions
Test A	1	0	0	0	0	0	0	
	2	0	0	0	0	0	0	
	3	0	0	0	0	0	0	
	4	0	0	0	0	0	0	
	5	1	1	1	1	1	1	
	6	1	1	1	1	1	1	
Etudiants	7	1	1	1	1	1	1	
	8	1	1	1	1	1	1	

NQ = 6 (nombre de questions)
σ^2 = 9 (variance)
$\Sigma\ (p\ .\ q)$ = 1,5

On a donc[1] : KR20 = $\frac{6}{5}\ \frac{(9\ -\ 1{,}5)}{9} = \frac{6}{5}\ .\ \frac{7{,}5}{9} = 1$

Dans le test ci-après B, NQ = 6, σ^2 = 3,9375, Σpq = 0,656 et KR20 = 1.

[1] En toute rigueur, il n'est pas permis de calculer les corrélations (à la base de l'estimation de la fidélité), ni les formules KR20 et KR21 à partir d'un aussi petit nombre de sujets. Les valeurs des indices de fidélité sont donc purement indicatives.

b) La KR21

La formule KR21 est d'un maniement plus aisé que la KR20, mais exige que toutes les questions soient approximativement de la même difficulté. Il faut que l'on calcule $\overline{pq}$, c'est-à-dire la moyenne de ces NQ produits. Or $\overline{pq} = \overline{p}.\overline{q}$. Dans le test A, $\overline{p} = 0{,}5$; $\overline{q}$ vaut donc 0,5 et M (p.q) vaut 0,25.

$$KR21 = \frac{NQ}{NQ-1}\left(\frac{(\sigma^2 - (NQ \cdot (\overline{p}.\overline{q})}{\sigma^2}\right)$$

Dans le test A, NQ = 6, $\sigma^2 = 9$, NQ. $\overline{pq} = 6 \times 0{,}25 = 1{,}5$.

On a donc[1] : $KR21 = \frac{6}{5} \cdot \frac{(9 - 1{,}5)}{9} = 1$

Le résultat est donc, ici, identique pour la KR20 et la KR21. Il l'est aussi pour la matrice B :

		1	2	3	4	5	6	Questions
Test B	1	0	0	0	0	0	0	
	2	0	0	0	0	0	0	
	3	0	0	0	0	0	0	
	4	0	0	0	0	0	0	
	5	1	1	1	1	1	1	
	6	0	0	0	0	0	0	
Etudiants	7	0	0	0	0	0	0	
	8	0	0	0	0	0	0	

Ce n'est cependant plus le cas pour les matrices plus complexes : les deux formules donnent des résultats légèrement différents.

3. La formule de GUILFORD basée sur les rpbis

On considère les NQ corrélations bisériales de point entre les items et le test (rit). On calcule la moyenne mrit de ces corrélations.

La formule proposée par GUILFORD (1956, p. 463) pour estimer la fidélité (rtt) est la suivante :

$$rtt_{(G)} = \frac{NQ(mrit)^2}{1 + (NQ-1)(mrit)^2}$$

Le symbole G entre parenthèses rappelle le nom de GUILFORD

Cette formule vise à compenser le fait que la question est NQ fois plus courte que le test et que mrit tend à sous-estimer la fidélité.

Pour le test A, nous aurions (voir note à la p. précédente)

$$rtt = \frac{6\,(1)^2}{1 + (5)\,(1)^2} = \frac{6}{6} = 1$$

4. Fidélité et nombre de questions

On peut prévoir la fidélité qui résultera de l'*allongement d'un test*[1].

Si l'on désigne par rll la fidélité du test de départ (longueur l) et par rnn la fidélité du test allongé n fois (cœfficient d'allongement = n), alors,

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

(GUILFORD et FRUCHTER, 1978, p. 426).

On reconnait dans cette formule générale celle qui a été appliquée auparavant à $(mrit)^2$ et Spearman-Brown.

[1] A condition que les questions ajoutées soient parallèles aux questions de départ (c'est-à-dire aient le même indice de difficulté).

D. Le nombre optimal de questions et de solutions

1. Validité et nombre de questions.

On peut se poser la question de la façon inverse : quel doit être le cœfficient d'allongement n du test pour atteindre une fidélité donnée (par exemple 0,80 ou 0,90) ? On y répond par la formule ci-dessous (Guilford et Fruchter, 1978, p. 432) :

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

Les formules correspondantes pour la validité[2] sont les suivantes. La corrélation entre un critère (désignons-le par y) et un test x allongé a fois se note ry (ax) ; 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).

2. Nombre de solutions proposées et fidélité

En se basant sur cinq principes théoriques, EBEL (1969, p. 566) montre qu'un nombre élevé de solutions est de nature à améliorer la fidélité des scores au test.

Pour estimer cette fidélité (rtt), il propose la formule suivante ; qui, on le remarquera, est indépendante des caractéristiques psychométriques (indice de difficulté) des questions posées :

$$rtt = \frac{NQ}{NQ-1}\left[1 - \frac{9\ (k+1)}{NQ\ (k-1)}\right]$$

où k = nombre de solutions
NQ = nombre de QCM

[2] 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 cœfficient de validité ne peut excéder le cœfficient de fidélité : on ne peut imaginer un test mieux corrélé avec un autre test qu'avec une forme parallèle de lui-même.

Pour un test de 100 questions, les fidélités calculées sont :

k	*rtt*
2	0,74
3	0,83
4	0,86
5	0,874
6	0,883

EBEL calcule comme suit le nombre de QCM nécessaires pour obtenir une fidélité donnée :

$$NQ = \frac{9}{1\text{-}rtt}\left[\frac{k+1}{k-1}\right]$$

où rtt = fidélité attendue

Ce qui donne, pour rtt = 0,90 :

Nombre de solutions	*Nombre de questions nécessaires*
2	270
3	180
4	150
5	135
6	126

Le tableau qui suit fournit, calculés selon cette formule d'Ebel, les nombres de questions à choix multiple (QCM) nécessaires pour obtenir une fidélité donnée, quand toutes les QCM de l'épreuve ont le même nombre (k) de solutions proposées.

FIDELITE DU TEST

NOMBRE DE SOLUTIONS PROPOSEES (k) PAR QCM	0,50	0,55	0,60	0,65	0,70	0,75	0,80	0,85	0,90	0,95
2	54	60	67	77	90	108	135	180	270	540
3	36	40	45	51	60	72	90	120	180	360
4	30	33	37	43	50	60	75	100	150	300
5	27	30	34	38	45	54	67	90	135	270
6	25	28	31	36	42	50	63	84	126	252
7	24	27	30	34	40	48	60	80	120	240
8	23	26	29	33	39	46	58	77	115	231
9	22	25	28	32	37	45	56	75	112	225
10	22	24	27	31	36	44	55	73	110	220

3. Nombre de solutions, nombre de questions et fidélité

Lorsque la réponse consiste à choisir *une* des solutions présentées, il est clair que plus la question présente de solutions, plus il faut de temps pour la lire et donc pour y répondre.

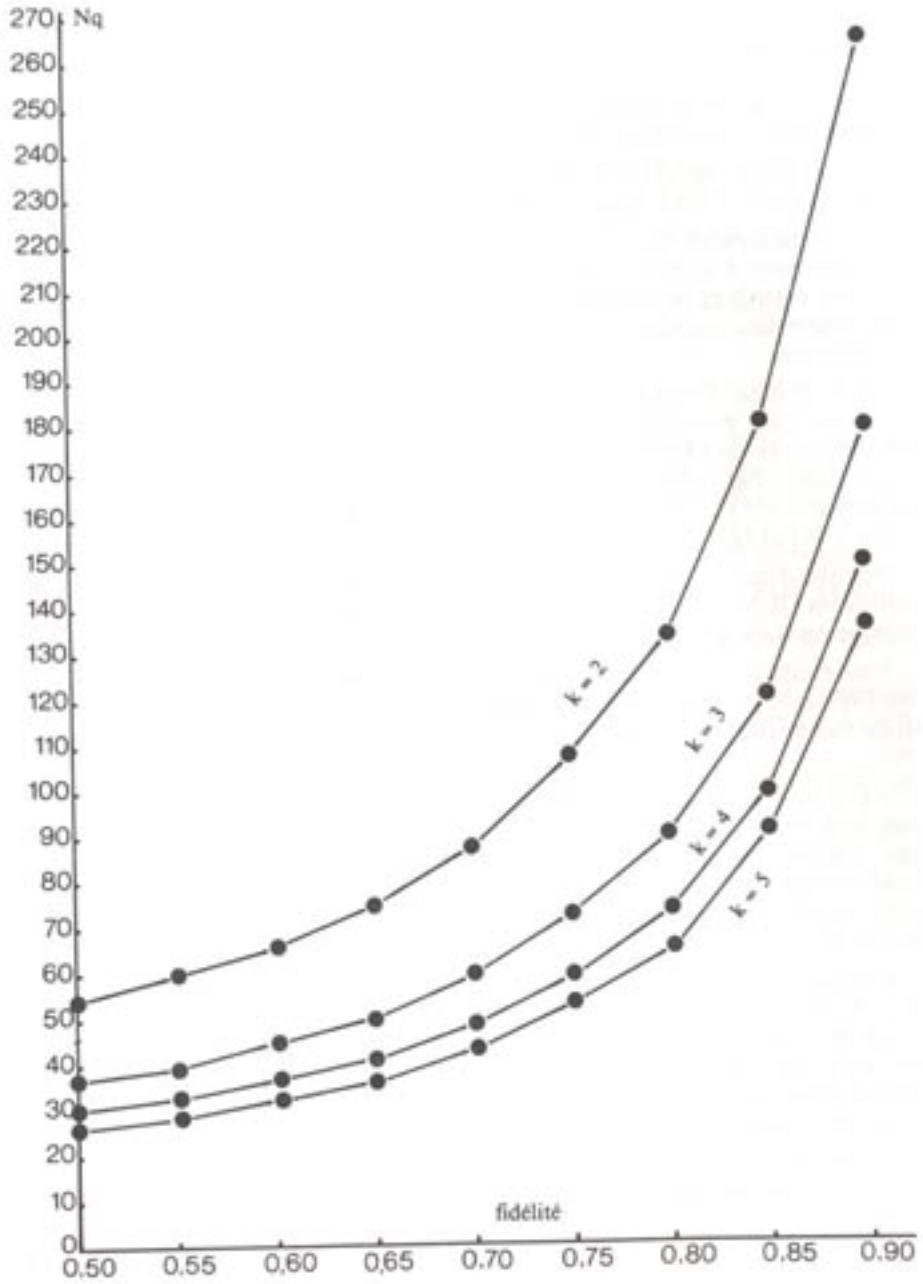
Donc, dans une unité de temps donnée (une heure, par exemple) on pourra obtenir beaucoup plus de réponses à des questions vrai-faux qu'à des QCM à cinq solutions proposées. Si la vitesse de réponse est *double* pour une question vrai-faux que pour une QCM à cinq solutions, alors le recours à l'un ou l'autre de ces deux types de questions a le même impact sur la fidélité des scores au test, comme le montrent les exemples ci-dessous.

Ainsi, imaginons qu'en une heure, on puisse répondre à 60 questions vrai-faux, tandis qu'on ne peut répondre qu'à 30 QCM à 5 solutions proposées. Si on consulte le tableau ci-devant, on constate que dans les deux cas, la fidélité prévue par la formule d'Ebel vaut 0,55.

De même, au bout de trois heures, on pourrait répondre à 180 questions vrai-faux, à 90 QCM à 5 solutions proposées et, dans les deux cas, la fidélité prévue par la formule d'Ebel vaut 0,85.

Le graphique ci-dessous permet de lire, pour k (le nombre de solutions proposées) valant 2, 3, 4 ou 5, les nombres de QCM nécessaires pour obtenir (selon la formule d'Ebel), un cœfficient de fidélité donné (compris entre 0,50 et 0,90).

Ce graphique peut aussi être lu autrement : étant donné k et un nombre de QCM, quelle est, toujours selon la formule d'Ebel, la fidélité prévisible ?



4. Discussion

Dans une unité de temps donnée, plus les solutions proposées sont nombreuses, moins le nombre de questions pourra être élevé.

Or, ces deux variables (NQ et k) interviennent toutes deux dans la première formule d'Ebel. Quelle solution faut-il dès lors préférer ?

Ferme défenseur des questions vrai-faux, Ebel prétend qu'un étudiant peut répondre à une telle question en moins de temps que pour une QCM à quatre solutions proposées. Il prétend aussi qu'un enseignant peut créer deux questions vrai-faux dans le temps nécessaire à créer une QCM à quatre solutions.

De leur côté, Oosterhof et Glanapp (1974) ont prouvé expérimentalement qu'il est nécessaire d'utiliser entre 1,5 et 4,5 fois plus de questions vrai-faux que de QCM (avec K=4) pour produire une épreuve de fidélité équivalente. Par ailleurs, Frisbie (1973) trouve, en ce qui concerne le temps de réponse de l'étudiant, un rapport de 1,5 (et non de 2) entre la QCM (avec k=4) et la question vrai-faux.

Partant d'autres principes qu'Ebel, Tversky calcule que, dans le cas où toutes les QCM d'un test comportent le même nombre de solutions proposées, le recours à *trois solutions* assure au test une fidélité maximale.

Une étude de Costin (1970) corrobore, sur certains points, la conclusion de Tversky : les tests à trois solutions se sont avérés plus « puissants » et discriminatifs que leur version à quatre solutions. La conclusion de Costin est :

« Les professeurs choisissent souvent quatre solutions sans doute parce que ce nombre apparaît comme un compromis. Trois solutions paraissent peu fiables et cinq s'avèrent difficiles à trouver dans maints cas. S'il se confirme qu'avec trois solutions, on augmente la puissance et la discrimination, les enseignants gagneront du temps en rédaction et les étudiants du temps de réponse ».

Ramos et Stern (1973) ont comparé des QCM à quatre solutions avec des QCM à cinq solutions ramenées à quatre solutions (par élimination de la solution la moins *attractive*). Ces deux types de questions donnent des résultats comparables, avec une légère chute de la fidélité dans le second type d'épreuve. Les auteurs suggèrent de répéter l'expérience en éliminant cette fois le distracteur le moins *discriminatif*.

En résumé, on ne dispose pas actuellement de réponse ferme au problème du nombre optimal de solutions proposées.

E. L'entraînement des étudiants

Avoir fait, à plusieurs reprises, l'expérience d'un certain type d'épreuve est un atout important pour un étudiant. Le professeur veillera donc à ménager cette possibilité à tous ses étudiants.

Des conseils peuvent être utiles. Les suivants, inspirés d'Ebel (1979, p. 180) concernent particulièrement les QCM.

1° Lisez (puis écoutez) les instructions avec attention et respectez-les.
2° Repérez les questions qui (éventuellement) ont plus de poids que d'autres.
3° Examinez le barème des tarifs. Quelle est la conséquence d'une réponse incorrecte et d'omissions ?
4° Quand le barème G (*correction for guessing* classique) et d'application rappelez-vous qu'une réponse au hasard a le même score attendu que l'omission, c'est-à-dire zéro. Quand vous avez la moindre information permettant de privilégier une réponse, utilisez-la et répondez.
5° Que le correcteur soit le professeur ou une machine, la lisibilité et la netteté sont toujours appréciées : prenez donc la peine de bien écrire vos réponses.
6° Ménagez-vous du temps pour *envisager* toutes les questions. Ne passez pas trop de temps sur une au détriment des autres.
7° Si vous devez répondre sur une feuille séparée, vérifiez fréquemment que les codes de la feuille de réponses correspondent bien à ceux de la feuille de tests.
8° Réservez du temps pour relire vos réponses et corriger les fautes d'inattention.

F. La fraude et sa prévention

1. Les diverses formes

La fraude à une épreuve peut prendre des formes multiples. Ebel (1979, p. 185) énumère les suivantes :

1° Le coup d'œil sur la feuille du voisin.
2° La collusion entre deux étudiants ou plus pour échanger des informations durant le test.
3° La préparation et l'utilisation d'un support illicite de la mémoire.
4° La copie illicite ou le vol d'une question, d'une page ou d'un fascicule de test.
5° L'achat de questions d'examen.
6° Le remplacement illicite d'un individu par un autre (substitution).

Les QCM facilitent le coup d'œil sur la feuille du voisin, mais le grand nombre de questions rend néanmoins la fraude malaisée.

2. Procédures contre la fraude

Procédure 1 : L'isolement des étudiants.

Isoler les étudiants, acoustiquement et visuellement, est la solution la plus radicale, mais elle est souvent impossible à réaliser.

Procédure 2 : Distribution fractionnée.

Le professeur divise l'épreuve en plusieurs sous-épreuves. Par exemple, on peut diviser un test de 30 questions en trois parties de 10 questions. Chaque étudiant reçoit un tiers différent de celui de ses voisins et rend ses réponses (et la sous-épreuve) après le tiers du temps. Une seconde sous-épreuve lui est alors remise, puis la troisième. Cette procédure est lourde et demande un plan de distribution à respecter soigneusement.

Procédure 3 : Formes parallèles composées de questions différentes.

Admettons que l'on crée trois épreuves parallèles A, B et C, comportant des questions différentes, mais jugées équivalentes. Si dans une classe donnée, on craint la fraude entre élèves voisins, on distribuera les formes dans l'ordre A B C, ABC, ABC, ABC, ABC... et non dans un ordre aléatoire (comme ABBCCAAABCAABCCCBB, cas où les voisins disposant de la même forme seront probablement nombreux).

Cette procédure est efficace, mais par définition, empêche de poser les mêmes questions à tous. C'est pourquoi on recourt souvent à diverses *formes parallèles comportant les mêmes questions*, mais dont l'ordre (et/ou l'ordre des solutions) sont différents. On part d'une forme originale, puis on brouille l'ordre des questions pour les autres formes. Les procédures 4 à 9 décrivent ces façons de faire. Pour la commodité de la correction, tout système de brouillage de l'ordre des questions doit être accompagné d'un système correspondant de remise en ordre.

Procédure 4 : Brouillage aléatoire de l'ordre des questions.

Cette procédure dissuade efficacement la fraude, mais alourdit considérablement la remise en ordre ultérieure. Il faut, en effet, établir des tables de correspondance, pour les diverses épreuves parallèles.

Exemple de quinze questions constituant trois formes parallèles.

Exemple de quinze questions constituant trois formes parallèles.

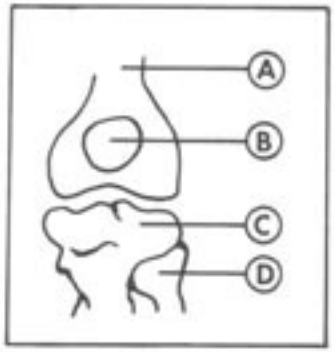
Forme	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Forme B	11	14	3	15	1	8	12	5	2	10	9	4	7	13	6
Force C	12	5	11	13	4	10	1	6	9	2	14	7	15	3	8

Même lorsque l'on dispose d'un programme d'ordinateur pour remettre les réponses en ordre, cette procédure est assez lourde et les risques d'erreurs de transposition sont grands. Elle exige, au préalable, la dactylographie de chacune des différentes formes ou différents montages par photocopie.

Le brouillage aléatoire des questions pour l'ensemble de l'épreuve rend la remise en ordre lourde et peu fiable. Une correction séparée pour chaque forme est donc recommandée.

Procédure 5 : Brouillage aléatoire des solutions proposées.

Il arrive que l'on soit contraint de poser à tous les étudiants une même question en même temps, par exemple à propos d'une *diapositive* projetée à l'ensemble de la classe. On peut alors utiliser des formes de tests où les questions varient non plus par leur ordre, mais par *celui des solutions proposées*. Le numéro de la réponse correcte diffère donc d'une forme à l'autre.



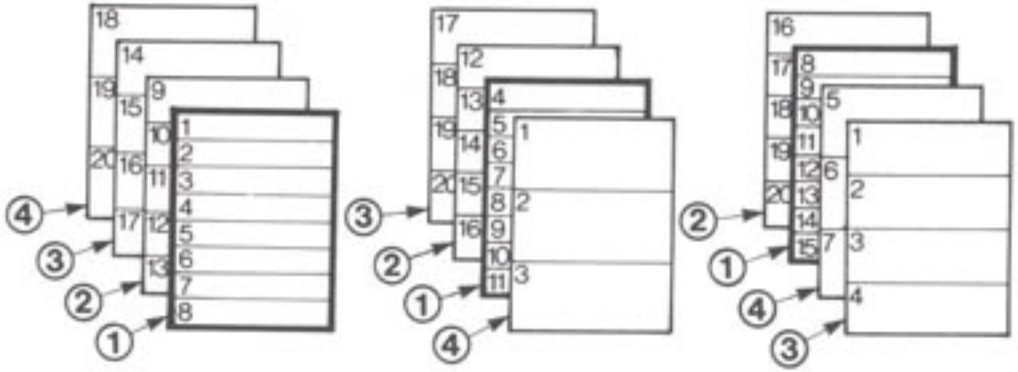
Forme A	*Forme B*
Le fémur est l'os 1. A 2. B 3. C 4. D	Le fémur est l'os 1. C 2. D 3. A 4. B
De quel genou s'agit-il ? 1. Du droit. 2. Du gauche. 3. On ne saurait le dire.	De quel genou s'agit-il ? 1. On ne saurait le dire. 2. Du gauche. 3. Du droit.
Le ligament rotulien va de : 1. L'os A à l'os B. 2. L'os A à l'os C. 3. L'os B à l'os C. 4. L'os B à l'os D.	Le ligament rotulien va de : 1. L'os B à l'os C. 2. L'os A à l'os B. 3. L'os B à l'os D. 4. L'os A à l'os C.

Le brouillage aléatoire des solutions pour l'ensemble de l'épreuve rend elle aussi la remise en ordre lourde et peu fiable. Une correction séparée pour chaque forme est donc ici aussi recommandée.

Procédure 6 : Brouillage aléatoire de l'ordre des questions et des solutions.

Ce système est forcément encore plus lourd que les précédents.

Procédure 7 : Décalage de l'ordre des questions.



Forme A Forme B Forme C

Le décalage vers l'avant consiste à placer la dernière feuille (ici 4) en première position dans le paquet.

Le débrouillage (ou remise en ordre) est basé sur l'ampleur du décalage des numéros des questions par rapport aux numéros de la forme de départ (forme A). Dans l'exemple ci-dessus, la forme B est décalée de trois questions et la forme C de sept questions par rapport à la forme A.

Si N est le numéro d'une question dans la forme B, N-3 est le numéro de cette question dans la forme originale A... sauf pour les trois premières questions, auxquelles on applique l'expression NQ + (N – 3), NQ étant le nombre de questions du test.

Procédure 8 : Décalage de l'ordre des solutions d'un nombre NDS.

Comme pour la procédure 7, il ne s'agit pas d'un brouillage aléatoire, mais d'une rotation respectant la séquence originale entre les solutions, mais en décalant la lecture de cette séquence. L'ordre des questions est le même d'une forme à l'autre. L'exemple (diapositive du genou) proposé pour la procédure 5 présente, pour les questions 1 et 2 un décalage de deux solutions vers l'avant.

La remise en ordre est délicate.

Procédure 9 : Décalage de l'ordre des questions et de l'ordre des solutions.

Une telle procédure fournit les épreuves presque aussi dissemblables que la procédure 6, tout en permettant cependant le brouillage et le débrouillage par ordinateur, ce qui représente un avantage appréciable.

Règles générales concernant les brouillages (procédures 4 à 9)

Tout brouillage doit être annoncé de façon à décourager la fraude :

> Vous avez tous reçu les mêmes questions. Cependant, elles ne vous sont pas présentées dans le même ordre. De même, les solutions proposées figurent à des places différentes. Ne vous fiez pas à la ressemblance des textes. Les numéros, eux, sont brouillés ! Ne copiez pas sur votre voisin : une solution correcte pour lui est incorrecte pour vous.

On remarque que cette consigne ne précise ni le type de brouillage pratiqué, ni le nombre de formes parallèles, ni l'ordre de distribution des formes de l'épreuve.

3. Détection *a posteriori* de la fraude

a) En l'absence de brouillage

A l'aide d'un programme (FORTRAN) de correction de QCM posées à des étudiants en médecine, Vorkauf (1977) établit les taux d'*erreurs identiques* pour tous les étudiants pris deux à deux. Ces cœfficients sont interprétés grâce au plan indiquant le numéro du siège occupé par chaque étudiant dans la salle d'examen.

b) Après brouillage (formes parallèles)

Le programme de Smets (1976) calcule les scores de chaque étudiant en médecine selon les diverses formes parallèles. Si un étudiant obtient plus de points avec une autre forme que la sienne, il est suspecté de fraude. Ici, aussi, on consulte le plan d'occupation de la salle par les étudiants.

c) Conclusions

Le recours à des procédures de détection *a posteriori* de la fraude ne peut, à notre avis, être caché aux étudiants. Ceux-ci doivent en être avertis avant l'épreuve.

Certains voient dans les modifications apportées aux réponses (barrage et remplacement) un signe de tricherie possible et ont envisagé d'interdire ces corrections. On trouvera dans WOOD (1977, p. 234) une discussion de cette question. WOOD, qui se base sur les travaux de Pippert (1966), Copeland (1972), Foote et Belinsky (1972), Relling et Taylor (1972), Jacobs (1974), Pascale (1974) et Lynch et Smith (1975), concut qu'il est préférable de permettre aux étudiants de modifier leurs réponses en cours d'épreuve.

CONCLUSIONS

Dans nos sociétés modernes, l'homme est amené à APPRENDRE de plus en plus SOUVENT, VITE et BIEN. De sa capacité d'apprendre du nouveau dépend la prospérité de son entreprise, voire de son pays. Pour être compétitif économiquement, un pays aujourd'hui ne peut se reposer sur une élite technique. C'est l'ENSEMBLE DE LA POPULATION qui doit être capable de faire cette démarche d'adaptation et d'innovation constante.

Or l'apprentissage est un processus régulé où l'EVALUATION joue un rôle crucial. C'est l'évaluation qui permet de détecter objectivement les lacunes ou le degré de maîtrise de concepts, de performances. Ces MESURES sont alors comparées à des CRITERES pour décider d'opérations diverses (apprendre encore, changer les méthodes, changer les objectfs, etc.), bref pour mettre en œuvre des BOUCLES DE REGULATION.

Il est donc capital que les FORMATEURS, dans les écoles ou dans les entreprises, manipulent bien et comprennent en profondeur les instruments de cette évaluation de plus en plus nécessaire, qualitativement et quantitativement. Cela fait partie de leur PROFESSIONALITE de maîtriser la signification des SYSTEMES DE MESURE en édumétrie, de savoir de quoi dépend la validité de mesures, de savoir les interpréter, d'en connaître les limites.

Certains lecteurs seront rebutés par le caractère théorique de certaines démonstrations mathématiques. Il suffit qu'ils en retiennent les conclusions et qu'ils en comprennent la problématique et les concepts.

INDEX

BIBLIOGRAPHIE

BAKER, F.B. & MARTIN, T.J. (1969), FORTAP : A Fortran test analysis package, *Educ. Psychol. Measmt.*, 29, 159-164.

BRICKER, P. et CHAPANIS, A. (1953), Do incorrectly percieved tachistoscopic stimuli convey some information ? *Psychological Review*, 60, 181-188.

CARDINET, J. (1972), L'adaptation des tests aux finalités de l'évaluation. Exposé au 3e Congrès de l'AIPELF à Bruxelles, Document IRDP n° 7208, Neuchâtel.

CARNIDET, J. et TOURNEUR, Y. (1975), The generalization of surveys of educational outcomes, Second International Symposium on Educ. Testing, Montreux.

CARVER, R.P. (1974), Two dimensions of tests : Psychometric and edumetric, *Amer. Psychol.*, 29, 512-518.

CHERNOFF, H. (1962), The scoring of multiple choice questionnaire, *Annals of Mathematical Statistics*, 33, 375-393.

CHOPPIN, B.H. (1974[a]), *The Correction for Guessing on Objective Tests*, IEA Monograph Studies, n° 4, Stockholm.

CHOPPIN, B.H. (1975), Guessing the answer on objective tests, *Brit. Jour. Educ. Psychol.*, 45, 206-213.

COSTING, F. (1970), The optional number of alternatives in multiple-choice achievement tests : Some empirical evidence for a mathematical proof, *Educ. Psychol. Measmt.*, 30, 353-358.

CREHAN, K.D. (1974), Item analysis for teacher-made mastery tests, *Jour. Educ. Measmt.*, 11, 255-262.

CRONBACH, L., GLESER, G., NANDA, H., RAJARATNAM, N. (1972), *The Dependability of Behavioral Measurements*, New York, Winey.

CRONBACH, L. (1952), Coefficient alpha and the internal structure of tests, *Psychometrika*, vol. 16, 297-334.

CROSS, L. & FRARY (1977), An empirical test of Lord's theoretical results regarding formula scoring of multiple choice tests, *Jour. of Ed. meast.*, vol. 14, 313-321.

CURETON, E. (1966), The correction for guessing, *Journ. of Experim. Educ.*, vol. 24, 44-47.

D'AGOSTINO, R.B. & CURETON, E.E. (1975), The 27 percent rule revisited, *Ed. Psychol. Measmt.*, 25, 47-50.

DAS GUPTA, S. (1960), Point biserial correlation coefficient and its generalization, *Psychometrika*, 25, 393-408.

DAVIS, F.B. (1946), Analyse des items, Ed. Nauwelaerts, Louvain, 1966.

DAVIS, F.B. (1952), Item analysis in relation to educational and psychological testing, *Psychol. bull.*, vol. 49, 97-119.

DAVIS, F.B. (1967), A note on the correction for chance success, *Journ. of Exper. Educ.*, vol. 35, 42-47.

DE FINETTI, B. (1965), Methods for discriminating levels of partial knowledge concerning a test item, *Brit. Journ. of Mathem. and Statist. Psych.*, 18, 87-123.

DE LANDSHEERE, G. (1980), L'Association Internationale pour l'Evaluation du Rendement Scolaire, Recherches en cours, *Revue* de la Direction générale de l'Organisation des Etudes du Ministère de l'Education nationale, 1980, n° 8, 11-21.

D'HAINAUT, L. (1973), Etude d'une nouvelle variable pour l'analyse statistique des expériences pédagogiques, *Bulletin de Psychologie*, 305, vol. XXVI, 622-630.

DUNN, T.F. & GOLDSTEIN, L.G. (1959), Test difficulty, validity and reliability as functions of selected multiple-choice item construction principles, *Educ. Psychol. Measmt.*, 19, 171-179.

EBEL, R.L. (1965), Confidence-Weighting and Test Reliability, *Journ. of Educ. Measmt.*, 2, 49-57 B.

EBEL, R.L. (1968), Blind guessing on objective achievement tests, *Journ. of Educ. Measmt.*, vol. 5, 321-325.

EBEL, R.L. (1969), Expected reliability as a function of choices per item, *Educ. Psychol. Measmt.*, 29, 565-570.

FABRE, J.M. (1980), *Jugement et certitude*, Ed. Peter Lang, Berne.

FLANAGAN, J.C. (1952), The effectiveness of short methods for calculating correlation coefficients, *Psychol. Bull.*, 1952.

FOOTE, R. & BELLINKS, C. (1972), It pays to switch ? Consequences of changing answers on mutiple-choice examinations, *Psychol. Reps.*, 31, 667-673.

GARDNER, P. (1970), Test length and the standard error of measurement, *Journal of Educ. Measmt.*, vol. 7, 271-273.

GLASER, R. & NITKO, A.J. (1971), Measurement in learning and instruction, in THORNDIKE, R.L. (Ed.), *Educational Measurement*, American Council on Education, Washington.

A. GRISAY, Rendement de l'enseignement de la langue maternelle en Belgique francophone, Recherche en Education, n° 4, Direction générale de l'Organisation des Etudes, 1974.

GUILFORD, J.P. (1954), *Psychometric Methods*, New York, Mc Graw-Hill.

GUILFORD, J.P. et FRUCHTER, (1978), *Fundamental Statistics in Psychology and Education*, New York, Mc Graw Hill.

HALES, L.W. (1972), Method of obtaining the index of discrimination for item selection and selected test characteristics : A comparative study, *Educ. Psychol. Measmt.*, 32, 929-937.

HAMBLETON, R., ROBERTS, D. and TRAUB, R. (1970), A comparison of the reliability and validy of two methods for assessing partial knowledge on a multiple-choice test, *Journal of Educational Measurement*, n° 7, 75-82.

HAMILTON, C. (1950), Bias and error in multiple choice, *Psychometrika*, 15, 151-168.

HAMMERTON, M. (1965), The guessing correction in vocabulary tests, *Brit. Journ. of Educ. Psych.*, 35, 249-251.

HARDY, J.L. (1981), Adjustment of choice tests correlation in multiple choice item analysis, Série LPE, 81-07-001, 17 pages.

HARDY, J.L. (1983), Plusieurs solutions au problème du recouvrement dans la corrélation entre un test et un item dichotomique, in *Scientia Paedagogica Experimentalis*, 49-71.

HENRY, G. (1974), Rendement de l'enseignement des sciences en Belgique francophone, Recherche en Education, n° 8, Direction générale de l'Organisation des Etudes, Bruxelles.

HENRYSSON, S. (1963), Correction of item-total correlations in item analysis, *Psychometrika*, vol. 28, 211-218.

HENRYSSON, S. & WEDMAN, I. (1974), Some problems in construction and evaluation of criterion-referenced tests, *Scand. Jour. Educ. Res.*, 18, 1-12.

HEVNER, K. (1932), A method of correcting for guessing in true-false and empirical evidence in support of it, *Journal of Soc. Psychol.*, 3, 359-362.

HIVELY, W., PATTERSON, H.L. & PAGES, S.H. (1968), A « universe-defined » system of arithmetic achievement tests, *Jour. Educ. Measmt.*, 5, 275-290.

HOFFMAN, R.J. (1975), The concept of efficiency in item analysis, *Educ. Psychol. Measmt.*, 35, 621-640.

HORST, P. (1933), The difficulty of a multiple-choice test item, *Journ. of Educ. Psych.*, 24, 229-232.

IVENS, S.H. (1971), Nonparametric item evaluation index, *Educ. Psychol. Measmt.*, 31, 843-849.

JACOBS, S. (1971), Correlates of unwarranted confidence in responses to objective test items, in *Journal of Educational Measurement*, vol. 8, n° 1, 15-20.

JASPEN, N. (1965), Polyserial correlation programs in FORTRAN, *Educ. Psychol. Measmt.*, vol. 25, 229-233.

JOHNSON, R. & ROSENTHAL, E. (1968), Influence of guessing on measurements of immediate and delayed attention, *Journal of Education Measurement*, vol. 5, n° 2, 175-181.

KELLEY, T.L. (1969), The selection of upper and lower groups for the validation of test items, Jour. Educ. Psychol., 30, 17.

KOEHLER, R.A. (1974), Over confidence on probabilistic tests, *Jour. Educ. Measmt.*, 11, 101-108.

KUDER, G. & RICHARDSON, M. (1937), The theory of the estimation of test reliability, Psychometrika, vol. 2, 151-160.

LAPLANCHE, G. (1977), Guide technique d'utilisation du STEP, Université de Liège, 2 volumes.

LECLERCQ, D. (1973), Critique des méthodes d'application, de correction et de cotation des questions à choix multiple, *Scientia Paedagogica Experimentalis*, X, 1.

LECLERCQ, D. (1975), Banque de questions et indices et certitude, options docimologiques adaptées à l'enseignement secondaire, Education, 149-150.

LECLERCQ, D. (1977), L Matrices or the computation of consequences for confidence marking procedures in educational settings ; rationale, algorithm and FORTRAN program, paper presented at the *6th Research conference on subjective probability, Utility and decision Making*, Waarszawa.

LECLERCQ, D. (1978), L'auto-évaluation des compétences dans le domaine cognitif, in Revue, 13e année, n° 2, 3-20.

LECLERCQ, D. et PEREE, F. (1979), Une expérience de dossier automatisé d'étudiant, Communication à la *4e conférence de l'A.T.E.E.*, Pont-à-Mousson (France).

LECLERCQ, D. (1980), Sequential adaptive tailored testing and confidence marking, p. 306, in VANDERKAMP, LANGERAK & DE GRUIJTER, *Psychometics for Educational Debates*, Proceedings of the third Symposium on Educational Testing (Leyden, July 1977), Wiley and Sons, New York.

LECLERCQ, D. (1980), Computerised tailored testing : structured and calibred item banks for summative and formative evaluation in European Journal of Education, vol. 15, n° 3, 251-260.

LECLERCQ, D. (1983), Confidence marking, its use in testing, 106 p. in POSTLETHWAITE & CHOPPIN, Evaluation in Education, vol. 6, 161-287, Oxford : Pergamon Press.

LECLERCQ, D. (1986), La conception des questions à choix multiple, Bruxelles : Labor.

LECLERCQ, D. (1987), Auto-évaluation et connaissance partielle, Bruxelles : Labor.

LORD, F. (1975), Formula scoring and number-right scoring, *Journ. of Educ. Measmt.*, vol. 12, 7-11.

LORD, F & NOVICK, M. (1968), *Statistical theories of Mental Test Scores*, Reading (Mass.) Addison-Wesley.

LORET, M.T. (1980), Rendement de l'enseignement de l'anglais en Belgique francophone, Recherche en Education, N° 15, Direction générale de l'Organisation des Etudes.

LUMINGU, B. (1974), Etude préalable à la construction d'un test diagnostique sur la consultation du dictionnaire, Université de Liège, Mémoire de licence ronéotypé.

LYNCH, D.C. & SMITH, B.C. (1975), Item response changes : Effects on test scores, *Meas. Eval. in Guidance*, 7, 220-224.

MASSOZ, D. et HENRY, G. (1979), Connaissances et attitudes sociapolitiques d'élèves de l'enseignevent secondaire, Recherche en Education, n° 17, Direction générale de l'Organisation des Etudes.

MATALON, B. (1975), L'analyse hiéarchique, Paris, Dunod.

MICHAEL, W. et al. (1963), An experimental determination of the optimal scoring formula for a highly-speeded test under different instructions regarding scoring penalties, *Educ. and Psych. Measmt.*, vol. 23, 83-99.

NOIZET, G., CAVERNI, J.P. et FABRE, J.M. (1977), Note sur l'expérimentation hors du laboratoire en docimologie, *Psychologie française*, vol. 22, 55-60.

NUTTALL, D.L. & SKURNIK, L.S. (1969), *Examination and Item Analysis* Manual, National Foundacion for Educational Research, Slough.

OOSTERHOF, A.C. & GLASNAPP, D. (1974), Comparative reliability and difficulties of the multiple-choice and true-false formats, *Jour. Exper. Educ.*, 42, 62-64.

PIERON, H. (1963), *Examens et docimologie*, Paris, PUF.

PIKLER, E. (1971), Se mouvoir en liberté dès le premier âge, Paris, PUF.

PIPPERT, R. (1966), Final note on the changed answer myth, *Clearing House*, 38, 165-166.

RASCH, G. (1968), *A Mathematical Theory of Obtectivity and its Consequences for Model Construction.* Paper delivered at European Meeting on Statistics, Econometrics and Management Science, Amsterdam.

REILING, E. & TAILOR, R. (1972), A new approach to the problem of changing initial responses to multiple-choice questions, *Jour. Educ. Measmt.*, 9, 67-70.

REMMERS, H. & GAGE, N. (1955), Educational Measurement and Evaluation, New York, Harper.

RICHARDSON, M. & ADKINS, D. (1938), A rapid method of selecting test items, *Journ. of Educ. Psychol.*, vol. 29, 547-552.

RISSE, R. (1972), *Réflexion docimologique à propos de deux examens du premier cycle des études médicales.* Thèse de doctorat en médecine, Paris, 381 p. polycopiées.

SABERS, D. & FELDT, L. (1968), An empirical study of the effect of the corrections for chance success on the reliability and validity of an aptitude test, *Journal of Educ. Measmt.*, vol. 5, n° 3, 251-258.

SARASON, (1960),

SHOEMAKER, D. (1973), *Principles and Procedures of Mutiple Matrix Sampling.* Cambridge (Mass.), Ballinger Ed.

SHUFFORD, E., ALBERT, A. & MASSENGILL, N.E. (1966), Admissible Probability Measurement Procedures, *Psychometrika*, 31, 125-145.

SLAKTER, M.J. (1968a), The penalty for not guessing, *Journ. of Educ. Measmt.*, 5, 141-144.

SLAKTER, M.J. (1968b), The effect of guessing strategy on objective test scores, *Journ. of Educ. Measmt.* vol. 5, 217-222.

SWINEFORD, F. (1938), The measurement of a personality trait, *Journ. of Educ. Psych.*, 29, 289-292.

SWINEFORD, F. (1941), Analysis of a personality trait, *Journ. of Educ. Psych.*, 32, 438-444.

SWINEFORD, F. et MILLER, P.M. (1953), Effects of directions regarding guessing on item statistics of a multiple-choice vocabulary test, *Journ. of Educ. Psych.*, 44, 129-133.

TERS, F., MAYER, G. et REICHENBACH, D. (1973), *Echelle Dubois-Buyse d'orthographe usuelle française*, Paris, UCDL.

TERS, F. (1973), *Orthographe et Vérités*, Paris, Editions E.S.F.

THORNDIKE, R.L. & HAGEN, E. (1969), *Measurement and Evaluation in Psychology and Education*, New York, Wiley, 3e éd.

THORNDIKE, R.L. (1971), Educational measurement for the Seventies, in Thorndike, R.L. (Ed.) *Educational Measurement*, American Council on Education, Washington.

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

VOTAW, D. (1936), The effect of do-not-guess directions upon the validity of true-false or multiple-choice tests, *Jour. of Educ. Psychol.*, vol. 27, 698-703.

WILLMOTT, A.S. & FOWLES, D.E. (1974), *The Objective Interpretation of Test Performance : The Rasch Model Applied.* National Foundation for Educational Research, Slough.

WOLF, R. (1967), Evaluation of several formulae for correction of item-total correlations in item analysis, *Journ. of Educ. Measmt.*, vol. 4, 21-26.

WRIGHT, B. & STONE, . (1979), Best test design, *Mesa Press*, Chicago.

ZILLER, R. (1957), A measure of the gambling response-set in objective tests, *Psychometrika*, vol. 22, 282-292.

TABLE DES MATIERES

DANS LA MEME COLLECTION :
EDUCATION 2000 dirigée par G. DE LANDSHEERE

LA PEDAGOGIE PALEOLITHIQUE
OU PREHISTOIRE DE LA CONTESTATION
de H. BENJAMIN. TRADUCTION de G. DE LANDSHEERE

PSYCHOLOGIE DE L'ENFANT ET DE L'ADOLESCENT
de J.A. RONDAL et de F. HOTYAT – 1re édition

EVALUATION CONTINUE ET EXAMENS
PRECIS DE DOCIMOLOGIE
de G. DE LANDSHEERE – 5e édition

TECHNOLOGIE EDUCATIVE ET AUDIO-VISUEL
de T. DECAIGNY – 3e édition

LE TEST DE CLOSURE
de G. DE LANDSHEERE – 4e édition

COMMUNICATION AUDIO-VISUELLE ET PEDAGOGIE
de T. DECAIGNY – 2e édition

CONSTRUIRE UN COURS PROGRAMME
de D. LECLERCQ, J. DONNAY, R. DE BAL et P. LAMBRECHT – 2e édition

COMMENT MESURER LA LISIBILITE
de G. HENRY – 2e édition revue et corrigée

LA RELATIVE EDUCATIONNELLE
de A. CLAUSSE

L'ETUDE EN EQUIPES
de G. POIRIER

CONCEPTS ET METHODES DE LA STATISTIQUE. Vol. 1 et 2
de L. D'HAINAUT

LA MISE EN PLACE DES POLITIQUES EDUCATIVES :
ROLE ET METHODOLOGIE DE LA CARTE SCOLAIRE
de J. HALLAK

DES FINS AUX OBJECTIFS DE L'EDUCATION
de L. D'HAINAUT – 3e édition revue et augmentée

INTRODUCTION A LA PSYCHOPEDAGOGIE DE L'EXPRESSION. Vol 1 et 2
de J. FAURY

COMMENT SE DOCUMENTER
de J.-E. HUMBLET

CARACTERISTIQUES INDIVIDUELLES ET APPRENTISSAGES SCOLAIRES
de B. S. BLOOM

LE GROUPEMENT DES ELEVES EN EDUCATION
de A. YATES

LES COMPORTEMENTS NON VERBAUX DE L'ENSEIGNANT
de G. DE LANDSHEERE et A. DELCHAMBRE

LE GRAPHISME PHONETIQUE
de V.A. GLADIC

ANALYSE ET REGULATION DES SYSTEMES EDUCATIFS, UN CADRE CONCEPTUEL
de L. D'HAINAUT

ELEMENTS DE PSYCHOLOGIE : UNE INTRODUCTION A LA PSYCHOLOGIE GENERALE
de J.A. RONDAL

EDUQUER LES PARENTS
de J.-P. POURTOIS

DROGUES, UN PROGRAMME D'EDUCATION POUR LA SANTE
de Ernst SERVAIS

POURQUOI LOGO DANS UN CONTEXTE EDUCATIF ?
de J.-L. HARDY

LES JEUNES, L'ECONOMIE ET LA CONSOMMATION
Sous la direction de J.-M. ALBERTINI

ENTRER A L'UNIVERSITE
de J.-P. POURTOIS

AUTONOMIE ET CONDITIONNEMENT CHEZ L'ENFANT ET L'ADOLESCENT
de R. GUBBELS

L'ART ET LA SCIENCE DE L'ENSEIGNEMENT
de M. CRAHAY et D. LAFONTAINE

UNE METHODE D'ECRITURE DES DIDACTICIELS - PILOT
de C. DEPOVER

LA CONCEPTION DES QUESTIONS A CHOIX MULTIPLE
de D. LECLERCQ

LES HABITUDES TABAGIQUES – COMMENT LES DEMYSTIFIER
de M. FRYDMAN

ACHEVE D'IMPRIMER
LE 15 OCTOBRE 1987

IMPRIME EN BELGIQUE D/1987/258/31 ISBN 2-8040-0245-4