L'ANALYSE CLASSIQUE D'ITEMS

Marc Demeuse et Georges Henry

1. Introduction

Sans entrer dans tous les détails de procédures parfois complexes, nous analyserons, dans les pages qui suivent, quelques méthodes classiques utilisées dans la perspective de la construction des tests psychologiques ou pédagogiques.

L'***analyse d'items classique*** fournit d'habitude deux types d'informations : des indices de difficulté des items et des indices de discrimination. Ce dernier type d'indices se rapporte à la façon dont l'item mesure ou discrimine les sujets en accord avec le reste du test.

Après avoir soumis un ensemble d'items à un groupe de sujets analogues à ceux à qui le test est destiné, on procédera aux calculs des différents indices que nous venons d'évoquer de manière à savoir quels items seront conservés et quels items seront rejetés. Cela suppose donc que le nombre d'items lors du pré-test est plus élevé que le nombre d'items qui seront finalement conservés.

Le recours aux différents indices ne permet pas d'éviter la première étape, essentielle, de la construction d'un test : la définition aussi précise que possible de l'objet de la mesure. Cette définition, selon l'objet de la mesure, peut reposer sur différents référentiels (programmes d'enseignement, définition d'un concept théorique, autre test ou examen déjà validé). Cette première étape vise à assurer la validité *a priori* du test. Celle-ci sera établie de manière définitive après avoir administré le test à un échantillon de sujets (validation empirique). Nous avons abordé ce sujet dans le chapitre consacré à la validité des mesures, nous n'y reviendrons pas ici. Nous n'aborderons pas plus les problèmes posés par la rédaction des items. Il existe à ce sujet plusieurs ouvrages, notamment dans le domaine de la rédaction des questions à choix multiples (Leclercq, 1986) et d'autres cours avancés permettent aux étudiants de s'y intéresser.

Nous nous limiterons à la sélection des items qui seront conservés dans la version finale du test, en fonction de l'analyse classique d'items. Il existe à présent d'autres formes d'outils pour mettre au point un test. Nous envisagerons, dans la partie consacrée aux échelles de mesure, le ***modèle de Rasch*** qui, lui aussi, fournit une aide à la mise au point de tests. Nous ne développerons cependant pas cet aspect dans ce document introductif.

2. Indices de difficulté d'items

Si l'item est corrigé de manière dichotomique (0 ou 1), l'indice de difficulté le plus élémentaire est le pourcentage de réponses correctes ($\mathbf{p_i}$). A cet indice correspond la probabilité qu'un « sujet moyen » appartenant à la population a de réussir l'item i. L'indice $\mathbf{p_i}$ constitue un indice de difficulté moyen pour l'ensemble des individus testés. Par contre, nous ne disposons pas d'informations exactes sur la difficulté de cet item pour un individu particulier. Tout ce que nous savons, c'est que si un individu réussit un item, cet item présente certainement une difficulté relativement faible par rapport à l'aptitude de cet individu[1].

[1] On se rapportera à la partie consacrée à la psychophysique à ce sujet: on peut tester une seule fois un ensemble d'individus ou, au contraire, tester un seul sujet de manière répétée. Selon l'optique adoptée, on dispose d'une proportion de détection à un stimulus donné à travers un ensemble de sujets ou, au contraire, à travers les réponses d'un seul individu.

L'indice de difficulté p_i de l'item i est particulièrement simple à calculer dans le cas des corrections dichotomiques (réponse correcte ou réponse fausse) puisqu'il s'agit de comptabiliser le nombre de bonnes réponses enregistrées pour l'item i à travers l'ensemble des individus testés et de diviser ce nombre de bonnes réponses par le nombre total d'individus soumis au test. Malheureusement, cet indice ne reflète pas nécessairement bien la proportion de réponses correctes résultant d'une réelle compétence. En effet, lorsqu'on traite des résultats de questions à choix multiples ou de vrai/faux, on doit tenir compte de la possibilité d'enregistrer des réponses correctes « par chance ». Il existe plusieurs méthodes pour diminuer ou contrôler au mieux les phénomènes de réponses influencées par la chance dans le cas des questions à choix multiples :

- augmentation du nombre de distracteurs,
- ajout de distracteurs du type « toutes les réponses sont correctes », « toutes les réponses sont fausses », « la question présente une aberration logique »…[2]
- pénalisation des réponses fausses par une coefficient négatif,
- utilisation de degrés de certitude qui créditerons de manière plus ou moins généreuse ou sévère les bonnes et les mauvaises réponses des sujets en fonction de la confiance qu'ils déclarent accorder à leurs propres réponses[3].

Parmi les quatre solutions envisagées, nous n'aborderons que la pénalisation des réponses fausses par un coefficient négatif. Cette méthode est certainement la plus simple et la plus couramment utilisée, malgré ses lacunes au niveau du postulat de base, comme nous allons le voir.

Afin de bien comprendre le raisonnement, on se placera d'abord au niveau d'un individu particulier avant de raisonner au niveau de l'ensemble des sujets. Si, pour chaque question, il y a **k** solutions proposées, chaque individu aura une chance sur **k** de trouver la bonne réponse en choisissant purement au hasard et **k-1** chances sur **k** de se tromper. Supposons qu'un individu particulier ait répondu purement au hasard. Si on attribue un point à chaque réponse correcte fournie pour un test de longueur **n**, cet individu aura comme score le plus probable $\frac{n}{k}$ points. Or, ce score devrait, en toute logique, être nul, puisque l'individu a une compétence nulle[4]. Il faut donc pénaliser les réponses fausses afin d'arriver à ce que le total de ces réponses fausses produise un score de $-\frac{n}{k}$, annulant ainsi le bénéfice $\frac{n}{k}$ des réponses correctes, obtenues par hasard.

Pour connaître la pénalité *x* à attribuer à chaque mauvaise réponse, on résoudra l'équation suivante:

$$n\frac{(k-1)}{k}x = -\frac{n}{k}$$

2 On peut, à ce sujet, consulter Leclercq et *al.* (1998, pp. 173-174) à propos de ce qu'il appelle « solutions générales implicites ».

3 On peut, au sujet des indices de certitude, consulter Leclercq (1983). Nous ne développerons pas ce sujet dans le cadre de l'analyse d'items classique. Cette méthode est une manière plus élaborée d'attribuer les coefficients positifs et négatifs aux bonnes et aux mauvaises réponses, en se basant sur les informations fournies par les sujets sur leurs propres certitudes.

4 Dans notre exemple, il n'a pas mobilisé la compétence que l'on souhaite évaluer puisqu'il a répondu purement au hasard, ce qui nous fait écrire qu'il a une compétence nulle.

Ce qui signifie que, sachant qu'il existe *n* questions pour chacune desquelles les « chances » de répondre de manière erronée sont $\frac{k-1}{k}$, il convient de trouver *x*, la pénalité à appliquer à chaque mauvaise réponse de manière à ce que le total des mauvaises réponses produise un résultats de $-\frac{n}{k}$ points, contrebalançant un résultat positif équivalent, soit $\frac{n}{k}$ points, produit par les réponses correctes, mais dues au hasard.

Si on résout cette équation, on obtient la valeur de la pénalité, soit $x = -\frac{1}{k-1}$. Cela signifie que la pénalité (négative) doit être égale à l'unité divisée par le nombre de réponses proposées moins une, pour les situations où une bonne réponse est créditée d'un point. Cette approche suppose que chaque distracteur a un pouvoir d'attraction identique. Ce n'est cependant pas toujours le cas : certains distracteurs peuvent être éliminés de manière évidente par les sujets parce qu'ils sont très peu plausibles. Dans ce type de situations, le choix au hasard ne va alors s'effectuer que sur les distracteurs restants et la bonne réponse, ce qui augmente les chances de réussite. On ne peut malheureusement pas neutraliser facilement ce biais, si ce n'est en pré-testant les instruments et en éliminant les distracteurs peu attractifs[5].

Exemple 1

Supposons un test de 20 questions pour lesquelles 5 solutions sont proposés. On enregistre, chez un sujet particulier, 4 réponses correctes et 16 réponses fausses. On calcule simplement le score brut de ce sujet.

Score brut du sujet : 4 points

La correction pour choix au hasard (score corrigé) sera, par contre différent:

Score corrigé du sujet $4-\frac{16}{5-1}=0$ *point*

Ces deux résultats peuvent être exprimés sur 20 points.

Exemple 2

Supposons à présent un autre test de 20 questions, mais qui n'offre que 4 solutions à chacune des questions. On enregistre, pour un sujet particulier, 13 réponses correctes et 3 réponses fausses. Le sujet omet de répondre à 4 questions. On obtiendra :

Score brut du sujet : 13 points

Score corrigé du sujet : $13-\frac{3}{4-1}=13-1=12$ *points*

Ici aussi, ces deux résultats peuvent être exprimés sur 20 points.

Comme on le constate, les omissions ne sont jamais pénalisées. En effet, en omettant, l'élève ne se donne aucune chance de réussir l'item par hasard.

Il convient de souligner que le score brut constitue une borne supérieure (on fait l'hypothèse que l'élève connaissait réellement les bonnes réponses à certaines questions et qu'il s'est

[5] On conservera malgré tout certains distrateurs peu attractifs s'ils permettent, dans une perspective diagnostique, d'identifier les sujets qui ne maîtrisent réellement pas des connaissances essentielles et qui sont capables de se rallier à des propositions particulièrement aberrantes ou dangereuses. C'est un peu l'idée qui est mise en pratique dans le cadre du permis de conduire : deux mauvaises réponses à des questions portant sur des problèmes graves entraînent le refus immédiat. Dans ce cas, c'est cependant plutôt les items que les distracteurs qui sont pris en compte.

trompé, pour toutes les autres, sans choisir au hasard). Par contre, le score corrigé constitue une borne inférieure, puisqu'on se place dans l'hypothèse la plus défavorable au sujet : on fait l'hypothèse que toute réponse correcte à une question a pu être produite aléatoirement et que toute réponse fausse résulte également d'un choix aléatoire. La compétence vraie du sujet se situe quelque part entre ces deux bornes, mais on ignore où puisqu'on ne sait rien des processus de décision qu'il a mis en œuvre. Le sujet a-t-il répondu au hasard lorsqu'il ne connaissait pas la bonne réponse ou croyait-il la connaître, mais se trompait ?

Le même raisonnement peut s'appliquer à un ensemble de sujets et aux pourcentages de réussites, d'erreurs ou d'omissions.

Si **p** est le pourcentage brut de réponses correctes, $\mathbf{p_e}$ le pourcentage d'erreurs, $\mathbf{p_o}$, le pourcentage d'omissions et **k** le nombre de solutions proposées, le pourcentage corrigé de réponses correctes $\mathbf{p_c}$ sera :

$$p_C = p - \frac{p_e}{k-1}$$

Exemple

Supposons un test pour lequel on obtient 62% de réponses correctes, 26% de réponses incorrectes et 12% d'omissions. Chaque question du test offre 5 réponses possibles. Soit:

$p = 62\ \%$

$p_e = 26\ \%$

$p_o = 12\ \%$

$k = 5$

On peut calculer le pourcentage corrigé de réponses correctes comme

$$p_c = 62 - \frac{26}{5-1} = 62 - 6{,}5 = 55{,}5\%$$

Le pourcentage moyen de compétence vraie de la population testée[6] se situe donc dans la fourchette 55,5% - 62%.

Mais, il existe encore d'autres problèmes liés au calcul d'un taux brut de réponses correctes à un item. En effet, on a défini le pourcentage brut de réponses correctes à un item :

$$p_i = \frac{\text{nombre de réussites à l'item}}{\text{nombre de sujets essayant de résoudre l'item}}$$

mais les derniers items d'un test sont parfois résolus par les sujets, faute de temps (voir à ce sujet la discussion à propos des tests de vitesse et des tests de puissance). En effet, on peut soit considérer que tous les sujets ont eu l'occasion de tenter chaque item (tests de puissance), soit admettre que le test a une certaine composante de vitesse (ce qui est, malheureusement, souvent le cas). Dans cette dernière hypothèse, certains sujets cessent de répondre, au fur et à mesure que l'on s'approche de la fin du test. On parle alors d'items non atteints (**NA**). Ces omissions ne sont pas comparables à celles qui se produisent en début ou au milieu du test parce que le sujet ignore la réponse correcte. On traite donc, malheureusement, la plupart du temps les items non atteints de deux manières qui s'avèrent toutes deux insatisfaisantes.

[6] Il faudrait encore tenir compte de l'erreur d'échantillonnage si les pourcentages mentionnés ont été établis à partir d'échantillons ! Nous ne tenons ici en compte que le problème spécifique de l'erreur de mesure due aux choix au hasard dans le cas d'une population.

a) $p_i = \frac{\text{nombre de réussites}}{\text{nombre total de sujets}}$

Dans ce cas, il y a sous-estimation des p_i puisqu'un certain nombre de sujets n'ont pas eu l'occasion d'essayer de répondre à l'item, alors qu'ils auraient peut-être été capables de le résoudre.

b) $p_i = \frac{\text{nombre de réussites}}{\text{nombre total de sujets-nombre de sujets n'ayant pas atteint l'item}}$

Dans ce cas, il y a surestimation des p_i, car ce sont en général les sujets les plus aptes qui sont arrivés à la fin du test (il n'y a généralement pas indépendance totale entre vitesse et puissance).

A nouveau, et comme c'était le cas lorsqu'on corrigeait les pourcentages de réussite pour choix au hasard, la vérité se situe entre ces deux bornes, sans qu'il soit possible de déterminer celle-ci de manière précise.

3. Indices de discrimination des items

Les indices de discrimination des items que nous décrirons ci-dessous sont, en fait, des indices de consistance interne puisqu'il n'y a pas, le plus souvent, de critère externe utilisé[7].

3.1. Caractéristiques psychométriques des items

Ferguson propose une méthode qui n'est pas sans analogies avec la méthode des stimuli constants. Supposons que l'on veuille analyser un item donné. Sur la base du score total, on réparti les sujets en sept groupes, avec des intervalles constants de **0,6 σ_t** par exemple[8].

Figure 1 - *Répartition des sujets en sept groupes selon leur score au test.*

Très faible	Faible	Moyens Faibles	Moyens	Moyens Forts	Forts	Très forts
-1,5 σ	**-0,9 σ**	**-0,3 σ**	**M**	**+0,3 σ**	**+0,9 σ**	**+1,5 σ**

Si on considère un item particulier, réussi par exemple à 50 %, les sujets forts ou très forts à l'ensemble du test auront théoriquement une probabilité de réussite beaucoup plus élevée et les sujets faibles une probabilité de réussite beaucoup plus basse. Ceci sera vrai pour autant que l'item mesure la même chose que l'ensemble du test. Après avoir calculé le taux de bonnes réponses pour chacun des sept groupes, on peut alors confronter les résultats à cette hypothèse d'un taux croissant en fonction des résultats globaux. Si celle-ci ne peut être étayée, il convient de s'interroger sur ce que mesure réellement l'item considéré, en regard de ce que mesure l'ensemble du test. Cette méthode très intuitive a cependant pu être améliorée avec la mise à disposition d'outils informatique.

C'est ainsi que, dans une publication de 1982, Chopin préconise une procédure similaire pour analyser en détail le fonctionnement des items, mais sans recourir à un système de répartition en 7 groupes. Il ne se contente pas d'analyser la réponse correcte; il examine aussi le fonctionnement des distracteurs. En outre, disposant des scores au test pour plusieurs milliers de sujets, il travaille score par score, sans être obligé d'effectuer des regroupements par

[7] Certains parlent néanmoins improprement d'indices de validité ou de qualité des items.

[8] 0,6 σ_t, c'est-à-dire 60% de la valeur de l'écart-type du score total au test pour l'ensemble des sujets testés.

classes comme c'est le cas ci-dessus. Prenons l'exemple de l'item B24 (Sciences, Recherche IEA, Australie)

Par laquelle des méthodes suivantes le temps géologique peut-il être mesuré avec la meilleure précision ?

A. La proportion d'isotopes d'uranium dans certaines roches.

B. L'épaisseur des couches de roches sédimentaires

C. Le volume de fossiles.

D. Le taux d'accumulation saline de l'océan.

E. Les températures du manteau de la terre.

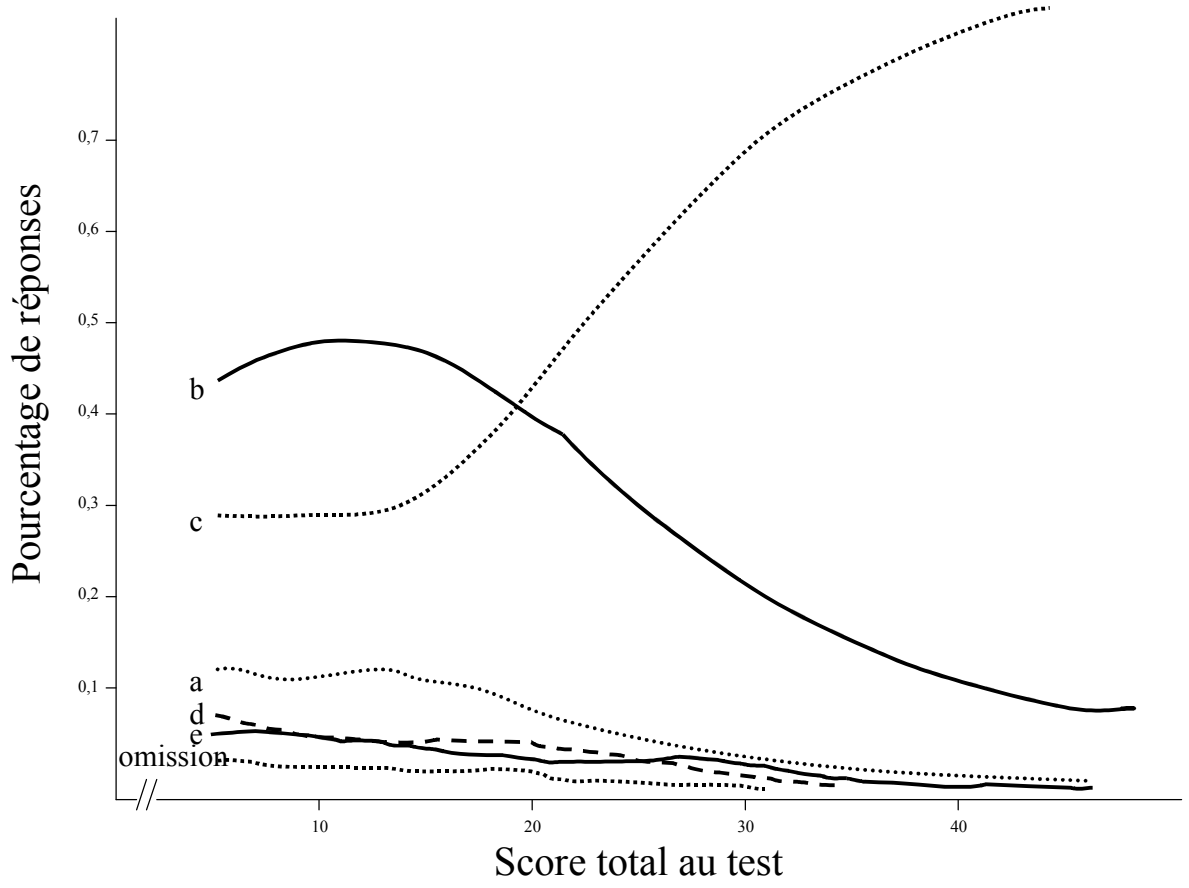
Si on analyse les réponses à cet item, on constate que pratiquement tout se joue entre la réponse correcte ***C*** et la réponse moins précise ***B***. Seules ces deux réponses sont réellement attractives, la réponse ***C*** étant meilleure que la réponse ***B***, les autres étant manifestement erronées. Si on analyse de gauche à droite la figure suivante (figure 2), c'est-à-dire depuis les élèves les plus faibles jusqu'aux élèves les plus forts, on constate tout d'abord que, parmi les élèves très faibles (obtenant moins de 15 au score total), environ 30 % choisissent la bonne réponse, alors que l'on pourrait attendre environ 0 %. Il y a là un effet typique de plancher communément constaté dans le cadre de tests à choix multiple[9]. On obtient rarement pour un item à 5 choix un taux de réponses correctes en dessous de 15 %. Dans cet exemple, l'effet de plancher se manifeste à environ 30 % et est apparemment causé par la non-attractivité des réponses ***A***, ***D*** et ***E***[10]. Il est également intéressant de constater que ce n'est qu'à partir d'un score total de ± 15 que la proportion de réponses correctes devient de plus en plus importante et que, parallèlement, le distracteur ***B*** devient moins attractif, la réponse correcte prenant réellement le pas au-delà d'un score de 20.

Une telle analyse fournit des renseignements précieux sur les caractéristiques psychométriques de l'item. On constate en effet, une évolution différente du choix des distracteurs selon les compétences des élèves. Il existe des logiciels qui permettent à présent de réaliser ce type d'analyses relativement simplement, comme par exemple ***TestGraph*** développé par J.O. Ramsay, de l'Université McGill. De telles analyses permettent par exemple, de rejeter des distracteurs très peu attractifs, d'identifier des distracteurs particulièrement choisis par les élèves les plus faibles et d'en déduire des informations sur les réponses erronées ou même d'identifier des distracteurs particulièrement choisis par les sujets les plus performants à l'ensemble du test, au détriment de la bonne réponse et de s'interroger sur les raisons de ces réponses aberrantes.

9 Voir à ce sujet aussi la correction pour choix au hasard.

10 Dans la section consacrée à la correction pour choix au hasard, nous n'avons pas pris en compte le caractère plus ou moins attractif de certains distracteurs. Nous avons considéré que les réponses étaient simplement réparties de manière égale entre les distrateurs et la bonne réponse, ce qui est rarement le cas. Dans les faits, même si cinq réponses sont proposées, le sujet qui a une compétence faible peut, malgré tout, éliminer certaines possibilités et accroître ses chances de répondre correctement.

Figure 2 - *Distribution des taux de réponse pour chacune des solutions proposées à un item du type choix multiple en fonction du score total au test (item B24 d'un test de Science de l'IEA en Australie, d'après Choppin, 1982).*



3.2. Indices de discrimination entre groupes « forts » et « faibles »

A côté des analyses graphiques, une série de formules existent qui se fondent toutes sur le principe qu'un item discrimine s'il est mieux réussi par le groupe fort que par le groupe faible. La force des groupes est, le plus souvent, déterminée par rapport aux résultats globaux au test (Exemple : méthode de Davis). Ces méthodes sont de moins en moins utilisées, compte tenu des progrès de l'informatique, au profit des indices de corrélation des résultats d'un items avec ceux obtenus à l'ensemble du test, comme nous allons le voir.

3.3. Indices de corrélation

Ces indices, de plus en plus utilisés, permettent de déterminer dans quelle mesure la réussite ou l'échec à un item donné est en relation avec le score total obtenu par l'élève. En d'autres termes, on recherche quelle part prend l'item dans la composition du score total. Lorsque ce score total est obtenu par sommation de résultats dichotomiques aux items du test, on peut utiliser la formule de la ***corrélation bisériale de point*** ou r_{pbis}. Cette corrélation repose sur l'idée suivante : un sujet qui a répondu correctement à un item présente une plus grande probabilité d'obtenir un résultat élevé au test que celui qui a échoué.

Formule

$$r_{pbis} = \frac{M_r - M_e}{\sigma_t}\sqrt{p_i q_i}$$

où M_r= Moyenne des scores totaux obtenus à l'ensemble des items du test par les sujets qui ont répondu correctement à l'item considéré.

M_e= Moyenne des scores totaux obtenus à l'ensemble des items du test par les sujets qui ont fourni d'autres réponses (erreurs ou omissions) à l'item considéré.

σ_t = Ecart-type de la distribution des scores totaux.

p_i = Proportion des sujets qui ont répondu correctement à l'item considéré.

q_i = Proportion des sujets qui n'ont pas répondu correctement à l'item considéré (erreurs ou omissions).

Interprétation

Le r_{pbis} équivaut au ***coefficient de corrélation de Bravais-Pearson*** lorsqu'on corrèle une échelle d'intervalles avec une distribution binomiale (1/0). L'interprétation est donc la même (marge de variation : -1 à +1).

Théoriquement, la moyenne obtenue au score total par les élèves qui ont bien répondu à l'item i doit être supérieure à la moyenne des élèves qui ont répondu incorrectement ou qui ont omis ($M_r - M_e > 0$). Il s'ensuit que s'il y a consistance interne de l'item par rapport à l'ensemble du test, $r_{pbis} > 0$ puisque σ_t et $\sqrt{p_i q_i}$ sont également positifs.

Il existe des formules complexes pour établir la valeur qu'un r_{pbis} devrait atteindre, en tenant compte, par exemple, du nombre d'items dans le test[11]. On considère cependant comme étant acceptables les items dont le $r_{pbis} > 0{,}20$ ou $0{,}25$.

Par ailleurs, on peut aussi utiliser cet outil pour analyser en détail le comportement des distracteurs. Normalement, les distracteurs doivent présenter une corrélation négative avec le score total. On procède donc de la même manière, mais en considérant respectivement M_r et M_e comme les moyennes obtenues à l'ensemble du test par les sujets qui ont choisi un distracteur particulier et ceux qui n'ont pas choisi ce distracteur. Les valeurs du r_{pbis} ainsi obtenues, bien que négatives en théories, peuvent être faibles, d'autant plus que les distracteurs sont nombreux et exercent un même pouvoir d'attraction sur l'ensemble des sujets faibles en fonction de leur score total.

[11] Le problème essentiel réside dans le fait que l'item qu'on analyse participe au calcul du score total et donc, on observe une corrélation "mécanique" due à ce seul fait. Ce lien est d'autant plus fort que le test compte peu d'items. Retirer l'item dans le calcul du score total altère la composition du test et ne constitue pas non plus une solution absolument satisfaisante. On peut, à ce propos, consulter l'article de Hardy (1983).

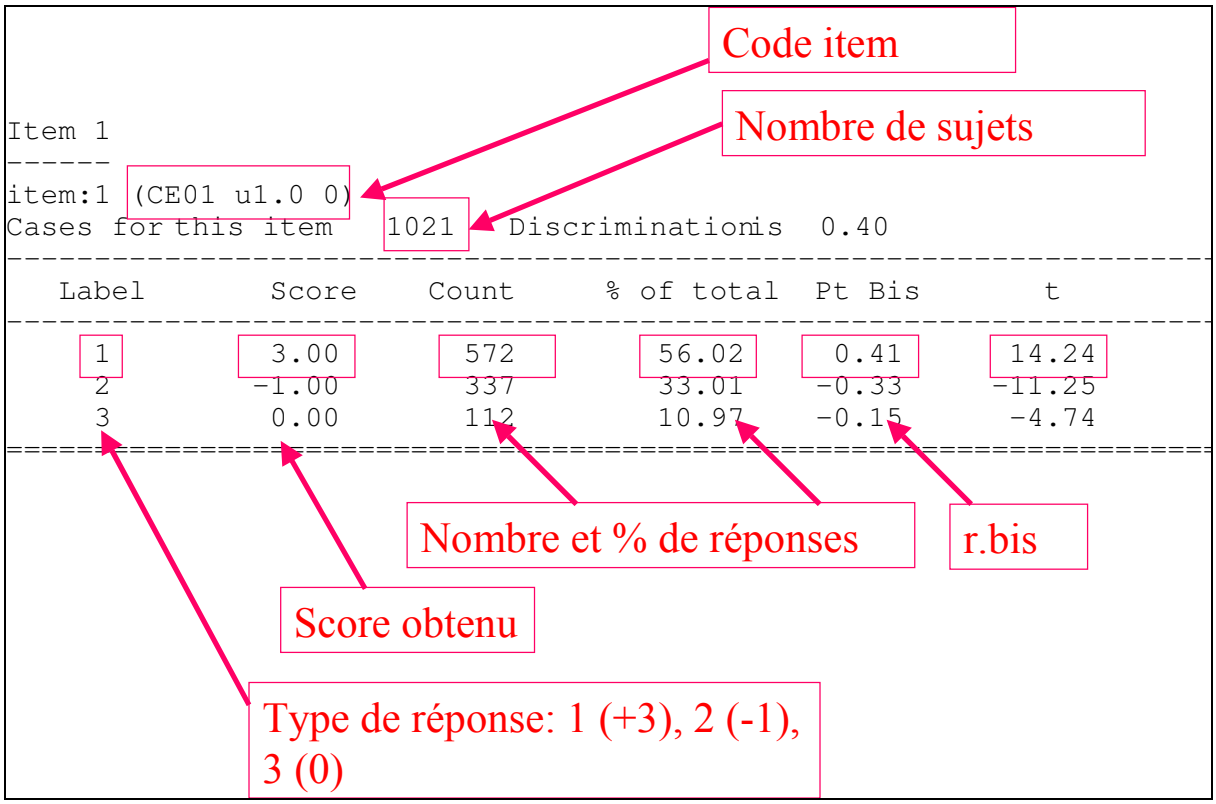


Figure 3 - *Extrait d'une analyse d'item par le logiciel ConQuest* (Wu, Adams & Wilson, 1998). Chaque solution (1, 2 ou 3) reçoit un nombre de points particulier (3 points, -1 point ou 0 points). La dernière colonne fournit une valeur t permettant d'éprouver la signification de la valeur du r_{pbis}.

4. Problèmes spécifiques

4.1. Effets de recouvrement

Lorsque le score total à un test est le critère, la corrélation item/score total est surestimée à cause de la part prise par l'item dans la composition du score total, surtout si le test compte un petit nombre d'items. Il existe plusieurs méthodes de correction qui ont notamment été décrites par Hardy (1983), nous avons évoqué ce problème ci-dessus.

4.2. Effet de contraction (*shrinkage*)

Lorsqu'on calcule une corrélation item/score total, une erreur d'échantillonnage est associée au r_{pbis} trouvé, puisque celui-ci a été calculé au départ d'un échantillon et est donc différent de la vraie valeur existant au niveau de la population. Lorsqu'on choisit, parmi de nombreux items, ceux qui présentent les corrélations les plus élevées, on peut penser qu'un certain nombre de ces corrélations ont pu être surestimées à cause d'erreurs d'échantillonnage allant dans le sens d'une augmentation de la corrélation. Lorsqu'on administre par après la version définitive du test qui a été mise au point en se basant sur ces corrélations, on obtient souvent des coefficients de fidélité plus faibles que ceux auxquels on s'attendait (effet de « shrinkage »). On peut utiliser des procédures de validation croisée, identiques à celles utilisées dans le cadre de la prédiction multiple, pour déceler cet effet de « shrinkage » dès le prétest et pour ne sélectionner que des items dont la stabilité s'avère élevée à travers des échantillons différents.

4.3. Effets de la chance sur les r_{pbis}

Lorsqu'on utilise des questions à choix multiple, il existe une probabilité de réussite au hasard d'autant plus grande qu'il y a peu de solutions proposées. On a vu qu'on peut corriger $\mathbf{p_i}$ pour obtenir $\mathbf{p_c}$.

Un effet parasite du même ordre existe dans le calcul des corrélations. En effet, lorsque le test est difficile et que le nombre de solutions proposées est peu élevé, les sujets auront une forte tendance à choisir au hasard. Une variance d'erreur importante est ainsi introduite dans les résultats et les corrélations items-score total seront sous-estimées. Si le même test est administré à une population présentant un degré d'aptitude plus élevé, les corrélations vont augmenter, car les choix au hasard vont diminuer.

4.4. Effets de la vitesse sur les r_{pbis}

La vitesse influence de manière nocive les corrélations items/score total. Si on compte les items non atteints comme des échecs, il y aura surestimation des corrélations. Si on neutralise les items non atteints, il y aura sous-estimation.

4.5. Comment procéder pour améliorer la validité ?

Pratiquement, on peut procéder de deux façons pour accroître la validité d'un test.

a) On fait l'hypothèse que le test tend vers l'univocité (une seule dimension est mesurée, le test est unidimensionnel). Sa validité sera, dès lors, accrue si on améliore sa fidélité. Cependant, on sait que ce type de test est très peu fréquent. Néanmoins, si cette approche est retenue, on peut élaborer des batteries de tests dont les différentes composantes seront pondérées de façon optimale.

b) Si on renonce à l'utilisation de batteries de test et si on veut assurer à un seul test une validité maximale, il faut rechercher des corrélations items-critère (ρ_{ic}) élevées et des inter-corrélations entre items (ρ_{ij}) faibles.

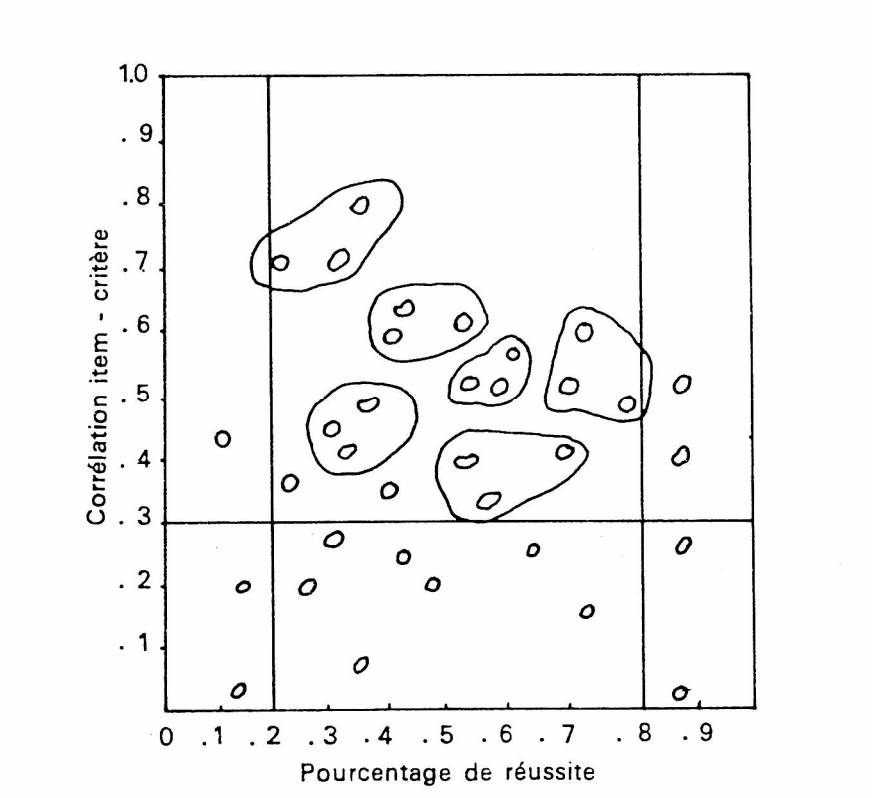
La première méthode est plus sûre, mais, de loin, plus coûteuse. Elle présente cependant un avantage important. Si les sous-tests qui composent la batterie sont de longueur suffisante pour assurer une fidélité minimale (>0,50) de ceux-ci, l'interprétation des résultats peut fournir des hypothèses de diagnostic, c'est-à-dire conduit à identifier des domaines problématiques. La validité des hypothèses doit cependant être vérifiée en utilisant des tests plus fidèles, mais on obtient malgré tout une sorte de "dépistage" ou de premier crible.

4.6. Choix des items en vue de la construction de formes parallèles

D'un point de vue strictement psychométrique, on définit comme formes parallèles, des tests ayant des moyennes et des variances égales ainsi que des intercorrélations élevées.

Pour construire des formes parallèles, on se sert de diagrammes semblables à celui présenté ci-dessous (figure 3).

Figure 3 - *Diagramme permettant d'effectuer le choix d'items destinés à construire des formes parallèles (extrait de De Landsheere, 1988). Chaque petit cercle représente un item particulier. Certains items sont regroupés à l'intérieur d'un ensemble, en raison de leur similitude.*



Ce diagramme de points représente les pourcentages de réussite (p_i) et les corrélations item-critère (r_{pbis}) habituellement utilisées lors de la sélection des items et montrant comment ces derniers peuvent être choisis en vue de produire trois formes parallèles (procédure de Gulliksen décrite, par exemple, par De Landsheere, 1988). Les items dont le $\mathbf{p_i}$ est supérieur à 0,8 ou inférieur à 0,2 sont éliminés, car ils sont peu susceptibles d'être discriminants. Il en va de même des items dont le r_{pbis} est inférieur à 0,30. A la suite de cette procédure, on répartira les items dans les différentes formes en choisissant pour, chacune des formes, des items dont les caractéristiques sont voisines.

4.7. Facteurs susceptibles d'introduire des biais dans les réponses

4.7.1. Facteurs liés au sujet et à ses dispositions mentales

On considère qu'il y a « biais de réponse » lorsque la réponse à un item a tendance à être altérée par un quelconque élément étranger à ce que l'item est supposé mesurer. Les biais que nous décrivons ci-dessous sont le plus souvent causés par un certain nombre de dispositions mentales spécifiques à l'individu et qui altèrent la mesure que l'on prétend obtenir au départ de l'item. Ces biais constituent donc une influence négative sur la validité des scores de test que l'on obtient. Ils peuvent aussi bien perturber la mesure individuelle, chez un sujet particulier, que des mesures différentielles, du fait de l'existence de biais spécifiques chez certains groupes de sujets.

a. La tendance à « deviner »

Lorsqu'il s'agit de répondre à des questions à choix multiple, certains sujets auront, plus que d'autres, tendance à répondre au hasard. Cette tendance différentielle existe tant au niveau des individus que des cultures. Le score obtenu mélange donc dans sa composante vraie, la

composante relative au test et la composante relative à la tendance à deviner. Les circonstances peuvent aussi influencer le sujet: importance de l'enjeu du test, sanction ou non des mauvaises réponses…

b. Interprétations sémantiques

Lorsqu'on utilise des catégories telles que « d'accord », « parfois », etc. on laisse la place à bon nombre d'interprétations individuelles. Ces interprétations peuvent différer d'un individu ou d'un groupe à l'autre.

c. Impulsivité

Il s'agit de la tendance à fournir beaucoup de réponses, que ce soit lors d'un questionnaire à choix multiple ou lors d'épreuves ouvertes. Certains sujets peuvent, plus que d'autres, être tentés de répondre rapidement et le plus possible, alors que d'autres se "bloquent" sur certaines questions lorsqu'ils ne connaissent pas la réponse correcte.

La forme des questions peut influencer cette tendance et se marquer différemment selon certains groupes d'appartenance. Nous avons déjà signalé, par exemple, la proportion plus grande de non-réponses chez les garçons lorsqu'ils sont confrontés à des questions ouvertes, plutôt que des questions du type QCM, bien qu'ils soient, comme les filles, moins habitués à ces dernières, dans l'enseignement secondaire belge francophone. Le même type de différence se marque aussi en termes de résultats entre, par exemple, les élèves flamands et francophones, en Belgique : les premiers réussissant mieux les questions ouvertes que les questions fermées, alors que c'est l'inverse chez les jeunes francophones du début d'enseignement secondaire en sciences, du moins dans la seconde étude internationale sur les mathématiques et les sciences (Monseur et Demeuse, 1998).

d. Tendance à acquiescer

Tendance à dire « oui » plutôt que « non ». Cette tendance se marque jusque dans les questionnaires cognitifs à choix multiple lorsque le choix est laissé entre « vrai » et « faux ». De nouveau, il existe des différences inter-individuelles et inter-culturelles (possibilité d'une tendance inverse dans certains types de populations particulières ou tendance à l'opposition).

e. Vitesse et exactitude

Ce problème a été longuement analysé dans les chapitres qui précèdent, nous n'y reviendrons donc pas ici.

f. Désirabilité sociale

Le sujet cherche, parmi les solutions qui lui sont proposées, celle qu'il croit correspondre à l'attente de la personne qui le questionne. Le sujet peut donc, par ce moyen, augmenter son score, sans que ce soit en rapport avec sa compétence vraie. C'est principalement dans les échelles d'attitude que ce biais peut se marquer de manière importante.

g. Fatigue, stress et altération de l'état mental du sujet

Fatigue et stress, comme d'autres altérations de l'état mental du sujet (euphorie, imprégnation médicamenteuse ou alcoolique…), modifient les réponses des sujets, soit de manière récurrente, soit de manière passagère. C'est tantôt les réponses qui deviennent inconsistantes au sein d'une même épreuve, par exemple, à cause d'une dégradation des capacités du sujet au fil de l'épreuve, c'est tantôt le niveau général des performances qui est altéré sur l'ensemble de la session de test, c'est encore une instabilité de l'état du sujet qui entraîne des résultats variables d'une passation à une autre. Ces différentes modalités peuvent aussi bien altérer la fidélité, par consistance interne ou lors de situation de test-retest, que la validité, modifiant la

nature même de ce qui est réellement mesuré (par exemple, à cause d'un accroissement des réponses au hasard).

Dans certains cas, des effets de maturation peuvent aussi se produire, liés au développement du sujet, surtout chez les sujets jeunes ou âgés[12]. On sera ainsi parfois amené à attribuer erronément une amélioration des résultats observés lors d'un post-test au traitement qui a été administré alors que celui-ci n'a produit aucun effet. Il importe, dans ce type de situation, d'adopter des plans expérimentaux appropriés.

h. Effet de testing

Certains sujets peuvent être plus ou moins habitués à certains tests. Ils en mémorisent les principes et consignes, ce qui facilite leur travail, mais ils mémorisent aussi certaines solutions ou méthodes de résolution. Les scores sont donc améliorés d'une passation à l'autre et certains groupes peuvent obtenir des scores plus élevés parce qu'ils sont plus familiarisés avec ce type d'épreuves.

Ce type d'effet peut aussi se marquer à l'intérieur même d'une épreuve: les questions deviennent de plus en plus familières, le sujet comprend mieux ce qui lui est demandé… On observe ainsi un effet d'ordre, lié à la position de l'item dans le test. Certains sujets ont aussi besoin d'un nombre plus ou moins grand d'essais avant d'atteindre leur "vitesse de croisière". Cet effet d'ordre peut être contrebalancé par le fait que le sujet est confronté à un test de vitesse et que les derniers items ont moins de chance que les premiers d'être atteints.

4.7.2. Facteurs liés à la situation d'évaluation elle-même

La situation de test elle-même peut altérer la mesure. Il peut aussi bien s'agir de facteurs liés à la différence de conditions, certains sujets étant plus ou moins bien placés dans la salle de test, la salle étant plus ou moins bruyante et la session, perturbée par des événements extérieurs. Il peut également s'agir de problèmes qui résultent de l'interaction spécifique entre certains sujets et les conditions particulières de la situation, par exemple, certains groupes d'élèves sont testés par leur enseignant, alors que d'autres sont testés par des examinateurs externes qui ne parviennent pas à maintenir la discipline dans les classes les plus bruyantes…

Le test lui-même peut influencer le sujet par son apparence extérieure ou le matériel utilisé : tous les sujets ne sont pas encore familiarisés avec l'informatique et un test automatisé peut présenter un caractère intimidant. De même, les mises en situation ou l'habillage de certains problèmes peut réduire ou accroître la réussite de certains sujets, indépendamment de leur compétence, parce qu'ils sont familiers ou non de ce type de sujet. On peut par exemple observer ce genre de situation lorsqu'on prépare des tests d'embauche en rapport direct avec le contexte de l'emploi à pourvoir, alors même que certains candidats sont peu familiarisés avec cet univers (ce qui tend à favoriser les candidats internes par rapport aux candidats externes).

On devrait encore ici évoquer le rôle essentiel que jouent les personnes chargées de l'évaluation: leur présentation et l'image qu'elles donnent, le cadre général, l'accueil qui a été réservé aux sujets, la qualité de l'information préalable à travers les contacts téléphoniques éventuels, la convocation… Si le test est une situation standardisée, il convient au mieux d'en contrôler les différents paramètres, y compris en dehors de la situation immédiate de test. C'est en cela qu'il s'agit bel et bien d'un travail réservé à des professionnels.

On peut tenter de remédier aux variations existant entre individus et relatives aux dispositions mentales ou aux interactions avec la situation de test si on prend un certain nombre de précautions :

[12] Dans ce cas, on ne parlera plus de maturation, mais d'une dégradation éventuelle, liée à la sénescence.

. Identification des dispositions susceptibles d'intervenir.

. Structuration suffisante du test. Précision dans les consignes[13].

. Présentation adéquate des items. Exemple : la réponse correcte doit être présentée aléatoirement dans différentes positions de manière à éviter de faciliter les déductions et les choix construits sur d'autres bases que la compétences à mesurer.

. Formulation correcte des questions (voir par exemple Leclercq, 1986, pour les questions à choix multiples) et utilisation d'un système de correction fiable dans le cas du recours à des questions à réponses rédigées.

. Utilisation d'une formule adéquate de correction pour choix au hasard et information des sujets testés.

. Mise en condition des sujets, accueil correct, positionnement confortable et adéquat dans la salle de test. Lorsqu'il s'agit de tests collectifs, vérification des conditions optimales pour chacun des sujets.

. Recours à d'autres instruments (par exemple, l'observation en milieu naturel) lorsque des biais trop importants sont susceptibles d'invalider les résultats de tests.

Bibliographie

Hardy, J.L. (1983). Plusieurs solutions de recouvrement dans la corrélation entre un test et un items dichotomique. *Scientia Paedagogica Experimentalis. XX,* 28-50.

de Landsheere, V. (1988). *Faire réussir, faire échouer. La compétence minimale et son évaluation.* Paris : Presses universitaires de France.

Leclercq, D. (1983). Confidence Making: Its Use in Testing. In B.H. Choppin, T.N. Postlethwaite (eds.). *Evaluation in Education. 6(2).* 161-287.

Leclercq, D. (1986). *La conception des questions à choix multiples.* Bruxelles: Labor.

Leclercq, D., Denis, B., Jans, V., Poumay, M., Gilles, J.L. (1998). Chapitre 7. L'amphithéâtre électronique. Une application: le LQRT-SAFE. In D. LECLERCQ (éd.) *Pour une pédagogie universitaire de qualité.* Sprimont (Belgique): Pierre Mardaga éditeur. 161-186.

Monseur, C., Demeuse, M. (1998). Apports des études internationales à la réflexion sur la qualité d'enseignement nationaux: une analyse de l'éducation scientifique en Communauté française de Belgique. *Bulletin de la Société Royale des Sciences de Liège, 67(5),* 261-280.

Wu, M.L., Adams, R.J., Wilson, M.R. (1998). *ACER ConQuest. Generalised Item Response Modelling Software.* Melbourne: Australian Council for Educational Research (ACER).

[13] Plus les consignes sont ambiguës, plus le sujet devra interpréter et donc, plus on introduira des dimensions différentes et des effets parasitant.