CHAPITRE I
ÉLÉMENTS DE DOCIMOLOGIE

La docimologie ou science des examens est relativement récente. Elle doit le jour aux travaux d'H. Laugier et H. Piéron[1]. Elle s'est développée en France, en Belgique, aux États-Unis, au Québec, dans les facultés des sciences de l'éducation et les ministères. Depuis peu, sous le nom d'édumétrie, elle tente de prendre de l'extension dans une direction plus formative, par l'analyse d'unités d'apprentissage comme la question à choix multiple (Q.C.M.). Elle est en passe de devenir la science de la mesure en éducation[2].

Dans les écoles et les lycées, la tendance est encore de penser qu'elle ne touche que les mathématiciens ou les éducateurs férus de statistiques. Et pourtant tous les enseignants n'ont-ils pas affaire à des listes de résultats d'examens ou de tests? Ils deviendraient plus aisément docimologues si la docimologie, abandonnant ses habitudes langagières mathématiques, abstruses aux non initiés, se laissait traduire en un langage plus accessible[3].

Présentons les choses concrètement. On réunit, disons, une soixantaine de Q.C.M. dans un questionnaire qu'on administre à un groupe d'une centaine d'usagers représentatifs de la population visée. Les réponses saisies peuvent entrer dans un tableau à deux dimensions à raison d'une ligne par individu et d'une colonne verticale par question.

---

1. Cf. H. Laugier, *Études docimologiques sur le perfectionnement des examens et concours*, Paris, Conservatoire National des arts et métiers, 1934, 88p. H. Piéron, *Examens et docimologie*, Paris, P.U.F., 1963.

2. Voir notamment D. Leclercq, *Qualité des questions et signification des scores, avec application aux QCM*, Bruxelles, éd. Labor, 1987, 174p.

3. Sont ici présentés les travaux d'une équipe de recherche, à l'Université de Montréal, composée notamment de Serge Normand (Service Pédagogique), de Norman Molhant (ÉcoSystématica, société pourvoyeuse de logiciels sur mesure), de Michaël Strobel (professeur à la Faculté des sciences de l'éducation). Les premières pages de ce chapitre sont un résumé de l'enseignement de M. Strobel. Le modèle logistique qui suit doit beaucoup aux travaux de N. Molhant (nwm@cafe.edu).

```
RISBERG | 314+344++33233324132+12++3++323+3324+1113242++12441442211113
VERHAEG | 31314244+342432132244212424412244322423143343221234242342121
LECUYER | 4++1223+234+3+3222132321122+112421+411313122441241414424424 4
BOUMAZA | 113334441323432212132323244424344241223332433332323242344422
THORINS | 3+213433313242224223212133311344222131323124122121434234142-
POMEL V | 12413-41132-24-24233432-42333432414122-33211423-341223-2-424
ESPIAGO | 1121444+4313+3324224+22232113123222344113324421232314211344 4
BELGHER | 444142411313-321+2144112131111313314422134224331141411232432 1
BOUCHET | 4234234++31242+1421323211243334144242111424444421211142441 24
TENFICH | 34343342233224343413443321432123423132313244332441133321232 4
HARCETE | 1141314211--33324233-242224434+334412--232-4--2-4213-23414--
MAHUREV | 11423122+14+433242132142221434331441444 14-44-4221123-231+-4-
NEXERER | 34+2244+133442443233441244224 33-444123214444433244132234242 4
RIPOLLE | 1241324+134442221212232323143243131122113242432212114213341 2
ALLAMER | 3133324++433423442134322333211 2-112-413332444222424142243222
LAUGAAM | 34413243134++212113321331322332312113424412443222441441222 43
```

Ce tableau contient non seulement les numéros de réponse des individus à chaque question, mais des abstentions (-) quand on reconnaît ne pas savoir, et des rejets (+), quand on estime la question mal posée.

Ranger les données dans un cadre, les voir en deux dimensions, c'est déjà se faire une idée plus précise du genre de problème que pose l'analyse des phénomènes de langue envisagés sous l'angle de la statistique. Voyons ce que représentent les lignes, horizontales, et les colonnes, verticales.

À gauche, verticalement, les données sont précédées des titres de lignes (les noms des répondants). Au sommet, les données sont normalement surmontées des titres de colonnes (ici, la ligne de titre des colonnes n'apparaît pas, faute d'espace pour les numéros de question dans le questionnaire, qui vont de 01 à 60). Chaque «donnée» est donc identifiée, par sa position : réponse de l'individu N à la question x. La donnée consiste en un seul caractère : 1, 2, 3, 4, -, +, puisqu'il s'agit des quatre choix de réponses d'une question fermée, à quatre distracteurs[1], plus l'abstention ou le rejet.

Chaque nom précède ainsi une ligne de numéros qui sont les réponses choisies par l'étudiant, pour une série de questions, dans leur ordre. L'intérêt de ce tableau est qu'il montre toutes les données telles quelles, simplement, sans leur faire subir encore aucun traitement.

L'information recueillie par un questionnaire expérimental est donc constituée au départ d'une masse de détails : un nombre x de choix pour un nombre y de répondants. La

---

1. On appelle **distracteur** une réponse proposée en vue d'écarter (de distraire) le répondant de la réponse la meilleure, car celle-ci ne doit être retenue que par ceux qui l'identifient distinctement, sans deviner (*guessing*). Notons d'emblée, toutefois, que nous ne fixons pas d'avance la solution (réponse à considérer comme bonne). Nous considérons plutôt le questionnaire pédagogique comme une sorte d'enquête. Dans ce cas, toute réponse vraisemblable devient une solution potentielle. Autrement dit, tous les choix, au moment du choix, sont à égalité et pourraient s'appeler distracteurs (comme c'est le cas pour **distractor**, en anglais). La détermination de la norme validée ne doit plus se faire a priori (voir plus loin).

multiplicité et la probable complexité sous-jacente de cette information exigent un traitement, mathématique et interprétatif. Pour réduire la masse des données, il s'agit d'extraire des indices généraux, par exemple la note obtenue par chaque étudiant, puis la moyenne de ces notes, mais aussi le nombre d'étudiants à avoir choisi telle réponse plutôt qu'une autre, la moyenne des « bonnes » réponses; etc.

Les choix de réponse.

ANADIST commence comme tout logiciel de statistiques par remplacer les réponses par leur valeur présumée (1 pour ce que les rédacteurs considéraient comme la meilleure solution, 0 pour la pire erreur, > et < pour les deux réponses de valeur intermédiaire[1]). Il additionne ensuite les valeurs par ligne (résultat de l'étudiant) et par colonne (difficulté de la question).

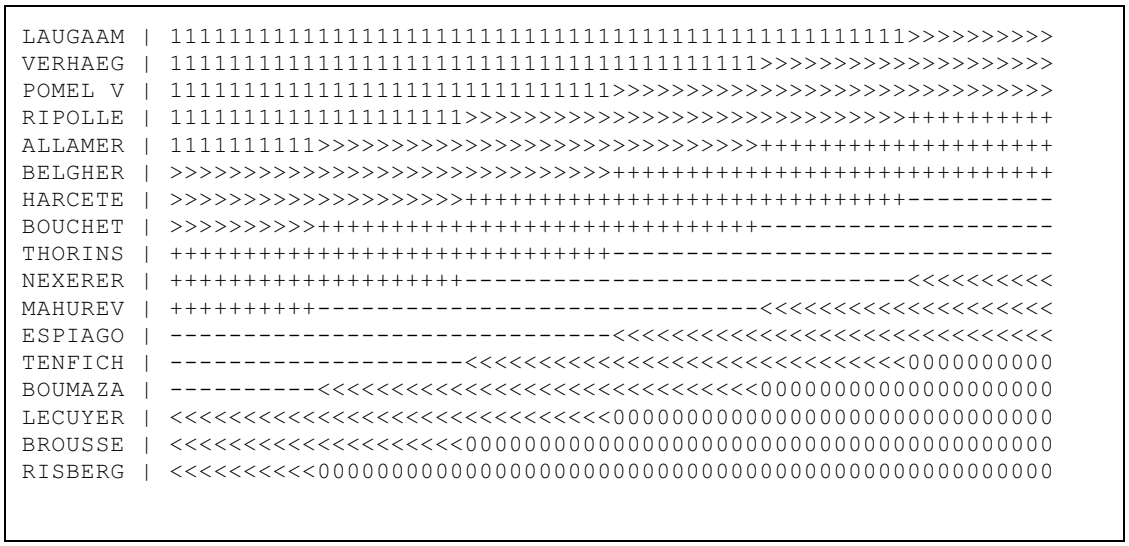
Maintenant, un brin de ménage pour y voir plus clair[2] : on trie les lignes suivant les résultats (les plus habiles en haut) puis on trie les colonnes (les questions les plus faciles au début). Ainsi les bonnes réponses se trouveront-elles concentrées dans le coin supérieur gauche. Réciproquement, le coin inférieur droit rassemblerait une majorité d'erreurs.

Si la connaissance, dans un groupe, était parfaitement homogène, c'est-à-dire strictement corrélative à l'habileté des répondants comme à la difficulté des questions, non seulement toutes les moins bonnes solutions et tous les moins bons étudiants se trouveraient dans la partie inférieure droite, mais les frontières entre les meilleures solutions, les presque meilleures, les rejets, les abstentions, les presque pires et les moins bonnes seraient des lignes de séparation nettes et sans ambiguïté... comme c'est le cas dans le tableau, fictif, ci-dessous.

---

1. Il s'agit des valeurs provisoires attribuées par les correcteurs. On verra plus loin comment ces valeurs se modifient pour rejoindre progressivement une échelle qui reflète le jugement du groupe des répondants.

2. La procédure qui suit est empruntée à Michaël Strobel. Sur la permutation des lignes et des colonnes d'une matrice de réponses et les «méthodes visuelles», cf. Benzécri, t.1, p.82-3.



On peut constater que LAUGAAM est passé au premier rang vu la majorité de bonnes réponses qu'il avait, tandis que RISBERG a un maximum d'erreurs graves. Si nous avions indiqué les numéros des Q.C.M. au sommet de chaque colonne, on verrait qu'elles ne sont plus dans l'ordre 123456789... mais dans l'ordre de leur difficulté croissante. La première colonne a le plus grand nombre de 1 et le moins grand nombre de 0, et ainsi de suite.

Naturellement, il existe quantité de causes de distorsions : le «guessing[1]» par exemple, c'est-à-dire le fait que ceux qui ne savent pas quoi répondre osent prendre le risque de se tromper, et courir leur chance plutôt que de s'abstenir; mais aussi le peu de validité de certains distracteurs, les choix offerts ne convenant pas toujours à des sous-groupes délimités par leur compétence spécifique; et encore l'hétérogénéité possible du groupe, dans lequel tous les niveaux de compétence ne sont pas nécessairement représentés. Les frontières sont d'habitude floues, apparemment tout à fait floues, à un point tel qu'il est souvent presque impossible de les reconnaître visuellement.

Un statisticien (Cronbach) a pu calculer le degré de cohérence des matrices de réponses aux tests : la consistance des frontières, en quelque sorte. L'indice alpha de Cronbach est égal à 1 pour des diagonales parfaites. En revanche, si le nombre de bonnes réponses, en dépit des remaniements dans les rangs des lignes et des colonnes, restait aléatoirement réparti, cet indice vaudrait zéro. En pratique, en dessous de 0.50, ou bien on s'est trompé de grille de correction, ou bien le groupe est composé de sous-groupes hétéroclites, ou bien le questionnaire mesure des choses trop différentes les unes des autres.

Le « Cronbach » –– comme on dit entre docimologues –– est un indice global remarquable. On peut le considérer comme une mesure de la convenance réciproque entre

---

1. Terme technique, le plus souvent non traduit. D. Leclercq distingue habilement réponse « au hasard » et réponse **devinée**. Voir *Qualité des questions et signification des scores* , p.93 et 94.

un groupe et un questionnaire. Supposons, par exemple, qu'un changement dans la clé de correction ait pour effet de faire monter le Cronbach : cela voudrait dire qu'aux yeux du groupe ce changement rend la clé plus conforme à la norme qu'il reconnaît.

Comme tout logiciel de statistiques, ANADIST fait aussi la somme des résultats, il la divise par le nombre des répondants pour obtenir une moyenne, il la met en pourcentage, il calcule les écarts de chaque répondant par rapport à cette moyenne du groupe, il fait la somme de ces écarts et la divise par le nombre des répondants pour obtenir l'écart moyen (écart moyen à la moyenne!). Celui-ci est un indice de la dispersion du groupe. L'écart-type, plus connu, même des non-docimologues, est plus long à expliquer mais il remplit à peu près les mêmes fonctions et n'est pas très différent quant à sa valeur numérique.

Intérêt de l'écart-type.

Au lieu de compter les «fautes» ou de transcrire une note proportionnellement (sur 10, sur 20, en lettres ou en pourcentage), l'enseignant qui effectue un contrôle de l'apprentissage réalisé peut mesurer l'acquis en se servant de l'écart-type comme unité. Ainsi, on situera un résultat en disant qu'il est à 1,3 écart-type au-dessus du résultat moyen et en écrivant +1,3. Au lieu de noter de zéro à cent, on a donc une échelle pour laquelle 0 ne désigne pas la nullité mais la moyenne des résultats du groupe; +1 est une note excellente et -1 une note faible (la limite de la réussite, normalement).

L'intérêt de la notation en écart-type est de fournir une échelle de mesure dite «pondérée», c'est-à-dire appropriée à la fois au niveau de difficulté du questionnaire et au niveau d'habileté du groupe. Par exemple, un étudiant qui aurait une note faible à cause de l'extrême difficulté de l'examen aurait, en écart-type, une note plus proche de la moyenne du groupe.

Il y a un possible inconvénient. En remplaçant la notation normative, avec son nombre de points préétabli pour chaque élément de contenu, par une notation pondérée (une échelle ajustée à l'habileté moyenne du groupe), le résultat de chacun se met à dépendre du groupe. Or le groupe est un critère valable dans ses propres limites, par exemple à des fins seulement formatives. Pour situer des notes ainsi obtenues dans un cadre plus large, par exemple à des fins évaluatives, il devient nécessaire de vérifier ce qu'on appelle la représentativité. Le résultat en écart-type est généralisable quand on peut s'être assuré que le groupe n'est pas marginal, par exemple qu'il n'a pas été réuni de façon prédéterminée artificiellement. Il doit avoir été pris sans aucun critère particulier et donc au hasard, dans l'ensemble de la population visée, dont il est, comme on s'en assure dans les sondages un «échantillon représentatif».

Dans la notation en écart-type, si la répartition suit une courbe «normale», on peut observer qu'entre 0 et +1 se trouve un bon tiers des effectifs. Entre 0 et -1, un autre tiers.

Au-delà de +1, il ne reste que 17% environ du groupe : les meilleurs. En deçà de -1, il ne reste aussi que 17% environ du groupe : les pires.

Incidemment, il faut remarquer que quiconque préfère une notation de 0 à 100 n'est pas obligé de s'en tenir pour autant à l'échelle arbitraire préétablie. Il est relativement aisé de retransformer une échelle pondérée. On donne par exemple au centre 0 la valeur de 50 (ou de 55, 60, la moyenne souhaitée pour le groupe) et on recalcule les notes en attribuant à l'écart-type une valeur de 15, 20, 25, 30 points selon la dispersion plus ou moins considérable que l'on veut voir entre les notes des individus. On obtient ainsi des résultats pondérés, normalisés[1], mais apparemment semblables aux notes en pourcentage traditionnelles, liées au seul jugement des responsables du système.

Tenant compte du groupe, la mesure en écart-type devrait être jugée plus sûre, mais souvent les élèves eux-mêmes s'en défient car elle ne favorise pas nécessairement ceux qui enregistrent le plus minutieusement les opinions de l'enseignant. On s'en tient donc plus volontiers à un système simple, où tout se décide d'avance.

Pourtant l'échelle en écart-type peut servir de mesure non seulement de l'habileté des répondants mais, réciproquement, de la difficulté des questions. Par exemple, une Q.C.M. dont la réponse prévue est le choix de répondants dont la moyenne est, disons, +1.4, aura une difficulté correspondante. Quand la moyenne des répondants qui ont pris la meilleure solution s'écarte de la moyenne générale de plus que la moyenne des écarts, la question est trop difficile ou trop facile. Si l'écart-type est supérieur à +2 ou inférieur à -2, la question est beaucoup trop difficile ou beaucoup trop facile. La précision de cette façon de noter nous l'a fait adopter pour les graphiques qu'on trouvera plus loin. Les écarts-types vont constituer la graduation de l'axe des habiletés, horizontalement. Les habiletés ne vont pas plus loin que 3 écarts-types car, déjà au-delà de 2, les courbes deviennent assez hypothétiques. Elles ne sont plus tirées des réponses attestées mais de leur prolongement théorique.

---

1. Appelés cote Z. Prenons comme exemple un groupe dans lequel un individu est à deux écarts-types au-dessus de la moyenne. Sa cote Z se calcule comme suit. Il faut d'abord choisir une valeur pour la moyenne du groupe, par exemple 50, et une valeur pour l'écart-type, par exemple 15. La note de +2 devient 50 + (2 x 15) = 80. L'individu dont l'écart-type est de -0.67 aurait une note Z de 50 - (0.67 x 15) = 39.5. On joue à son gré sur le point de départ et la dispersion en déplaçant la moyenne et les écarts. Avec une moyenne mise à 60 et un écart mis à 10, l'individu qui a +2 conserve sa note de 80 (60 + 2 x 10) mais celui qui a -0.67 passe à 60 - (0.67 x 10) = 53.3. Il a donc réussi! Et ce n'est que juste car il ne faudrait pas, "normalement", arrêter plus de 17% des étudiants. En effet, ceux qui sont dans les limites d'un écart-type (entre -1 et +1) sont dans le peloton.

Que valent les questions?

Quand on entend parler de moyenne et d'écart-type, on pense qu'il s'agit toujours des résultats des étudiants. Pourtant, dans le tableau où nous avons rangé les données, rien n'empêche de calculer deux moyennes et deux écarts-types suivant qu'on opère sur les totaux des lignes (ce sont les étudiants) ou sur ceux des colonnes (ce sont les questions). Pourquoi n'a-t-on pas l'habitude de faire les moyennes et les écarts-types des questions? Sans doute parce qu'on a généralement affaire à des matières à enseigner dont le contenu est indiscutable ou qu'on ne tient pas à le mettre en discussion, alors qu'en revanche, il fait partie de la structure pédagogique, de la relation maître-élève, de considérer les capacités des répondants comme foncièrement discutables.

Dans un domaine comme celui des langues en contact ou en évolution, au lieu de juger seulement les étudiants en les confrontant à des normes parfois livresques ou passéistes[1], on a avantage à tenter d'évaluer aussi la norme. C'est possible à partir des opinions, de préférence les plus récentes, des spécialistes (norme individuelle éclairée) ou à partir des capacités qui se révéleront dans un groupe (norme collective).

Si l'on pouvait disposer d'une échelle des habiletés solide et sûre, il suffirait de mesurer l'habileté moyenne de ceux qui optent pour chacun des choix. On aurait ainsi une clé de correction conforme à celle du groupe, avec une cohérence interne maximale pour ce test. Mais comment établir cette échelle des habiletés sans partir justement de ce qu'on cherche à vérifier : la valeur des réponses...

Il y a, dans les échelles d'habileté qu'on établit, des fluctuations considérables dès que l'on cesse de se fier à une norme arbitraire pour les «bonnes réponses». Telle est l'origine de la rigidité des normes culturelles, qui ne se modifient que par soubresauts. Les groupes sociaux ont besoin d'une norme indiscutable (penser au code orthographique) mais ils ne peuvent la puiser dans les groupes, où pourtant elle réside, faute d'une méthode assurée pour établir les habiletés. On se rallie à un code a priori. Toute évolution est assimilée à une faute. L'écart entre la réalité et la vérité officielle s'accroît, créant un malaise que le conservatisme fait grandir.

Or, le critère qui permettrait de sortir du cercle existe : c'est la cohérence interne, l'indice de Cronbach.

Le critère de cohérence.

Si la clé de correction initiale est trop écartée de celle qui correspondrait le mieux aux habiletés réelles, cet indice sera peu élevé. Au contraire, si la clé de correction utilisée est

---

1. Les aspects positifs de la défense du «bon» langage sont analysés notamment par Cl. Hagège, *le français et les siècles*, p.156-7 et passim.

conforme aux opinions du groupe, si elle convient aux capacités des répondants, cet indice sera élevé. Certains tests ont obtenu jusqu'à 0.96 (le maximum possible est 1.00). Les mêmes tests, pour des populations complètement différentes, ont pu tomber à 0.25 (alors que, au-dessous de 0.50, nous considérons les indices obtenus comme inutilisables).

Peut-on améliorer le Cronbach d'un test? Il faudrait arriver à modifier la clé de correction initiale (la valeur attribuée aux réponses par les rédacteurs) de façon qu'elle se rapproche de la clé validée sur le groupe, de celle qui résulte du jugement de tous les intéressés, dans la mesure où on peut le connaître. Quand il s'agit de langue, si les groupes sont représentatifs de toute la population, c'est dans cette direction que se situe l'évolution, l'actualité du rapport entre les formes et le sens.

Voici de quelle façon nous avons procédé. Certaines valeurs, initiales, en l'occurrence +1, +0.5, -0.5 et -1, ont été attribuées par l'enseignant aux quatre choix de réponses. Prenons l'exemple d'une question à choix multiple, rédigée à Bangui (République centrafricaine).

Je lance un appel à mes camarades _____ au village.
1 de retourner
2 pour retourner
3 à retourner
4 pour qu'ils retournent[1]

Le distracteur 4 a, au départ, la cote +1; le 2, +0.5; le 1, -0.5 et le 3, -1. On part de la clé professorale car elle se révèle le plus souvent une bonne approximation.

De l'expérimentation d'un nombre suffisant de Q.C.M. de ce genre, on tire une échelle des habiletés en procédant comme suit. Chaque répondant a obtenu un résultat. Celui-ci permet de calculer, pour chaque réponse possible, une valeur qui remplacera +1, +0.5 etc. Cette nouvelle valeur est la moyenne des résultats au test pour tous les étudiants qui ont choisi cette réponse. Une colonne ayant pour titre **Moyenne** figure donc dans un tableau de résultats (ci-dessous).

La question a été placée dans un questionnaire présenté à un groupe de finissants de l'École normale supérieure de Yaoundé. Voici le tableau des indices obtenus pour chacun des choix. Ils ne sont pas dans l'ordre initial mais dans l'ordre décroissant des moyennes qu'ils ont obtenues. Ainsi, on a d'abord la bonne réponse du groupe (qui est celle du sous-groupe supérieur), puis celle du deuxième sous-groupe, etc.

---

1. Un «corrigé» est remis par la suite aux interrogés. Il comporte la réponse considérée comme bonne et des contre-exemples. Voici le corrigé de la Q.C.M. ci-dessus.
Réponse pour qu'ils retournent
Mais (les autres choix de réponse auraient été de bonnes réponses sous la forme suivante) : pour attirer leur attention; pour **les** inciter à retourner...
Et On leur défend **de** retourner; on les incite **à** retourner au village.

Questionnaire AG16. Groupe Cameroun. Q.C.M. 25.

	Nombre	%	Moyenne	Z	Écart-type
Réponse 4*	56	70%	0.23	62	0.98
Réponse 1	8	10%	-0.07	59	0.61
Réponse 2	6	7%	-0.44	56	0.74
Réponse 3	6	7%	-0.72	53	0.78
Abstention	3	3%	-1.41	46	0.88
Rejet[1]	1	1%			

Les réponses apparaissent donc dans un ordre qui est celui des valeurs décroissantes des moyennes de leurs adhérents. Cette moyenne, en effet, est une bonne approximation de leur éventuelle validation (plus les étudiants habiles sont attirés par une réponse, plus elle a de chances d'être effectivement la meilleure). Rappelons que la cote Z est une transposition de la moyenne sur une échelle plus traditionnelle (allant de 1 à 100 plutôt que de -1 à +1). La colonne **Nombre** donne les effectifs des sous-groupes. Trois étudiants se sont abstenus et un seul a rejeté la question.

Pour faciliter la lecture et l'interprétation des indices, un astérisque (*) accompagne le numéro de la réponse présumée la meilleure.

On voit que se présente en premier lieu la réponse 4*. Elle est étoilée et donc elle est aussi la bonne aux yeux des rédacteurs, avant l'expérimentation. Le nombre des étudiants à avoir pris cette réponse est de 56, ce qui représente les 70% du groupe, qui compte 80 personnes. La moyenne des résultats à l'ensemble du test pour le seul sous-groupe de ces étudiants-là (les 70%) est de +0.23 (sur une échelle où 0 représente la moyenne du groupe et où +1 ou -1 sont l'écart moyen par rapport à ce 0). Pour la lisibilité, on reporte ce 0.23 sur une échelle traditionnelle (notée Z) où la moyenne serait 60 (la note de passage) et où l'écart-type serait 10 (ce qui donne six écarts-types entre 0 et 60 et quatre entre 60 et 100). Ainsi le niveau ressemble-t-il à une note d'examen, où la plupart des notes seraient situées entre 50 et 70. L'indice suivant est l'écart-type, toujours pour le sous-groupe des étudiants qui ont donné cette réponse-là. Ainsi peut-on voir s'ils sont regroupés ou non, du point de vue de leurs habiletés. Ici, l'écart-type de la réponse 4 est de 0.98, donc il est presque égal à celui du test dans son ensemble[2].

D'autres détails du tableau seront commentés plus loin, de même que la signification d'indices supplémentaires plus révélateurs. Pour le moment, nous avons entrepris d'exposer comment on peut tenter d'améliorer la cohérence interne des réponses d'un groupe à un test.

---

1. Rejet de la Q.C.M. par l'étudiant (il répond 5) quand il la considère comme inutile ou mal posée.

2. On sait que l'écart-type du test est noté **1** par définition (+1 ou -1). En effet, la moyenne des écarts est prise en-dessous et au-dessus de la moyenne générale (notée 0). Cet écart, divisé en centièmes, peut servir ensuite d'échelle de mesure pour les habiletés, pour chaque répondant ou chaque sous-groupe de répondants.

Des itérations?

La cohérence est maximale quand les réponses permettent de trier au mieux les compétences, c'est-à-dire quand se constituent des sous-groupes dont les traits sont bien distinctifs. Or le tri dépend des valeurs attribuées à chaque distracteur (puisque c'est par là qu'on peut évaluer les compétences). Et ces valeurs sont un peu grossières au départ (+1, +0.5, -0.5, -1) sinon arbitraires (elles reflètent le jugement des rédacteurs). Il suffirait de préciser ces valeurs en fonction du jugement du groupe pour que la cohérence augmente, comme on l'a vu à la page précédente. Peut-on aller plus loin? Existe-t-il une clé de correction qui maximisera la cohérence?

La moyenne des habiletés des adeptes de chaque réponse, que nous avons substituée aux cotes magistrales, c'était déjà des valeurs mesurées, donc une révision réaliste de la clé initiale. Toutefois, cette révision avait été possible à partir de moyennes tirées d'une correction effectuée avec la clé initiale : elle restait donc directement tributaire de celle-ci et elle n'en était pas très éloignée. Ce n'était donc qu'un pas dans la bonne direction.

Que serait la clé « réelle », celle du groupe, si on parvenait à l'établir indépendamment d'une clé initiale? Avant d'aller plus loin, soulignons que la clé initiale, celle des rédacteurs, constitue tout de même la meilleure approximation accessible. Elle offre plus de garanties, en tout cas, que n'importe quelle autre, qui ne serait pas moins arbitraire. Et elle est sûrement beaucoup plus proche de la clé mesurée que ne le serait une clé de départ purement aléatoire. Mais puisqu'il a été possible de l'améliorer en se servant des réponses effectives du groupe et d'un critère mathématique, donc sans intervention subjective unilatérale, que peut-on tenter pour aller encore plus loin?

On a donc maintenant une nouvelle clé, établie à l'aide des moyennes obtenues. Il s'agit déjà de valeurs beaucoup plus fines que la cote initiale (elles ont deux décimales) et surtout ce sont des mesures prises sur le terrain. On peut dire que la nouvelle clé est un compromis heureux entre les points de vue des deux parties en présence, l'enseignant et les enseignés. Il y a donc progrès. Progrès mesurable, d'ailleurs. Le test AG16, qui avait un Cronbach de 0.25 (très insuffisant), est monté à 0.48 par cette modification des valeurs. Une partie de l'écart entre la cote initiale et la cote qui permettra une cohérence maximale a dû se résorber. Mais cela ne suffit peut-être pas. Un Cronbach de 0.48 est à peine satisfaisant.

Pourquoi s'arrêter sur une bonne voie? Une fois les cotes modifiées, une nouvelle échelle d'habileté, assez différente, va s'élaborer. Le rang de certains répondants se modifie. Par la suite, cette échelle améliorée engendre à son tour de nouvelles moyennes... donc de nouvelles valeurs pour chaque distracteur. Résultat enthousiasmant. Le test AG16, en 9 itérations, avait atteint une cohérence interne tout à fait remarquable: 0.84. L'ordinateur

n'ayant pas à se faire prier pour recommencer les mêmes choses, nous lui avons fait faire jusqu'à cent itérations[1]. On arrêtait quand les modifications cessaient d'être significatives.

La première version du logiciel comportait quelques faiblesses. Dans les groupes disparates, les itérations rénouvelées pouvaient faire diverger les résultats. Ce défaut a pu être corrigé et les résultats ont fini par si bien converger que même en partant d'une clé de correction quelconque, totalement aléatoire, on aboutissait, à la longue, à des clés tout à fait similaires. Nous avions le sentiment de Christophe Colomb arrivant en vue d'un nouveau continent. La cohérence interne du groupe était donc un critère plus solide que tous les autres. Elle pouvait se laisser atteindre même en l'absence de clé professorale initiale. On entrevoyait la possibilité de décrire avec précision un ensemble de règles pratiquées d'un bout à l'autre de la francophonie, le "français commun d'aujourd'hui", et de repérer les divergences structurelles. Mais il allait falloir mener d'innombrables enquêtes, comme Gilliéron arpentait les campagnes pour faire la géographie linguistique de la France, une carte pour *abeille*, une autre pour *ive* ou *jument*, etc.

Jusqu'où fallait-il poursuivre les itérations? On adopta comme critère le moment où plus aucune Q.C.M. ne voyait l'ordre de ses réponses se modifier. « Solution stable! »

Le logiciel actuel, ANADIST, donne donc des clés de plus en plus proches d'un centre qui est constitué par le jugement du groupe. Il donne les valeurs pour lesquelles les réponses trient le mieux les compétences. D'itération en itération (elles reçoivent un numéro d'ordre), on peut voir les courbes se préciser, se redresser (et la raideur est un signe de meilleure discriminance). On peut aussi voir changer la validation de quelques-unes des réponses considérées d'abord comme bonnes, et inversement voir surgir des distracteurs qui finissent parfois par prendre la place de la solution qu'on croyait correcte...

À chaque itération est franchie une partie (environ la moitié) de la distance qui reste entre la clé en cours et la clé recherchée, idéale pour le groupe. Les indices de **moyenne**, sans cesse revus, se succèdent. Les jugements implicites du groupe sur les choix sortent de l'ombre tandis que l'incompréhension projetée par la loi quand elle est encore inassimilable (ou les divergences insoupçonnées avec l'enseignant) sont écartées provisoirement. Les variations du Cronbach (et d'autres indices, de fiabilité) sont là pour témoigner des progrès éventuels, pour les mesurer. Le test qui avait monté de 0.25 à 0.48 était à 0.58 à la troisième itération. Dès ce moment, un grand pas avait été franchi. On voyait un test qui paraissait inutilisable devenir pertinent à la seule condition de modifier partiellement, mais d'une manière bien précise, sa clé de correction.

Pour donner une idée de l'amélioration d'échelle obtenue par les itérations, voici un diagramme des habiletés (dans un test d'orthographe présenté à des lycéens de Metz, Dijon et Reims), d'abord avec la clé de correction professorale initiale puis après 89 itérations.

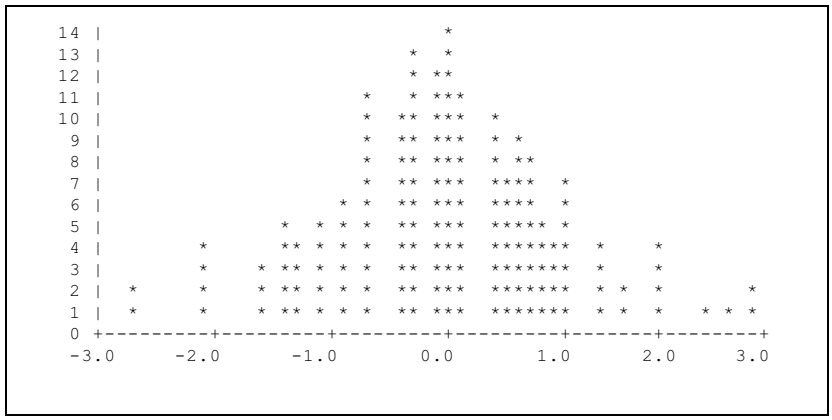
---

1. «Un cerveau humain ne peut accomplir une synthèse multidimensionnelle sans faire de nombreux choix arbitraires qui ôtent souvent toute signification au résultat. Il faut donc l'aide d'une calculatrice» souligne J.-P. Benzécri (*«l'Analyse des données*, t.1, p.16-17).

Distribution des habiletés.

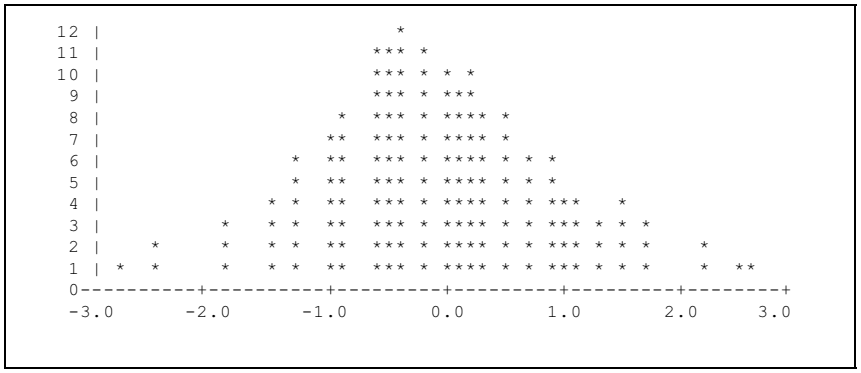
Avant toute itération[1], la première phase de correction donne des habiletés distribuées comme suit : on lit en abscisse (horizontalement) les habiletés représentées dans le groupe (entre -3.0 et +3.0, du minimum au maximum) sur une échelle en écart-type. Il y a 26 tranches d'habiletés[2]. Une seule des tranches est remplie à 100% avec ses 14 répondants. C'est la plus proche de la moyenne du groupe en habileté.

On peut voir ici une courbe « normale », en forme de cloche. Elle est assez bien représentée mais elle a de fortes irrégularités. Celles-ci diminueraient sans doute si l'on augmentait le nombre des répondants.



Toutefois, pour améliorer l'échelle elle-même, c'est à la clé de correction, révisée, qu'il faut toucher.

Voici ce que l'on obtient comme distribution pour le même test après 89 itérations. Mentionnons tout d'abord que le nombre d'étudiants retenus n'a pas changé[3]. A-t-on fait un gain, globalement, en cohérence interne ? Le Cronbach est passé à 0.66. Augmentation déjà notable (10%).



---

1. Test EF111-114. Groupe LYCÉES. Itération -1. Nombre d'étudiants : 143. Nombre de questions : 85. Cronbach : 0.56. Note minimale : 23; moyenne : 35.37; maximale : 51.

2. Elles ont été mises à sept centiles (sept centièmes du nombre de répondants).

3. Il est rarissime que les itérations écartent quelqu'un. Pour qu'un répondant cesse d'avoir une note mesurable, il faudrait qu'il ait bien répondu à toutes les questions validées (ou qu'il se soit abstenu à toutes celles-ci, ou qu'il les ait rejetées). Par contre, des hypothèses permettant le regroupement des déviances restent à développer. Les groupes ne sont jamais totalement homogènes et cela nuit à la précision des indices.

Les changements apportés par la clé revue et réadaptée au groupe à chaque itération entraînent comme conséquence qu'il ne reste que 62 Q.C.M. (au lieu de 85) à contribuer à l'établissement de l'échelle des habiletés. Ceci explique que la note minimale soit devenue 13. La moyenne est passée à 25.66 (en nombre absolu, car pour la moyenne pondérée, elle est toujours de 0, milieu entre -1 et +1; cela équivaut à 50 ou 60 entre 0 et 100 dans les cotations courantes). La note maximale est maintenant de 42. Cette fois, le maximum par tranche d'habileté n'est plus que de 12 répondants[1] bien qu'il y ait toujours 26 tranches d'habileté.

On voit que les courbes se sont améliorées puisqu'elles sont un peu plus semblables au modèle, et plus régulières; et que le Cronbach s'est élevé.

Toutefois, en s'approchant ainsi de l'opinion des groupes (chacun a sa vérité), on risque de perdre la rigidité du conformisme et d'accroître la multiplicité des positions possibles, selon la provenance, l'état de préparation et l'habileté des groupes. L'importance de la représentativité des échantillons par rapport à la population visée (pour la recherche ou pour l'enseignement) s'en trouve accrue.

Le choc des cultures.

Rien de tel qu'une bonne mise en situation historique pour clarifier la théorie. Dans quel cadre les réflexions ci-dessus se sont-elles mises à naître et à croître[2] ? Racontons-le brièvement, en faisant un léger détour dans le passé.

En 1961, je fus chargé d'un cours de rattrapage en langue écrite à l'Université de Montréal. À cette époque, le frère Untel avait lancé son pamphlet, exacerbant les tensions entre le français dit normatif, importé, valorisé par la partie internationaliste de la classe intellectuelle, et un français québécois, qui commençait à s'affirmer en littérature, et qui se faisait vilipender sous le nom symbolique de **joual**. Il fallait opérer sur le vif, délicatement, et le Belge d'origine à qui sa Faculté remettait le scalpel ne pouvait ni se fier à lui-même ou à ses connaissances acquises à Paris, ni trop céder aux revendications émanant de son nouveau milieu.

---

1. Ouvrons ici une parenthèse sur le décentrement du sommet par rapport à la moyenne (0.0). Il y a plus d'étudiants à droite du sommet qu'il n'y en a à gauche, donc trop ont «bien» répondu. Avec la clé de correction du groupe, le test est devenu un peu trop facile.

2. Circonflexes superflus depuis la Réforme...

Quel français enseigner? Où trouver la description d'une langue qui soit conforme aux pratiques effectives et aux vœux des futurs écrivants nord-américains? Il fallut peser le pour et le contre, y aller cas par cas, chercher des critères qui résistent aux critiques, contradictoires, de relâchement ou de colonialisme (culturel).

Je commençai comme tant d'autres par des tests d'orthographe sous forme de questions à choix multiple. Des logiciels de correction de tests, récemment venus des États-Unis, firent bientôt ouvrir des yeux ronds à ma gendelettrerie. Par les indices fournis, il était parfois quasiment possible de constater *de visu* ce qui se passait dans l'esprit des étudiants. Jamais, malgré des heures de discussions approfondies dans les classes, je n'avais pu saisir aussi nettement les causes et les conditions des difficultés éprouvées. Des choses apparemment complexes étaient bien assez connues, ne demandant pas plus ample commentaire. Des choses même simples, au contraire, que ce soit à cause de l'influence de l'anglais ou de l'isolement culturel d'avant la Révolution tranquille, offraient des obstacles à mes yeux inexplicables. Il devenait possible de se poser les questions réelles. Et surtout, les clés de correction étant validées, je pouvais exposer avec conviction les points sur lesquels la tendance foncière du groupe me donnait d'avance un accord de principe, éviter le risque d'augmenter la confusion inhérente au conflit des langues en contact.

Au début, ces indices me permirent de faire une mise au point de mon enseignement et de l'adapter aux besoins ressentis par le groupe. Peu à peu, cependant, en adepte de la phénoménologie, je fus amené à «reformer et reformuler» (comme dit Merleau-Ponty) cette perception. La relativité de la connaissance dans l'histoire des sociétés devenait autre chose qu'un sujet de discours : nous étions face à face et elle était mesurable. Elle prenait une dimension sociale, « intersubjective », parfaitement tangible. Cela modifiait non seulement l'approche didactique mais la vision du phénomène de la langue. Il fallait y inclure les collectivités, leurs échanges, leurs situations respectives, leur évolution.

Car les bonnes réponses validées par les groupes étaient des synthèses minimales bien partagées, justifiées dans leur cadre. À travers les épineux problèmes de langue, elles reconstituaient un **français correct «d'ici»** (au Québec), aussi conforme aux vœux des enseignants, qui commençaient à s'organiser (l'Association québécoise des professeurs de français), qu'aux proclamations du Ministère de l'Éducation. Il ne s'agissait ni de joual ni du français d'aujourd'hui, mais d'un sous-ensemble de la langue soignée, d'un système cohérent aux yeux de Nord-Américains fidèles à leurs origines du Grand Siècle, et décidés à les préserver par dessus tout de l'anglicisation.

Un exemple? *Le trappeur se demanda quelle pouvait être l'impression produite par le monde des _____ sur les animaux*. (**hommes** ? **humains** ?) Environ la moitié des personnes interrogées ont opté pour **humains**, à Montréal comme à Paris, mais à Montréal,

ce 50% se place parmi ceux qui ont le mieux répondu au reste du test (discriminance[1] de +0.16 contre 0.00 à **hommes**) alors qu'à Paris (et c'était aussi notre avis personnel) les plus habiles répondent **hommes** (discriminance de +0.35 contre -0.46 à **humains**). Voici donc un point où les valeurs d'emploi de *homme / humain* ont bougé au Québec. **Homme** étant un substantif, **humain** un qualificatif, employé comme substantif, on peut avancer l'interprétation suivante : au Québec, on voit dans ce qui caractérise notre espèce une différence moins de nature (de substance) que de qualité; et ce glissement n'est pas perçu comme un anglicisme (bien qu'il puisse venir, comme idée ou comme usage, du monde anglo-saxon).

Le C.A.F.É.

Avec les années[2], naquit le Cours autodidactique de français écrit (C.A.F.É.), qui se répandit «à distance» (par la poste) sur tout le territoire québécois. Outre la validation des Q.C.M. que des milliers de secrétaires se mirent à éplucher dans les cahiers d'exercices, le cours comportait une graduation (tri sur difficulté croissante) et une individualisation (chacun commençant au niveau d'habileté qui est le sien).

La méthode pourrait se résumer comme suit. Parmi les fautes à corriger courantes, il y a un sous-ensemble de points sur lesquels le groupe en général est d'accord implicitement (ce sont les Q.C.M. validées). Ces points peuvent être enseignés en priorité à ceux qui ne les maîtrisent pas. Les expérimentations préalables permettent de les identifier, de leur donner un niveau sur une échelle progressive, de les présenter dans un ordre d'apprentissage mesuré. Un simple cahier d'exercices peut alors suffire à assurer le progrès de celui qui est animé de la volonté d'avancer par lui-même[3].

Le processus de montage de tels ensembles didactiques, une fois les questions à choix multiple expérimentées, était le suivant. 1. Vérifier le Cronbach. 2. Écarter les Q.C.M. dont la bonne réponse prévue n'avait pas un taux positif de discriminance. (En effet, si la bonne réponse prévue est choisie par les moins habiles, elle forme une strate de compétence inférieure en ce qui regarde l'ensemble du test; elle ne peut plus être une bonne réponse pour l'ensemble du groupe.) 3. Trier les Q.C.M. validées suivant leur difficulté croissante (cet indice n'était autre, aux débuts, que le pourcentage de bonnes réponses). 4. Bâtir les cahiers gradués en répartissant les Q.C.M. selon leur niveau de difficulté et leur sujet. On fait normalement dix chapitres d'environ 100 Q.C.M. par cahier : orthographe, morphologie, accords, syntaxe et vocabulaire. 5. Réserver, dans chaque chapitre, dix Q.C.M.

---

1. La discriminance ou sélectivité est élevée quand ceux qui ont pris cette réponse-là sont en majorité parmi les meilleurs répondants (et inversement). Une Q.C.M. est sélective quand elle effectue à peu près le même tri parmi ses répondants que le test tout entier. Voir p.25.

2. Et sous la vigoureuse impulsion de M. Jacques D. Girard.

3. Il serait plus exact de dire **celle qui est animée**, les deux tiers des inscrits étant du sexe féminin.

représentatives du reste et très discriminantes afin de constituer un test de cheminement. En passant le test dès son inscription, l'étudiant reçoit une feuille de route imprimée par l'ordinateur, qui lui indique où commencer, pour chaque chapitre, suivant son habileté au départ.

Tel fut le C.A.F.É. des années soixante-dix, encore achalandé après vingt-cinq ans. Il représentait un progrès car les divergences entre cultures et subcultures avaient été prises en compte. Du français, on n'enseignait plus aux Québécois qu'un sous-ensemble approprié, selon leur compétence mesurée; et, de ce sous-ensemble, chacun ne recevait que la partie nécessaire, c'est-à-dire le programme individuel composé par ordinateur à la suite de son test de cheminement. Une synthèse se faisait jour, identifiable dans les réactions collectives. Devenaient accessibles au grand public les points de langue (pas seulement l'orthographe) les plus valorisés dans la population visée.

L'objectif était limité, au départ, mais la nouvelle approche ouvrait trop de possibilités. La linguistique avait donc affaire aux grands nombres... Eh bien, les grands nombres suivent des lois, qu'étudie le calcul des probabilités. Or les docimologues américains venaient justement d'intégrer ce type de calcul à leurs préoccupations.

La connaissance de la façon d'apprendre allait s'approfondir.

Le phénomène d'apprentissage.

La recherche de notre équipe[1] se tourna vers les diverses formules docimologiques et explora surtout celles «du trait latent». Il s'agissait de repérer dans la dispersion naturelle des réponses la ou les dimensions les plus caractéristiques. Ce type de modélisation devait aider à l'affinement des mesures de niveau (habileté ou difficulté) et d'écart-type. Nos spécialistes essayèrent la formule de Rash (à un paramètre) et celle de Birnbaum[2] (à deux paramètres).

Reprenons une vue d'ensemble du problème. À partir d'ici, ce sont les graphiques qui parsèment le présent ouvrage qui vont être explicités.

On cherche donc à savoir comment un apprentissage se réalise. On examine selon les méthodes de l'analyse statistique le comportement d'échantillons représentatifs, suffisamment variés, d'une «population».

Étant donné que les chances de bien répondre croissent avec la compétence, il ne s'agit plus d'une distribution comme celle de la fréquence des résultats ou celle de données aléatoires, avec la traditionnelle courbe en cloche. Les «bonnes réponses», et même les

---

1. V. p.5, note 3.

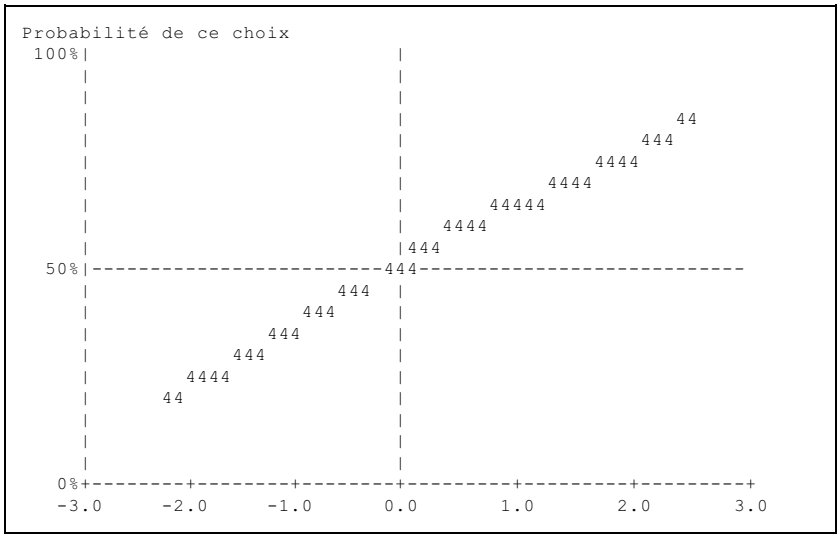
2. V. Lord et Novick, *Statistical Theories of Mental Tests Scores*, p.357 et sv.

autres ont quelque chose de commun avec la compétence mesurée dans le groupe, elles portent sur des savoirs ou des habiletés et vont former une courbe dite **logistique**.

En docimologie, la distribution logistique donne les chances de bien répondre (la probabilité des choix) en fonction des tranches successives d'habileté des répondants. La courbe prend place dans un espace à deux dimensions, qui sont les habiletés en écart-type, comme précédemment, en abscisse, et en ordonnées le pourcentage des adhérents au choix de réponse concerné. Quand on parle ici de pourcentage des adhérents, soulignons bien que ce n'est pas par rapport au grand total des répondants mais seulement par rapport au sous-total des répondants qui se situent dans la tranche d'habileté concernée.

Pour la clarté du déchiffrement de ces graphiques, nous avons carrément pris le numéro du choix, ici 4, pour tracer la ligne. Ce n'est que la représentation approximative, en caractères d'imprimerie, d'une courbe continue. De cette façon, chaque ligne est identifiée immédiatement. On peut observer que s'il y a plusieurs chiffres 4 de suite, horizontalement, ils ne forment pas un nombre : c'est simplement dû au fait que la courbe ne monte pas de façon raide et escarpée à cet endroit (ce qui signifie que le choix a moins de discriminance à ce niveau d'habileté-là, ce sera expliqué plus loin).

Les habiletés sont normalisées de la même façon que dans les graphiques examinés précédemment : elles ont 0.0 (zéro) pour centre (la moyenne du groupe) et vont de -3.0 (moins trois) à +3.0 (plus trois) écarts-types. Le pourcentage des adhérents va de 0% (aucun) à 100% (le maximum pour chaque tranche d'habileté).



Le moment intéressant est le milieu de la courbe, l'endroit où elle franchit la ligne de 50%. Cette ligne délimite la moitié du groupe et donc le moment où la minorité devient une majorité. Sur la courbe ci-dessus, on voit que 50% environ des adeptes de la réponse 4 sont d'une habileté moyenne. En effet, la courbe du distracteur 4 franchit la ligne des 50% en abscisse précisément au moment où elle atteint son degré moyen d'habileté, 0.0 en ordonnée. La courbe monte, ce qui veut dire que les étudiants les plus faibles (niveau inférieur à zéro donc indice négatif) sont moins de la moitié de leurs effectifs à prendre la réponse 4. Mais comme la courbe monte, moins ils sont faibles, plus ils sont nombreux. Au-delà de la moyenne (0.0), les étudiants forts sont plus de la moitié à prendre la réponse 4.

Plus ils sont forts, plus ils prennent cette réponse. Elle est donc plutôt bonne. Elle reflète un savoir du groupe.

Voici donc (pour le redire autrement) comment le tracé de ces courbes, qui révèlent des strates de comportement collectif, est obtenu. On trie les étudiants suivant leur note au test et on les répartit en sous-groupes, par tranche d'habileté. On regarde quelle proportion des effectifs de chaque tranche a pris chaque réponse (ici la réponse 4) et on place un point (ici le chiffre 4) sur le graphique, à l'intersection correspondante. Si les points ainsi définis esquissent un mouvement ascendant ou descendant, on arrondit la courbe de façon que l'ensemble soit le plus cohérent possible. On doit faire cela parce que l'échantillon d'étudiants dont on dispose n'est pas nécessairement homogène en habileté. Mais aussi, il est normal de le faire parce qu'on est en droit de supposer qu'il y a une corrélation entre l'habileté et le nombre de – relativement savantes – réponses : c'est la définition même de l'apprentissage.

Il est intéressant de mentionner ici l'analogie de cette courbe et de celle d'Ebbinghaus, qui note le nombre de succès à chaque nouvelle série de tentatives. Il y a dans les deux cas une accélération rapide au moment de la découverte de la tâche et un ralentissement progressif au moment où l'acquisition se rapproche du maximum possible. Ce qui distingue les deux courbes est que la courbe logistique prend le phénomène sous l'angle collectif alors que la courbe dite d'apprentissage d'Ebbinghaus relate les phases successives du travail d'un seul individu[1].

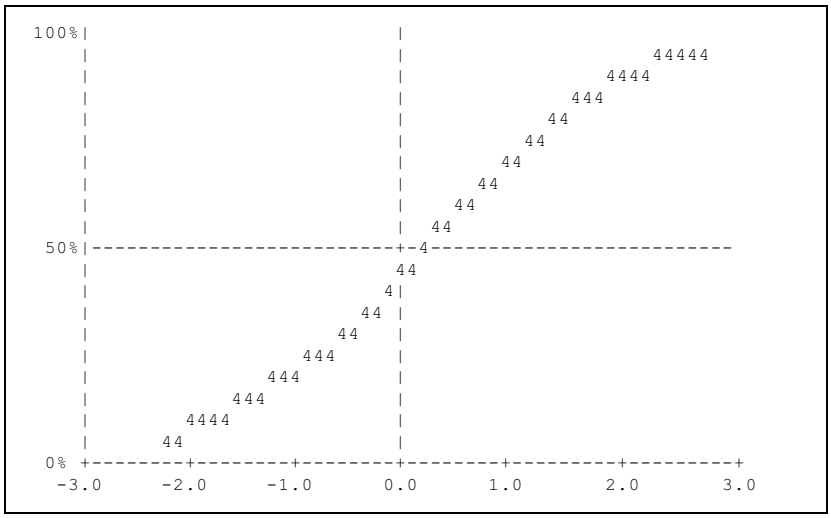
Remarquons que la courbe logistique peut concerner la bonne réponse mais aussi les distracteurs. Nous examinons les courbes de toutes les réponses possibles, au point de vue des « chances de donner cette réponse » selon le niveau d'habileté concerné. C'est méthodologiquement essentiel pour plusieurs raisons, tant théoriques que pratiques. Les distracteurs, dans l'apprentissage, ne sont des erreurs que si l'on se place au point d'arrivée, qui permet de les juger. Dans une perspective d'acquisition progressive, ils sont des ébauches successives, ils tendent vers une bonne réponse, qui est encore mais provisoirement hors d'atteinte. Ceci peut donc s'analyser dans le comportement du groupe. Il vaut la peine de l'aborder maintenant plus en détail et même techniquement.

Pour qu'une réponse ait une courbe logistique qui révélera un apprentissage, il faut (et il suffit) : primo, que ses adhérents soient plus rares dans les tranches d'habileté moindre (la gauche du graphique); secundo, que leur fréquence relative augmente de façon significative à un endroit donné de l'échelle des habiletés; tertio, qu'une forte majorité se manifeste au-delà de cet endroit où « tout change », et qui est le niveau de difficulté normalisé de cette réponse.

---

1. Les deux approches convergent du fait que l'ontogenèse (ou formation individuelle) est habituellement un raccourci de la philogenèse (ou formation de la tribu, historiquement). Autrement dit, on trouve dans le groupe, simultanément, toutes les phases successives possibles de l'apprentissage.

En pratique, il est rare que cette courbe soit quasi rectiligne comme dans le graphique ci-dessus. Elle est plutôt légèrement incurvée, ne prenant que progressivement son essor et ralentissant à l'arrivée au sommet.

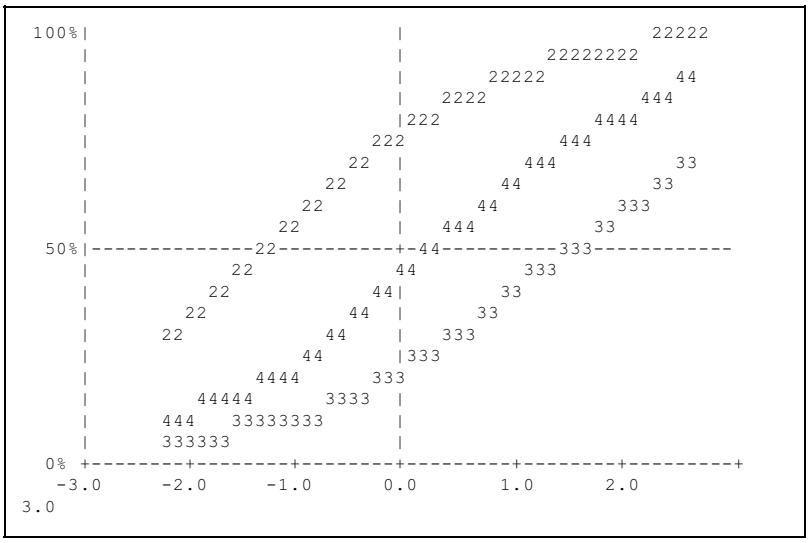


Ce que cette forme illustre, c'est que le savoir lié au distracteur 4 de cette Q.C.M. ne s'obtient pas aussi aisément , dans le groupe, à n'importe quel niveau d'habileté. Il est plus lent, ici, avant -1.0; plus rapide au moment où il atteint des effectifs de 50% (une majorité de connaissants); plus lent à nouveau quand l'habileté dépasse un seuil élevé (près de +2.00). L'explication de ce phénomène est assez simple. Il y a un degré d'habileté idéal pour l'apprentissage de chaque réponse. La courbe dévoile ce point à partir de l'habileté mesurée de tous ceux qui ont fait le choix considéré (qui n'est pas forcément le meilleur de tous mais qui doit être au moins meilleur que l'une des autres réponses données). C'est l'habileté correspondant à 50%, la moitié des effectifs, qui marque le plus exactement le niveau idéal d'apprentissage. Beaucoup plus bas (par exemple à une distance de 1 en écart-type), on apprend mal parce que la chose doit apparaître comme trop difficile. Beaucoup plus haut, on répugne à s'intéresser à quelque chose qui apparaît sans doute comme trop facile.

Il s'agit donc d'une courbe légèrement incurvée, appelée courbe « en S », qui se place grosso modo dans la diagonale du graphique. Elle se rapproche de la diagonale par son accélération (son raidissement, son escarpement) au moment de passer du bas (peu de réponses) vers le haut (beaucoup de réponses). La transition du bas vers le haut est acquise au moment où on atteint, parmi les répondants du sous-groupe qui a opté pour un distracteur donné (et qui forme une « **strate** ») une quantité, en comptant à partir des plus faibles, égale à la moitié, c'est-à-dire au moment où la ligne horizontale marquée 50% est franchie. Ce point correspond à un niveau d'habileté (mesuré en écart-type) qui est le « niveau d'apprentissage ». Le niveau d'apprentissage est d'une grande précision. Il est extraordinairement utile à la confection de cours puisqu'il indique à quel moment l'individu apprendra le plus aisément (et durablement).

Pour interpréter les graphiques, il est bon d'avoir à l'esprit cette forme : un S allongé dans la diagonale. Il arrive que les habiletés représentées dans le tableau ne soient pas exactement celles qui convenaient à la question. N'apparaît alors qu'une partie de la courbe en S : la base par exemple, ou le sommet. Il est aisé de les identifier, cependant, puisque la courbe s'arrondit en creux dans le premier cas, en bosse dans le second.

Des trois courbes ci-contre, laquelle est complète? laquelle est une base du S? laquelle est un sommet?



Le graphique est une visualisation des indices. Il donne une configuration tangible de la répartition des réponses. On a ici une courbe complète de niveau moyen (la réponse 4), une courbe qui est plutôt un début de S, une base et la moitié du reste (la 3, dont le niveau est 1.16) et une autre qui est un sommet avec la moitié de sa montée, car elle est de niveau faible (la 2, à -1.46).

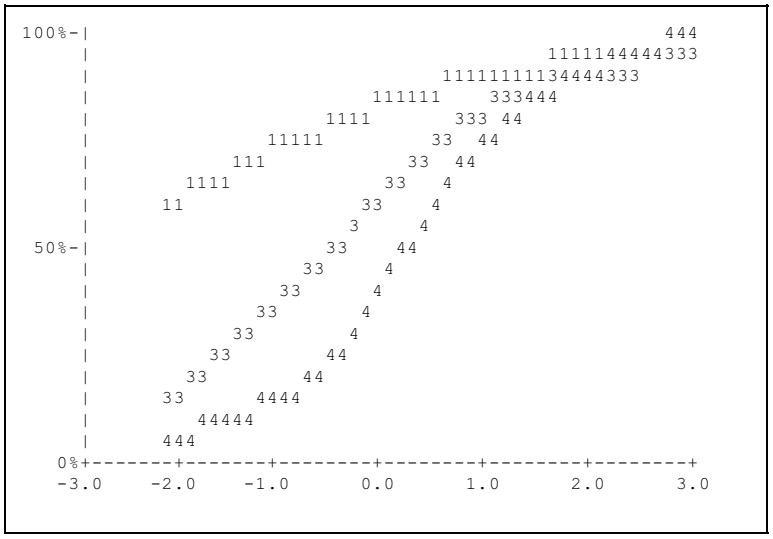
Normalisation du niveau d'habileté.

Y a-t-il un rapport entre cette courbe et celle des résultats, en forme de cloche? La courbe d'apprentissage ne va pas sans la courbe normale. Elle la présuppose en quelque sorte. Si l'on devait prendre le nombre absolu de réponses à chaque niveau, on ne pourrait jamais obtenir de courbe conforme à la théorie : les étudiants moyens restent plus nombreux que les forts. C'est proportionnellement aux étudiants présents qu'il faut déterminer le point à reporter sur le graphique.

Avec le temps, nous avons donc affiné la mesure du niveau de difficulté d'une Q.C.M., et ce à deux reprises. Ce n'était, d'abord, que le pourcentage des bonnes réponses. Par la suite, pour chaque réponse, nous avons eu la moyenne d'habileté de ses adeptes, mesure d'autant plus ajustée qu'elle pouvait tenir compte d'une échelle d'habileté mesurée dans le groupe. Maintenant, avec la projection d'une courbe d'apprentissage, s'offre la possibilité de compenser en partie les aléas de la variété des groupes censés représenter la population visée.

Verticalité.

Le niveau est un point central (comme la moyenne) et l'on aimera aussi savoir quelle est l'amplitude de la dispersion (comme l'écart-type tout à l'heure) pour cet événement collectif qu'est l'apprentissage. Voyons les choses de plus près sur l'exemple des courbes suivantes.



Laissons de côté la courbe de la réponse 1, qui est un sommet, et comparons la 3 et la 4. Cette dernière est très verticale. Sa discriminance est de 0.93 c'est-à-dire qu'elle départage presque parfaitement, verticalement, les faibles et les forts. Or plus une courbe est verticale, plus est significative dans le système d'expression du groupe la réponse qu'elle représente, mais aussi plus il importe de présenter cette Q.C.M. à un apprenant quand il a atteint et pas encore dépassé le niveau indiqué. On a besoin d'une mesure de la marge de manœuvre dont on dispose pour intéresser l'étudiant et lui rendre vraiment service.

Cette mesure n'est pas sans analogie avec l'écart-type. Elle en diffère toutefois parce qu'elle ne s'applique pas à tous les répondants : elle est prise sur la courbe et constitue donc une approximation vraisemblable de ce qui se passe à ce niveau.

Questionnaire EF118-20. Q.C.M. 36. Groupe LYCÉES. Itération 17.

		Nombre		**Courbe**	**Moyenne**	**Écart-type**	**Niveau**	**Sélectivité**
Réponse	4*	98	57%	B+	0.40	0.93	-0.24	0.93
Réponse	3	21	12%	B	-0.41	0.50	-0.89	0.68
Réponse	1	25	14%	B	-0.51	0.83	-2.82	0.40
Réponse	2	24	14%	P	-0.57	0.90		
Réponse	-	1	0%	P				
Réponse	+	1	0%	P				

On aurait pu mesurer la largeur de la tranche comprise entre les deux parties incurvées du S, ou donner le niveau en habileté du début et de la fin de la partie rectiligne, la plus verticale, de la courbe. Il a paru plus simple, aux mathématiciens, de prendre comme indice la pente de la courbe (l'angle formé par la courbe et par l'horizontale, à la moitié des effectifs). Tel est l'indice **de sélectivité**. On l'appelle plus souvent indice de **discriminance**, mais ce mot peut passer pour péjoratif dans la langue courante.

Quand la partie centrale du S est proche d'une diagonale (**sélectivité** autour de 0.36), les tranches d'habileté concernées sont assez nombreuses. La Q.C.M. est accessible à plus d'un sous-groupes. Mais quand le segment rectiligne du S est plus vertical, les tranches d'habileté «où l'on apprend» (où l'on a le maximum de chances de faire l'apprentissage de façon consciente et durable) forment une bande plus étroite. C'est le cas de la courbe 4 ci-dessus. Pour visualiser, on peut dire de l'indice de sélectivité qu'il donne la raideur de la courbe (sa verticalité).

Pour fixer les idées, on peut considérer le **niveau** comme une **moyenne** tenant compte d'une distribution sous-jacente qui soit « normale »; et la **sélectivité** comme une sorte d'**écart-type** normalisé, c'est-à-dire valable pour n'importe quel échantillon non biaisé.

Bonnes et mauvaises courbes.

Reprenons. Les courbes du graphique sont obtenues de la façon suivante. Pour chaque distracteur (choix de réponse), l'ordinateur va fouiller ses listes de réponses d'usagers (classés par tranche d'habileté) et il calcule combien (en pourcentage) l'ont choisi. On évalue d'après la pente à quelle tranche d'habileté cette proportion atteindra 50% pour mesurer un « niveau d'apprentissage ». Alors que les habiletés forment une courbe « en cloche », la courbe d'apprentissage est une oblique, puisque ce sont normalement les étudiants forts qui ont appris en plus grand nombre.

Évidemment, les courbes obtenues ne sont pas toujours incurvées de la bonne façon. Supposons qu'on ait affaire à un choix de réponse qui n'apprend rien et qui sert comme piège (distracteur au sens fort). Choisi par quelqu'un, il constituerait une faute dépourvue de toute signification (aux yeux du groupe). Il ne pourrait avoir un plus grand nombre d'adeptes dans les tranches d'habileté supérieures.

Il peut même arriver que la courbe des points mesurés soit inverse à celle des habiletés mesurées. Une telle réponse enseigne, pourrait-on dire, le contraire de ce que le reste du test enseigne. Le logiciel ANADIST l'indique (p.28) par la lettre **M** (mauvaise courbe), à gauche des indices. Un **B** signale une « bonne » courbe, qui indique que l'on apprend normalement. Un **B+** signale une bonne courbe qui a de plus servi à établir l'échelle des habiletés[1]. L'absence de courbe est indiqué par **P** (pas de courbe).

1. À l'itération précédente, la réponse était valide (elle avait une bonne courbe et sa moyenne était la plus élevée).

Niveaux hypothétiques.

La courbe d'apprentissage dessinée est parfois très peu incurvée, presque rectiligne (voir p.20). En vérifiant le niveau calculé, on peut remarquer qu'il se situe très en dehors du graphique : beaucoup plus haut ou beaucoup plus bas. Il arrive qu'il soit à 7 ou même 10 écarts-types de la moyenne alors que ne sont tracées que les portions de courbes qui vont de -3 à +3. L'aspect rectiligne vient du fait qu'on ne voit que la partie inférieure ou supérieure de la courbe, qui décrit la portion des strates où tout le monde ou presque ignore, ou connaît. Il faudrait agrandir la fenêtre que découpe le graphique sur le phénomène d'apprentissage dans le groupe pour voir la raideur au point le plus caractéristique, quand on franchit la ligne des 50%. Presque aucun de nos étudiants, d'ailleurs, ne se trouve à des endroits aussi éloignés de la moyenne que 3 écarts-types. De telles courbes sont donc des artifices mathématiques. Constatant que quelques étudiants dans le groupe peuvent esquisser un début ou une fin de courbe sigmoïde (en forme de S), l'algorithme (la formule mathématique) est appliqué. En pratique, tout ce qu'on peut tirer comme renseignement sur ces courbes trop étalées, c'est qu'il faudrait poser à nouveau la question à des groupes beaucoup plus forts ou beaucoup plus faibles, selon le cas.

Plusieurs bonnes courbes, donc plusieurs bonnes réponses?

Comment se fait-il que certaines Q.C.M. ne tracent qu'une courbe, ou même aucune, alors que d'autres en ont plusieurs?

Qu'il n'y en ait aucune vient probablement du fait que la Q.C.M., si elle mesure quelque chose, ne mesure pas la même chose que le reste du test. Et pour ne pas fausser la recherche des meilleures échelles d'habileté, on retire ces Q.C.M. des données retenues pour le calcul du Cronbach. (On les réintroduit plus tard, quand il arrivera que les échelles nouvelles permettent de tracer une bonne courbe avec une de leurs réponses.)

Qu'il n'y ait qu'une seule courbe, ce pourrait être la règle générale, à bien y penser. Une fois que plus de 50% des plus habiles ont choisi une réponse, qui trouve ainsi sa «bonne courbe», on voit mal comment une autre réponse pourrait rééditer la performance. Or il fallait trouver le moyen de faire apparaître le phénomène d'apprentissage à des niveaux d'habileté divers et pas seulement pour les meilleurs étudiants. Ce que nous avions pu observer, en effet, c'était qu'une mauvaise réponse pouvait assez souvent représenter une approximation intéressante. Par exemple, elle prenait, dans certains groupes, le comportement d'une bonne réponse si celle-ci était retirée des choix.

La solution adoptée fut, bien simplement, de simuler ce retrait. La réponse qui a obtenu une bonne courbe empêchant les autres d'en recevoir une à leur niveau, on la supprime provisoirement. Il n'est même pas nécessaire pour cela de soustraire du groupe les adeptes de cette réponse. Il suffit, en pratique, de mêler les sous-groupes des deux distracteurs concernés. Sur le graphique, le résultat est particulièrement clair puisque la

seconde courbe vient se placer au-dessus de la première, en sorte que les étudiants sont toujours représentés sur des points différents de l'espace, sans laisser de vide. On procède ensuite de la même façon pour les distracteurs suivants. La distance entre deux courbes, à quelque niveau que ce soit, représente le nombre des étudiants qui ont fait le choix indiqué par le chiffre (1, 2, 3, 4, -, +) formant la courbe qui s'ajoute (la plus haute). Comme les Q.C.M. sont corrigées dans l'ordre décroissant de la **moyenne** des réponses, les courbes ne peuvent se croiser qu'à leurs extrémités, ce qui ne dérange ni la lisibilité ni la vraisemblance puisque ce sont probablement des projections dès qu'on s'écarte à l'excès (plus de 2 écarts-types) de la moyenne générale.

Disposer ainsi les sous-groupes de chaque choix présente un intérêt tout particulier quand on met en présence plusieurs cultures ou subcultures, diversement valorisées par le groupe. Les strates intermédiaires mettent en évidence la présence de types de réponses qui pourraient parfaitement être les meilleures si le groupe n'était pas en train d'en valoriser d'autres (pour toutes sortes de raisons qu'il reste à inventorier).

Examinons un graphique.

Ayant terminé l'exposé des grandes lignes de la méthode suivie pour obtenir des graphiques sur le comportement des sous-groupes, nous pouvons maintenant examiner les résultats obtenus pour un certain nombre de Q.C.M. typiques. Voici les indices, le texte puis le graphique d'une question de linguistique qui présente une distribution très harmonieuse.

Questionnaire U0901, Q.C.M. 13, groupe Univ. de Montréal-Ét.Fr.

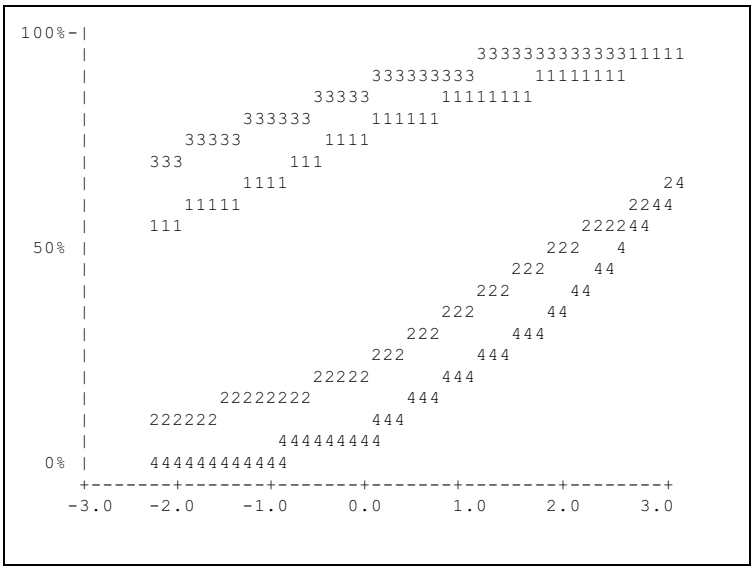
	Nombre		Courbe	Moyenne	Écart-type	Niveau	Sélectivité
Réponse 4*	28	21%	B+	0.92	0.90	1.56	0.64
Réponse 2	17	13%	B	-0.16	0.73	1.13	0.40
Réponse 1	59	45%	B	-0.17	0.86	-2.58	0.34
Réponse 3	14	10%	B	-0.26	0.96	-3.53	0.38
Réponse +	10	7%	P	-0.89	0.80		
Réponse -	2	1%	P				

La phrase se découpe en groupes de mots (ou syntagmes) dotés d'une fonction par rapport au verbe principal. Mais comment découper le syntagme ? Quels en sont les constituants immédiats ?

1 *Les mots grammaticaux*. 3 *Les mots lexicaux*.

2 *Les mots phonétiques*. 4 *Autre chose*

C'est une Q.C.M. de théorie grammaticale qui fait appel à des notions de linguistique déjà pointues[1]. **Syntagme** équivaut à **groupe de mots ayant une fonction** : la notion, sinon le mot, est courante dans les grammaires. **Constituant** et **immédiat**, en revanche, sont réservés au linguiste. Ils relèvent d'une méthode scientifique d'analyse[2].



Commentaire sur le graphique. Les plus habiles au test (**moyenne** : 0.92) prennent le choix 4 parce qu'ils savent que le syntagme est composé d'un mot lexical entouré de mots grammaticaux. Ils sont 28 et nettement pris surtout dans la tranche supérieure (raideur de leur courbe, **sélectivité** : 0.64). Vu que la Q.C.M. reçoit comme niveau d'ensemble celui de sa bonne réponse validée, celle-ci est fort difficile (**niveau** : 1.56).

Il est souvent intéressant de se pencher aussi sur les scores des distracteurs. La réponse 2 a une bonne courbe, très proche de celle de 4. Elle discrimine moins nettement (**sélectivité** : 0.40), ce qui est heureux pour l'enseignant, qui la considérait comme fausse. Le mot phonétique, en effet, est un segment délimité par des critères de rythme. Aux yeux du groupe, toutefois, la réponse 2 est presque valide (peut-être par le fait que le syntagme est souvent aussi un seul mot phonétique, avec un accent de longueur sur sa dernière syllabe).

L'intérêt du traçage de courbes d'apprentissage apparaît aussi dans la différence de niveau obtenue ainsi : 1.13 donc quelque chose de presque aussi difficile que la bonne réponse (alors que la simple moyenne des habiletés accuse, elle, une différence considérable : **moyenne** -0.16).

La courbe de la réponse 1 est révélatrice aussi mais à un autre égard. C'est une réponse deux fois plus prisée que la 4 (59 étudiants, 45%); sa **moyenne** est toute proche de celle

---

1. Voici le corrigé. *Réponse* Autre chose. Les mots (tout court).

Ou Les mots lexicaux, grammaticaux, syntaxiques, qui, le plus souvent, s'identifient à des mots graphiques (découpés par un espace typographique).

Mais Les mots phonétiques regroupent des syllabes et constituent une unité de rythme prosaïque. Ex. : Tous les jours (1) à la même heure (2), le maître d'école (3) ouvrait (4) les auvents (5) de sa maison (6).

2. On découpe des segments qui peuvent se réunir « immédiatement », « constituant » ainsi des segments étendus d'une nature distincte (les syllabes en mots, par exemple, ou les syntagmes en assertions).

de la 2 (-0.17) mais sa courbe place presque en dehors du tableau le niveau d'habileté requis de ses adeptes : -2.58! Ceux qui ont pris la réponse 2 ne savent pas grand-chose du syntagme, ils se doutent seulement que ce doit être quelque chose de grammatical... Quant à la réponse 3, elle offre les mêmes caractéristiques aggravées car elle semble impliquer une confusion entre grammaire et lexicologie.

Même les abstentions peuvent être validées.

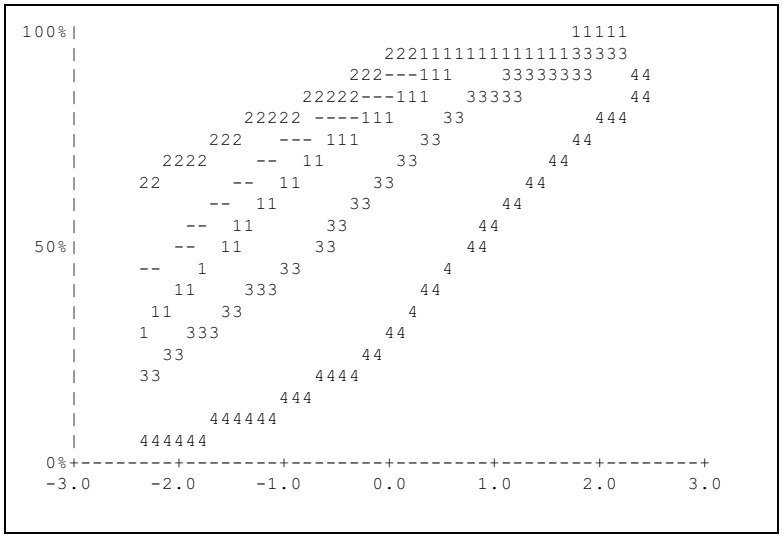
Voici une autre Q.C.M. prise dans le même test et qui se comporte de façon assez semblable[1].

Un code de transcription idéal propose un son par lettre, et une lettre par son. Où trouve-t-on un tel code?

1 *En français.*
2 *En anglais.*
3 *En latin.*
4 *Autre chose*.

	Nombre		Courbe	Moyenne	Écart-type	Niveau	Sélectivité
Réponse 4*	55	42%	B+	0.55	0.87	0.41	0.71
Réponse 3	36	27%	B	-0.17	0.91	-0.96	0.64
Réponse 1	18	13%	B	-0.38	0.81	-1.70	0.70
Réponse -	6	4%	B	-0.38	0.79	-2.16	0.60
Réponse 2	7	5%	B	-0.71	0.66	-3.30	0.44
Réponse +	8	6%	P	-1.23	0.58		

Pas moins de cinq courbes! Même les abstentions sont validées (-). Seuls les rejets (+) n'ont pas de courbe significative (P). La majorité a pris la bonne réponse cette fois, ce qui rend donc la Q.C.M. beaucoup moins difficile (**niveau** : +0.41). La réponse 3 appartient à ceux qui savent combien le français et l'anglais ont l'apanage des incohérences et des bizarreries dans les relations,



1. Questionnaire U0901, Q.C.M. 8. Groupes Univ. de Montréal-Ét. Fr. 7 à 9 (département d'Études françaises, années 1987, 1988 et 1989).

si archaïques, qu'y entretiennent l'orthographe et la prononciation[1]. Ils supposent que le latin est plus logique, ce qui n'est pas entièrement vrai, et ils ne songent pas à des codes artificiels. Ceux qui optent pour l'anglais sont plus faibles encore que ceux qui croient à la logique du français; et de fait, l'anglais est pire encore dans ses divergences entre la graphie et les sonorités. Pourquoi les abstentions s'établissent-elles à un niveau intermédiaire entre 1 et 2? Ignorance vaut mieux que «gourance».

L'intérêt de ces deux Q.C.M. est qu'elles ont de bonnes courbes pour chaque réponse, illustrant les strates de compétence. Chaque réponse est celle d'un sous-groupe dont le niveau est délimité avec une certaine précision (**sélectivités** élevées). Les distracteurs sont représentatifs de synthèses embryonnaires réalisées par les usagers avant l'apprentissage ultérieur[2].

---

1. Corrigé. *Réponse* Autre chose. Dans l'alphabet phonétique international (A.P.I.)

Ou En esperanto.

Et En espagnol (depuis la réforme de l'orthographe réalisée il y a une cinquantaine d'années). En ancien français (XII^{e} siècle), on a une graphie variable mais très proche de ce qui était prononcé.

Règle En anglais, lettres et sons ne correspondent pas exactement. En français, c'est déjà mieux, mais on a deux **i** (**i**, **y**), deux **f** (**f**, **ph**), trois **o** (**o**, **au**, **eau**), etc.

2. La relativité réciproque des performances individuelles et des normes linguistiques locales de l'époque est l'explication et la justification de l'approche adoptée ici. Cette relativité avait été observée déjà, notamment par J.-P. Benzécri et son équipe. Après avoir exposé ses méthodes de similarité, qu'il juge plus puissantes que l'analyse factorielle, celui-ci se rend compte que les «caractères» peuvent avoir un poids différent, et que les «individus» peuvent valoriser ou non certains caractères. Il parle alors (dans *l'Analyse des données*, t.1, p.65) d'une «hiérarchie» des caractères. Hiérarchie mais aussi interdépendance, comme l'ont souligné, dans le cas de «caractères» de nature linguistique, Saussure et le structuralisme (*Ibidem*, p.66). Nos analyses déboucheront sur un schéma qui inclut ces deux aspects, celui de la spirale (V.p.164).