Mesure et évaluation en éducation

La capacité discriminante d'un instrument de mesure

Louis Laurencelle

Volume 20, numéro 1, 1997

URI : https://id.erudit.org/iderudit/1091385ar
DOI : https://doi.org/10.7202/1091385ar

Aller au sommaire du numéro

Éditeur(s)

ADMEE-Canada - Université Laval

ISSN

0823-3993 (imprimé)
2368-2000 (numérique)

Découvrir la revue

Citer cet article

Laurencelle, L. (1997). La capacité discriminante d'un instrument de mesure. *Mesure et évaluation en éducation*, *20*(1), 25–39. https://doi.org/10.7202/1091385ar

Résumé de l'article

La « capacité discriminante » est cette propriété d'un test, instrument de mesure ou examen scolaire, grâce à laquelle on peut séparer les personnes ou objets mesurés selon leurs valeurs et les discriminer les uns des autres. Le concept, partiellement inspiré de l'indice de pouvoir classificatoire de Ferguson, est élaboré sur trois bases : la probabilité d'une catégorisation correcte, la largeur suffisante des intervalles de mesure et le nombre d'intervalles efficaces d'une distribution de mesures empirique ou théorique. La capacité discriminante, qui s'exprime comme une fonction du coefficient de fidélité de l'instrument de mesure, s'ajoute donc à l'arsenal conceptuel de la théorie classique des tests.

Mesure et évaluation en éducation
Vol. 20 n° 1, pages 25 à 39.

La capacité discriminante d'un instrument de mesure

Louis Laurencelle
Université du Québec à Trois-Rivières

La « capacité discriminante » est cette propriété d'un test, instrument de mesure ou examen scolaire, grâce à laquelle on peut séparer les personnes ou objets mesurés selon leurs valeurs et les discriminer les uns des autres. Le concept, partiellement inspiré de l'indice de pouvoir classificatoire de Ferguson, est élaboré sur trois bases : la probabilité d'une catégorisation correcte, la largeur suffisante des intervalles de mesure et le nombre d'intervalles efficaces d'une distribution de mesures empirique ou théorique. La capacité discriminante, qui s'exprime comme une fonction du coefficient de fidélité de l'instrument de mesure, s'ajoute donc à l'arsenal conceptuel de la théorie classique des tests.

(capacité discriminante - erreur de mesure - catégorisation correcte - indice de Ferguson)

« Discriminative capacity » is defined as a property of a test, measuring device or scholastic exam, which enables us to segregate and categorize objects or people according to their measured values. The concept, partially derived from Ferguson's index of classificatory power, is developed upon three bases : the probability of categorizing an object (or person) in its proper measuring interval; the sufficient length of measuring intervals; the number of efficacious intervals in an empirical or theoretical distribution of measures. Expressed as a function of the reliability coefficient of a measuring device, discriminative capacit appears as a brand new tool in the conceptual apparatus of classical test theory.

(discriminative capacity - measurement error - rightful categorization - Ferguson's index)

Une règle à mesurer de 1 mètre, graduée au 0,01 m, nous permet de catégoriser tous les objets selon leurs grandeurs en 101 groupes distincts: objets de longueurs inférieures à 0,01 m, objets de 0,01 m à moins de 0,02 m, *et cœtera* jusqu'aux objets de 1 m et plus. De même, un pèse-personne en kilogrammes, gradué au ½ kg et plafonnant à 150 kg, peut séparer les personnes et les objets selon leurs masses en 301 catégories distinctes. En combien de catégories distinctes une épreuve d'aptitude en mathématiques sépare-t-elle les élèves? Combien de degrés d'intensité vraiment distincts

nous permettent d'obtenir une échelle de tendances suicidaires? En combien de catégories un test de quotient intellectuel (QI) va-t-il répartir une population?

La « capacité discriminante » d'un instrument de mesure est le nombre de catégories de mesure qu'exploite cet instrument et parmi lesquelles il peut classer les objets. Pour un objet de mesure physique, tels la règle à mesurer et le pèse-personne, la capacité discriminante, qu'on peut dénoter D', est facilement obtenue par

$$D' = E / u , \quad (1)$$

c'est-à-dire en divisant l'étendue (E) de l'instrument par son unité de mesure (u)[1]. Cette opérationnalisation se heurte toutefois à deux difficultés. La première est que, pour la plupart des instruments, voire pour la règle à mesurer, l'étendue effective (E) reste indéfinie et ambiguë:

- indéfinie parce que, pour la règle par exemple, elle peut être appliquée bout à bout et que l'étendue virtuelle serait alors infinie et parce que, en général et pour les mesures biologiques, psychologiques et cognitives, il n'y a pas de vrai minimum ni de vrai maximum et qu'il est impossible de déterminer l'étendue;

- ambiguë parce que, pour une valeur donnée de E, la répartition des éléments d'un univers (d'objets ou de personnes) varie d'une catégorie à l'autre et est rarement uniforme, la zone centrale étant ordinairement plus densément occupée que les zones extrêmes, et que la capacité discriminante devrait refléter le nombre de catégories effectives (ou effectivement occupées) du système de mesure.

La seconde raison, très prégnante pour les édumétriciens et les psychométriciens, vient de ce que l'unité de mesure (u), dans un examen scolaire ou un test, n'est aucunement définie[2] et que la catégorisation d'une personne dans une mesure X d'unité u (ou X_u) est fortement incertaine, en raison d'une erreur de mesure importante caractérisant la plupart de nos instruments de mesure.

C'est pourquoi nous proposons d'aborder différemment le concept de capacité discriminante en tenant compte à la fois de l'erreur de mesure, en fait de la *fidélité* (ρ_{XX}) de l'instrument de mesure, et en invoquant une répartition normale (ou gaussienne) du caractère mesuré dans la population. Par « capacité discriminante » d'un instrument de mesure, nous entendons alors

le nombre d'intervalles de mesure efficaces, ou catégories efficaces, parmi lesquels l'instrument de mesure à fidélité ρ_{XX} peut répartir une population statistique normale de telle façon qu'un élément mesuré ait une probabilité d'au moins ½ d'être classé dans son intervalle propre.

La capacité discriminante, que nous dénoterons D, a pour formule générale:

$$D = 2{,}654 / \sqrt{(1 - \rho_{XX})} \; . \tag{2}$$

Dans les sections qui suivent, nous développons et justifions les bases mathématiques du concept de capacité discriminante, en payant d'abord tribut aux auteurs qui ont précédemment abordé la question.

Notes historiques

Les publications en psychométrie ne font pas mention d'un concept semblable à celui de la capacité discriminante. Nous retrouvons cependant quelques idées apparentées à notre concept et, notamment, l'indice de pouvoir classificatoire de Ferguson, dont nous avons tiré profit.

Bassière et Gaignebet, dans leur traité de *Métrologie générale* (1966), font seulement mention de la « capacité d'information » d'un instrument, définie comme *le nombre d'états distincts qu'il peut prendre*. « La capacité d'information de l'instrument dépend à la fois de son pouvoir de résolution et de son temps de réponse qui limite le nombre de mesure [sic] par unité de temps » (*op. cit.*, p. 140-141). L'implication du temps, ou d'un taux d'information par unité de temps, évoque en même temps le concept de « capacité de canal » en théorie de l'information (voir, par exemple, Ralston & Reilley, 1983) et crée, pour nous, une confusion notionnelle.

D'un autre côté, les auteurs en psychométrie et en docimologie reconnaissent depuis longtemps l'impact des niveaux de réussite des items d'un test sur le bon étalement des scores du test, donc sur sa capacité discriminante. Cette discussion, qui concerne l'analyse et la sélection des items devant constituer un test afin de produire un bon étalement des scores totaux, se retrouve par exemple dans Davis (1951) et Anastasi (1994); on peut la résumer comme suit. S'ils sont peu inter-corrélés, il faut sélectionner des items à niveaux de réussite (ou « indices de facilité ») moyens, c'est-à-dire des niveaux proches de ½ pour des items dichotomiques. Au contraire, si les items sont fortement corrélés et ont tendance à discriminer les mêmes répondants, on doit les choisir avec des niveaux de réussite gradués. Bien que pertinentes pour notre propos, ces considérations sur la construction d'un test

à base d'items se situent en amont du concept de capacité discriminante, une propriété globale de l'instrument de mesure.

Le vocable de « pouvoir discriminant » revient dans les textes de psychométrie classique, mais dans une acception restreinte, et il désigne alors l'efficacité d'un item pour discriminer les meilleurs des moins bons répondants, selon le caractère mesuré (Henrysson, 1971). Cette propriété d'item est quantifiée au moyen de divers indices, tels les soi-disant « indices d'homogénéité », le coefficient de corrélation bisériale entre item et score total, etc. (voir aussi Guilford, 1954). Elle se retrouve aussi dans la théorie des réponses aux items, précisement dans le paramètre de pente (« a ») de la courbe caractéristique d'item (Hambleton, Swaminathan & Rogers, 1991).

Ferguson publie, en 1949, un article intitulé « On the theory of test discrimination », dans lequel il propose de considérer et de quantifier, d'une certaine manière, l'étalement des scores d'un test ou des données d'un instrument de mesure; il propose aussi un « coefficient de discrimination » (voir aussi Guilford, 1954). Ferguson argue que, si n répondants se répartissent en k scores différents à raison de $f_1, f_2, \cdots, f_k$ répondants par valeurs de score, les discriminations obtenues sont égales au nombre de distinctions deux à deux que le test permet d'opérer parmi les n individus. Ce nombre, calculé par

$$N_d = f_1 \times f_2 + f_1 \times f_3 + \cdots + f_{k-1} \times f_k , \tag{3}$$

reflète donc l'efficacité de l'instrument de mesure pour discriminer les uns des autres individus ou objets. Le nombre de distinctions (N_d) s'obtient aussi par la formule équivalente:

$$N_d = \binom{n}{2} - \sum_{j=1}^{k} \binom{f_j}{2} . \tag{4}$$

Ce nombre N_d dépend fondamentalement de trois facteurs: le nombre total de mesures (n), le nombre de valeurs (ou catégories de mesure) disponibles (k), et la distribution des mesures d'une valeur à l'autre ($f_1, f_2, \cdots$). Pour n constant, le pouvoir discriminatif pourra donc croître avec k, mais il dépend de façon essentielle de la distribution des fréquences f_j. En effet, la seule adjonction de nouvelles catégories de mesure, au-delà de k, n'entraînera de distinctions supplémentaires que si ces catégories sont occupées. En outre, le nombre maximal de distinctions qu'un système à k catégories permet d'obtenir advient lorsque ces catégories sont occupées également, c'est-à-dire si la distribution des fréquences f_j est uniforme. Ce nombre maximum est alors, approximativement:

(5) $$\max N_d = \binom{n}{2} - k\binom{n/k}{2}$$

$$= \frac{n(n-1)}{2} - \frac{n(n-k)}{2k}$$

Le « coefficient de discrimination » δ, proposé par Ferguson, est alors le quotient du nombre réel de distinctions opérées sur le nombre maximal, soit

(6) $$\delta = N_d / \max N_d .$$

Thurlow (1950), dans un long article publié l'année suivante, réclame la copaternité du coefficient de discrimination tel que présenté par Ferguson. Fait intéressant, Thurlow rajoute la notion de « discriminations stables », c'est-à-dire des discriminations qui persistent du test au retest par contraste aux discriminations non stables, et il relie cette différence à la fidélité de l'instrument de mesure. Cependant, l'auteur n'opérationnalise pas davantage son idée.

L'élaboration du concept de « capacité discriminante »

Notre concept de capacité discriminante est essentiellement basé sur celui de Ferguson. Cependant, nous avons généralisé le concept emprunté à Ferguson, qui s'appuyait sur une distribution de fréquences, pour l'appliquer à toute distribution de probabilités d'intervalles, puis nous l'avons inversé afin de déterminer le nombre équivalent k^* d'intervalles (ou catégories de mesure) *efficaces*. La capacité discriminante, en outre, est tout aussi essentiellement basée sur la fidélité de l'instrument de mesure, dénotée par le coefficient habituel ρ_{XX}, ceci en faisant intervenir explicitement une probabilité (dénotée γ) de catégorisation de chaque mesure dans son intervalle propre. Les paragraphes suivants présentent les étapes de développement du concept.

Catégorisation correcte d'une personne (ou d'un objet)

Plaçons-nous d'emblée dans le contexte de la mesure d'un caractère dans un objet, ou dans une personne, au moyen d'un instrument de mesure. L'instrument, appliqué à l'objet, fournit une mesure X; l'unité de mesure n'est pas spécifiée. Le coefficient de fidélité ρ_{XX} est connu.

Pour un objet donné (ou une personne donnée) i, la valeur précise, ou valeur vraie V_i, de cet objet existe[3] et chaque mesure en constitue une approximation. En fait, la mesure X_i va s'écarter plus ou moins de V_i selon que la précision, la fidélité du procédé de mesure est petite ou grande. En fait, plus grand est l'intervalle de mesure formé autour de V_i, plus grande est

la probabilité que la mesure prise X_i s'y retrouve. Soit L la largeur de l'intervalle considéré et γ la probabilité que la mesure X_i de l'objet considéré s'y retrouve, nous avons

$$\gamma(L) = \Pr\{ X_i \in (V_i - \tfrac{1}{2}L, V_i + \tfrac{1}{2}L) \} . \quad (7)$$

La probabilité γ qu'un objet soit catégorisé *dans son intervalle propre*, c'est-à-dire dans le voisinage immédiat de sa valeur vraie V_i, varie directement avec la largeur d'intervalle considérée. À la limite, posant L = 0, on voit qu'il est (mathématiquement) impossible que X_i égale V_i : ainsi, il serait illusoire de croire, pour un instrument de mesure à échelle infiniment divisible, que la mesure obtenue X_i soit, jusqu'à la dernière décimale, la valeur exacte de l'objet mesuré. Pour catégoriser numériquement un objet avec une certaine plausibilité, il faut que la probabilité de catégorisation correcte soit déterminée et suffisante, ce qui implique en retour la détermination d'un intervalle de grandeur suffisante.

Détermination d'un intervalle de mesure suffisant

L'écart de la mesure X_i par son rapport à V_i est généralement nommé *erreur de mesure* et dénotée ε (Lord & Novick, 1968; Laurencelle, sous presse). La valeur attendue (ou espérance) de ε est 0, sa variance (ou espérance de ε^2) σ_ε^2 et sa distribution, symétrique. C'est expressément pour cet écart ε, entre mesure et valeur vraie, que le modèle de la loi normale fut élaboré au XVIII^e siècle (Stigler, 1986). Nous pouvons donc, en toute légitimité, faire se reporter la variable ε au modèle normal et en indiquer la distribution par

$$\varepsilon \sim N(0, \sigma_\varepsilon^2).$$

Dans le contexte indiqué, la mesure $X_{i,o}$ de l'objet i à l'occasion o s'écrira donc

$$X_{i,o} = V_i + \varepsilon_o . \quad (8)$$

De plus, dénotant par Z une variable normale de distribution N(0,1), nous pouvons ré-exprimer (7) plus explicitement, comme

$$\begin{aligned} \gamma(L) &= \Pr\{ V_i + \varepsilon_o \in (V_i - \tfrac{1}{2}L, V_i + \tfrac{1}{2}L) \} \\ &= \Pr\{ \varepsilon_o \in (-\tfrac{1}{2}L, +\tfrac{1}{2}L) \} \\ &= \Pr\{ Z \in (-\tfrac{1}{2}L/\sigma_\varepsilon, +\tfrac{1}{2}L/\sigma_\varepsilon) \} \\ &= 1 - 2 \times \Pr\{ Z > \tfrac{1}{2}L/\sigma_\varepsilon \} . \end{aligned} \quad (9)$$

Fixant la probabilité γ à une valeur déterminée, nous pouvons inverser (9) afin de trouver la grandeur d'intervalle L suffisante pour qu'une mesure X_i y

soit catégorisée avec une probabilité (d'au moins) γ. L'équation résultante est simplement

$$L(\gamma) = 2\sigma_e z_{½(1+\gamma)} \; ; \tag{10}$$

dans cette expression, $z_{½(1+\gamma)}$ dénote le centile $100\times½(1+\gamma)$ de la loi normale standard.

Nous pouvons enfin ramener la grandeur de l'intervalle suffisant L en valeur standard, c'est-à-dire sur une échelle standardisée de moyenne 0 et variance 1, en divisant les deux parties de l'équation (10) par l'écart-type de l'échelle de mesure, σ_X. La fidélité ρ_{XX} étant également définie par

$$\rho_{XX} = \sigma_V^2 / \sigma_X^2 \quad = 1 - \sigma_e^2 / \sigma_X^2, \tag{11}$$

nous pouvons écrire de façon équivalente $\sigma_e = \sigma_X\sqrt{(1-\rho)}$. Ainsi, la grandeur d'intervalle suffisante standardisée, $\lambda(\gamma)$, devient

$$\lambda(\gamma) = 2\sqrt{(1-\rho)}z_{½(1+\gamma)} \; . \tag{12}$$

Dans une échelle de mesure X catégorisée en intervalles de grandeur commune $\lambda(\gamma)\sigma_X$, la mesure X_i d'un objet serait placée dans son intervalle propre avec une probabilité γ. La question qui reste à résoudre est la suivante: combien une telle échelle comporte-t-elle d'intervalles suffisants, de catégories?

Le nombre d'intervalles d'une distribution de probabilité et les intervalles efficaces

Adoptons d'emblée un procédé de mesure idéal, relatif à une variable X de fidélité parfaite ($\rho_{XX} = 1$) et ayant dans une population hypothétique une distribution quelconque, de densité de probabilité $f(X)$ et de répartition F(X). La probabilité qu'une mesure X soit égale à une valeur précise x est nulle, par définition. Cependant, la probabilité p que X tombe dans un intervalle, par exemple $p_{a,b} = \Pr[\, X \in (a, b)]$, $b > a$, est facilement calculable par

$$p_{a,b} = \Pr[\, X \in (a, b)] = F(b) - F(a) \; ; \tag{13}$$

ce calcul peut être effectué pour tous les intervalles de X.

Maintenant, l'indice de discrimination de Ferguson, en particulier le nombre de distinctions N_d de l'équation (4), est basé sur les *fréquences* d'occupation des intervalles plutôt que sur leurs probabilités. Imaginons alors un échantillon de taille n, les n observations étant réparties à raison de la

fréquence f_j dans l'intervalle j. En valeurs attendues $\hat{f}_j$ ou pour la moyenne de tous les échantillons de taille n tirés de la population, nous avons

$$\hat{f}_j = E(f_j) = np_j . \tag{14}$$

Admettant un nombre quelconque (k) d'intervalles de mesure, l'expression de Ferguson (4) devient, en espérance et pour un échantillon hypothétique de taille n:

$$E(N_d) = \binom{n}{2} - \sum_j^k \binom{np_j}{2} = \frac{n^2}{2}\left(1 - \sum_j^k p_j^2\right). \tag{15}$$

La taille d'échantillon n étant arbitraire, de même que la constante « 2 » dans la formule, nous simplifions la formule en divisant (15) par $n^2/2$, obtenant enfin

$$C_k = 1 - \sum^k p_j^2 , \tag{16}$$

cette quantité étant *proportionnelle* (plutôt qu'égale) au nombre de distinctions opérées en vertu des fréquences des k intervalles considérés. À l'instar de Ferguson (1949) et Thurlow (1950), il est aisé de montrer que le nombre maximum de distinctions, ou la valeur maximale de (16), survient lorsque toutes les probabilités sont égales[4], c'est-à-dire lorsque $p_j = 1/k$ pour tout j, ce maximum étant

$$\max C_k = 1 - 1/k . \tag{17}$$

Le nombre maximum de distinctions survient donc quand toutes les fréquences, ou de façon équivalente, toutes les probabilités, sont égales: nous désignerons alors les intervalles de mesure ayant des probabilités égales *intervalles efficaces*, et nous désignerons provisoirement par le symbole k^* le nombre d'intervalles efficaces d'un système de mesure. La règle suivante permet de déterminer k^* dans un cas donné. Soit C_k défini en (16), le nombre de distinctions opérées par un système de mesure en k intervalles de fréquences $\{ f_j \}$. Alors, en égalisant $C_k = \max C_{k^*}$ et en inversant (17) pour obtenir k^*, nous avons

$$k^* = (1 - C_k)^{-1} ; \tag{18}$$

l'indice k^* indique le nombre d'intervalles efficaces, c'est-à-dire des intervalles virtuels à probabilités égales tels qu'ils déterminent le nombre de distinctions C_k constaté dans le système de mesure X. En d'autres mots, si notre système de mesure X était découpé en k^* intervalles à probabilités (ou fréquences d'occupation) égales, il serait caractérisé par un indice de discrimination C_{k^*} égal à l'indice C_k observé. La quantité k^* constitue ainsi une mesure standardisée du nombre d'intervalles d'un système de mesure, quelle que soit la forme de distribution de ses données.

Il est enfin important de noter que la détermination de k^* ne dépend pas du nombre k d'intervalles originaux du système de mesure; ce nombre k pourrait même rester indéterminé. Nous pouvons alors réécrire les définitions (16) et (18) en les généralisant pour qu'elles s'appliquent à des systèmes de mesure non bornés, ayant un nombre d'intervalles indéterminé: la distribution normale en est un exemple illustre. Les définitions généralisées sont simplement

(16') $$C = 1 - \sum_{-\infty}^{\infty} p_j^2 ,$$

et

(18') $$k^* = (1 - C)^{-1} .$$

Le nombre d'intervalles efficaces à catégorisation correcte

La discussion de la section précédente, portant sur le nombre k^* d'intervalles efficaces, concernait un système de mesure idéal, à fidélité parfaite: dans un tel système, les intervalles de mesure, ou l'unité de mesure, peuvent être définis *ad libitum*, d'une grandeur quelconque, et les objets mesurés y seront toujours catégorisés correctement. Toutefois, les mesures sont, en pratique, rarement parfaites, et le coefficient de fidélité ρ_{XX} qui les caractérise est habituellement inférieur à 1.

Ainsi, pour que les mesures découlant d'un procédé de mesure à fidélité ρ_{XX} donnent lieu à une catégorisation « fiable », il faut tenir compte de l'erreur de mesure comprise dans chaque évaluation. Nous avons montré plus haut qu'il est possible de déterminer une grandeur d'intervalle, $L_X(\gamma) = \lambda(\gamma)\sigma_X$, telle que la probabilité qu'un objet soit catégorisé dans son intervalle propre soit d'au moins γ. Nous pouvons alors découper d'une manière ou d'une autre l'axe des X en une séquence d'intervalles limitrophes de grandeur $L_X(\gamma)$. Reprenant la distribution hypothétique des X déjà mentionnée, nous pouvons trouver par (13) la probabilité d'occurrence de chaque intervalle j, puis calculer (16') et enfin (18'); ce dernier calcul fournit l'indice k^* - le nombre de catégories efficaces caractérisant notre procédé de mesure X -,

lequel est explicitement une fonction de γ (la probabilité de catégorisation correcte) et de ρ_{XX} (la fidélité du procédé de mesure).

L'invocation du modèle normal et la capacité discriminante

Les idées et formules élaborées plus haut permettent de déterminer k^* pour chaque procédé de mesure particulier, selon sa fidélité ρ_{XX}, en fonction de la distribution, empirique ou théorique, des mesures X_i dans la population, et pour une probabilité γ donnée de catégorisation correcte. Cependant, dans le but de parvenir à un concept mieux cadré et plus net, nous proposons deux simplifications importantes, l'une sur la distribution des X, l'autre sur la probabilité γ.

En premier lieu, il apparaît commode d'adopter un *modèle de référence* pour la distribution des mesures X dans la population, modèle qui tient lieu de la distribution réelle, obvie à son estimation et, en quelque sorte, l'idéalise. Nous faisons allusion, bien sûr, au modèle de la distribution normale, qui sert ordinairement à représenter la répartition des caractères comme l'intelligence (ou QI), les grandeurs en anthropométrie et bien d'autres caractéristiques dans la population[5].

En second lieu, le choix d'une valeur particulière de la probabilité de catégorisation correcte, γ, reste arbitraire. Des valeurs telles que $\gamma \geq ½$ paraissant souhaitables, nous retenons $\gamma = ½$ comme probabilité de référence. Toute autre valeur dans l'intervalle ($½ \leq \gamma < 1$) serait acceptable.

Adoptant donc le modèle de la distribution normale et la probabilité $\gamma = ½$ pour la catégorisation correcte, l'indice k^* devient une fonction de la fidélité ρ_{XX} seulement, et nous définissons ainsi la **capacité discriminante**, soit

$$D_\rho = k^*[f(X) \text{ normale}, \gamma = ½, \rho_{XX}] . \tag{19}$$

Pour une valeur donnée de $\rho = \rho_{XX}$, donc de $\sigma_e = \sqrt{(1-\rho)}$ en valeurs standardisées, la grandeur d'intervalle standardisée $\lambda(½)$ s'obtient par l'équation (12). Nous appliquons alors cet intervalle en segmentant l'axe standardisé X' en intervalles contigus centrés, produisant ainsi le système d'intervalles suivant:

$$-\infty \cdots (-\tfrac{5}{2}\lambda,-\tfrac{3}{2}\lambda), (-\tfrac{3}{2}\lambda,-\tfrac{1}{2}\lambda), (-\tfrac{1}{2}\lambda, \tfrac{1}{2}\lambda), (\tfrac{1}{2}\lambda, \tfrac{3}{2}\lambda), (\tfrac{3}{2}\lambda, \tfrac{5}{2}\lambda), \cdots \infty . \tag{20}$$

Les probabilités p_j associées à chaque intervalle j dans la distribution normale peuvent être calculées et leurs carrés sommés en vue de l'indice (16'). Partant de l'intervalle central, qui a la probabilité p la plus grande, et allant vers les intervalles plus extrêmes, à probabilités fortement décroissantes, la

sommation se poursuit jusqu'aux deux infinis en ce sens que la valeur de $\sum p_j^2$ converge et se stabilise rapidement près des ailes de la distribution.

L'inversion par l'équation (18') fournit enfin $D_\rho = k^*$ pour la valeur de ρ considérée. Le graphique de la figure 1 montre la relation entre k^* et ρ telle qu'établie à partir du modèle normal et avec la probabilité de catégorisation correcte $\gamma = ½$. La relation entre k^* et ρ devient presque parfaitement linéaire quand nous utilisons, en abscisses, la fonction $1/\sqrt{(1-\rho)}$, comme en fait foi la figure 2. Remaniant l'équation (11), nous voyons que cette fonction $1/\sqrt{(1-\rho)}$ est égale au quotient de l'écart-type des mesures X sur leur erreur-type, σ_X/σ_e.

L'étude de l'ensemble de données (k^*, ρ), pour $0 < \rho < 1$, tel que représenté à la figure 2, permet d'établir une équation de régression linéaire approximative, soit

$$(20) \qquad D_\rho = k^*[\gamma = ½] = 2{,}654/\sqrt{(1-\rho_{XX})} \; ;$$

le coefficient de détermination (r^2) obtenu pour cette régression est d'environ 0,99996. La relation est globalement linéaire, sauf aux extrémités. Ainsi, à l'extrémité basse, quand $\rho \to 0$, le nombre minimal d'intervalles à probabilité $\gamma = ½$ de catégorisation correcte est évidemment 2, violant ainsi la linéarité stricte.

Nous avons étudié la relation (k^*, ρ) pour d'autres valeurs de probabilité γ. À toutes fins utiles, les régressions linéaires obtenues sont:

$$(21a) \qquad k^*[\gamma = 0{,}75] = 1{,}576/\sqrt{(1-\rho_{XX})} \qquad (r^2 = 0{,}99994) \; ;$$

$$(21b) \qquad k^*[\gamma = 0{,}90] = 1{,}113/\sqrt{(1-\rho_{XX})} \qquad (r^2 = 0{,}9998) \; ;$$

$$(21c) \qquad k^*[\gamma = 0{,}95] = \max\{\, 1,\, 0{,}939/\sqrt{(1-\rho_{XX})} \,\} \qquad (r^2 = 0{,}9998) \; .$$

Enfin, d'après les équations (11), la forme générale obtenue pour notre indice du pouvoir discriminant d'un instrument est

$$(22) \qquad D_\rho \approx C_\gamma \, / \, \sqrt{(1-\rho_{XX})}$$

$$= C_\gamma \times \frac{\sigma_X}{\sigma_e} \; ,$$

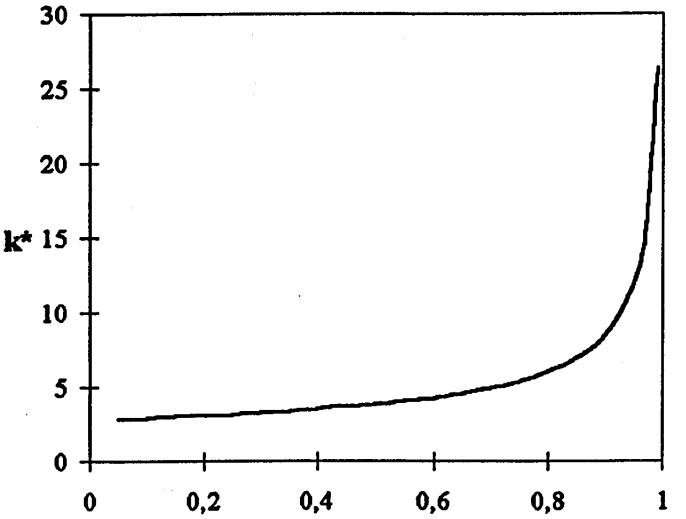

Figure 1 - Relation entre k^* et ρ selon un modèle $X \sim N(\mu_x \sigma^2_x)$

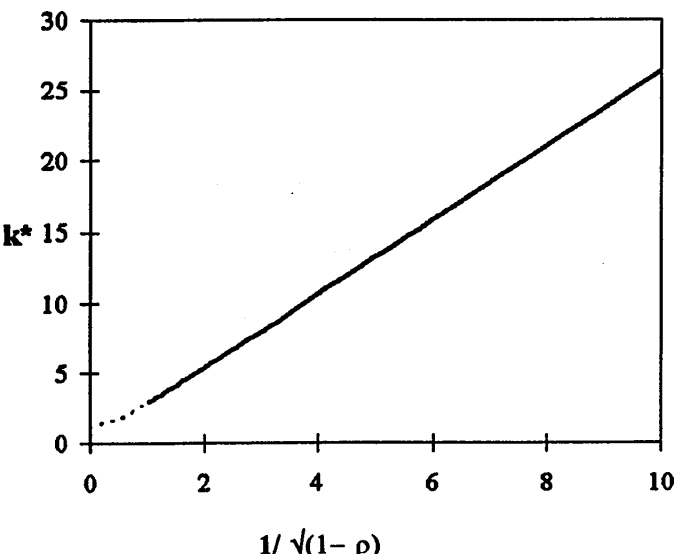

Figure 2 - Relation entre k^* et $1/\sqrt{(1-\rho_{xx})}$ $(= \sigma_x / \sigma_e)$ selon un modèle $X \sim N(\mu_x, \sigma^2_x)$

C_γ étant la constante de pente, fonction de γ et caractérisant l'équation de régression. Cette forme rappelle avec évidence l'indice naïf du départ, $D' = E/u$, avec au numérateur une quantité relative à la dispersion des mesures sur l'axe X et, au dénominateur, une quantité reflétant la précision, structurelle ou statistique, de l'instrument employé.

La forme (22) ci-haut suggère aussi une analogie plus profonde[6], cette fois avec la *fonction d'information* d'un test, $I(\theta)$ (Hambleton *et al.*, 1991, p. 94-95). On peut exprimer cette dernière par:

$$\sqrt{I(\theta)} = \frac{1}{\sigma_e(\theta)} ; \tag{23}$$

la valeur moyenne de cette fonction (dans le domaine de θ) est certainement corrélée avec la capacité (22), ne serait-ce qu'en raison de la parenté mathématique des concepts.

Un exemple d'interprétation

Prenons, comme exemple fictif, un test d'intelligence donnant une mesure de QI; le test, de moyenne 100 et écart-type 15, aurait une fidélité $\rho_{XX} = 0,85$. Réécrivant (10) autrement, la grandeur d'intervalle suffisante pour catégoriser une personne dans son intervalle propre avec probabilité γ est

$$L(\gamma) = 2\sigma_X\sqrt{(1-\rho_{XX})}z_{½(1+\gamma)} . \tag{24}$$

Appliquant nos valeurs de σ_X, ρ_{XX} et $\gamma = ½$, nous avons

$$L(½) = 2\times15\times\sqrt{(1-0,85)}\times z_{0,75} \tag{25}$$

$$\approx 7,84 ,$$

en utilisant $z_{0,75} \approx 0,6745$. L'unité, conventionnelle et artificielle, du QI est de 1 point; en redéfinissant cette unité (à travers les tableaux de conversion normative) pour qu'elle englobe une suite de 7,84 points originaux, la nouvelle échelle de QI serait telle que le score attribué à une personne serait son score vrai avec probabilité $\gamma = ½$. De plus, le nombre de catégories efficaces et à probabilité de ½ dont est capable ce procédé de mesure serait de $D = 2,654/\sqrt{(1-0,85)} \approx 6,85$; il permet donc de classer la population aussi efficacement que s'il la disposait dans environ 7 catégories de tailles comparables.

En conclusion

Le pouvoir de séparer et de catégoriser les objets selon leurs valeurs est l'une des propriétés fondamentales de la mesure. La « capacité discriminante » que nous proposons ici est une opérationnalisation directe de cette propriété, qui tient compte aussi de l'erreur de mesure telle qu'elle est conçue dans la théorie des tests.

Le développement du concept de capacité discriminante nous a conduit à définir une autre quantité intéressante, caractéristique d'un système de mesure: la largeur d'intervalle suffisante $L(\gamma)$. Dans les mesures en sciences humaines, pour lesquelles l'unité de mesure n'a de réalité que symbolique, la quantité $L(\gamma)$, qui est fonction de la probabilité γ de catégorisation correcte, constitue peut-être un palliatif.

La capacité discriminante, telle que nous la présentons, dénote le pouvoir qu'a un instrument de mesure de répartir les objets mesurés en catégories vraiment distinctes et de capacités (ou tailles) comparables. Le concept, qui exprime une propriété d'un système de mesure X à variation continue, pourrait aussi être généralisé à un système de catégories fermées tel qu'on retrouve dans les systèmes d'observation dits qualitatifs. Si une telle généralisation du concept de capacité discriminante était faite, nous serions enfin en possession d'un concept métrologique commun aux mesures exactes des sciences physiques, aux mesures relatives des sciences humaines et aux mesures dites qualitatives.

NOTES

1. La grandeur précise serait plutôt donnée par $D = 1 + \lfloor E / u \rfloor$, où $\lfloor x \rfloor$ dénote la partie entière de x. Ainsi, pour la règle à mesurer de 1 mètre graduée aux unités u de 0,01 m, $D = 1 + \lfloor 1/u \rfloor = 101$.

2. Elle est *stipulée* implicitement à la valeur 1 (*i.e.* le nombre de points, le nombre de réponses correctes), comme dans les échelles de QI, les examens scolaires, etc.

3. Une définition constructive de V_i est $[\Sigma X_{i,o}] / n_o \to V_i$ si $n_o \to \infty$, l'objet i étant mesuré dans une infinité d'occasions o équivalentes.

4. Une démonstration simple de ce résultat est la suivante. Soit $\text{var}(p_j)$, la variance entre les p_j, et $k \times \text{var}(p_j) = \Sigma p_j^2 - [\Sigma p_j]^2$. Puisque $\Sigma p_j = 1$, on a $k \times \text{var}(p_j) + 1 = \Sigma p_j^2$. Or, par définition de variance, $\text{var}(p_j) \geq 0$. La valeur minimale de Σp_j^2 correspond à $\text{var}(p_j) = 0$; toutes les valeurs p_j étant égales et leur somme égale à 1, on obtient $p_j = 1/k$ pour tout j.

5. Nous sommes bien conscient que, à l'opposé, plusieurs caractères mesurables ont une distribution dans la population qui s'écarte sensiblement du modèle normal, tels le revenu des particuliers, le temps d'exécution d'une tâche standardisée, le diamètre des axones.

6. L'auteur tient à remercier l'évaluateur anonyme pour cette intéressante suggestion.

RÉFÉRENCES

Anastasi, A. (1994). Introduction à la psychométrie (traduction par F. Gagné de Psychological testing, 6e éd., Macmillan). Montréal : Guérin Universitaire.

Bassière, M. & Gaignebet, E. (1966). Métrologie générale : théorie de la mesure, les instruments et leur emploi. Paris : Dunod.

Davis, F.B. (1951). Item selection techniques. In E. F. Lindquist (éd.), Educational measurement (pp. 266-328). Washington (D.C.) : American Council on Education.

Ferguson, G.A. (1949). On the theory of test discrimination. Psychometrika, 14, 61-68.

Guilford, J.P. (1954). Psychometric methods (2e édition). New York : McGraw-Hill.

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). Fundamentals of item response theory. London : Sage.

Henrysson, S. (1971). Gathering, analyzing, and using data on test items. In R. L. Thorndike (éd.), Educational measurement (2e édition) (pp. 130-159). Washington (D.C.) : American Council on Education.

Laurencelle, L. (sous presse). Théorie et techniques de la mesure instrumentale. Sainte-Foy : Presses de l'Université du Québec.

Lord, F.M. & Novick, M.R. (1968). Statistical theories of mental test scores. Reading (Mass.) : Addison-Wesley.

Ralston, A. & Reilley, E.D. Jr., éd. (1983). Encyclopedia of computer science and engineering (2e édition). New York : Van Nostrand.

Stigler, S.M. (1986). The history of statistics. The measurement of uncertainty before 1900. Cambridge (Mass.) : Harvard University Press.

Thurlow, W.R. (1950). Direct measures of discriminations among individuals performed by psychological tests. Journal of Psychology, 29, 281-314.