Mesure et évaluation en éducation

À la recherche de l'« unité de mesure » en psychométrie : réflexions sur la mesure en sciences humaines

Louis Laurencelle et Jim O. Ramsay

Volume 24, numéro 2-3, 2001

URI : https://id.erudit.org/iderudit/1091169ar
DOI : https://doi.org/10.7202/1091169ar

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. & Ramsay, J. O. (2001). À la recherche de l'« unité de mesure » en psychométrie : réflexions sur la mesure en sciences humaines. *Mesure et évaluation en éducation*, *24*(2-3), 41–52. https://doi.org/10.7202/1091169ar

Résumé de l'article

La mesure en sciences humaines, bien que largement répandue et pratiquée, a mauvaise presse chez les théoriciens et fait souvent figure d'enfant pauvre à côté de la mesure instrumentale qu'on retrouve dans les sciences physiques et biologiques. Les auteurs reprennent le débat, en reformulant la question et en revoyant les concepts de mesure, unité de mesure, étalon, grandeur et échelle. Ils posent quelques balises qui permettent de juger, sans passion ni préconception, la légitimité des mesures, que ce soit pour des tests psychométriques, des examens scolaires ou d'autres types d'indices socioculturels. La perspective générale, bien que critique, reste favorable et optimiste.

À la recherche de l'« unité de mesure » en psychométrie : réflexions sur la mesure en sciences humaines

Louis Laurencelle
Université du Québec à Trois-Rivières

Jim O. Ramsay
Université McGill

MOTS-CLÉS : Mesure, unité de mesure, échelles de mesure, métrique élastique, théorie des réponses aux items, unité de lecture

La mesure en sciences humaines, bien que largement répandue et pratiquée, a mauvaise presse chez les théoriciens et fait souvent figure d'enfant pauvre à côté de la mesure instrumentale qu'on retrouve dans les sciences physiques et biologiques. Les auteurs reprennent le débat, en reformulant la question et en revoyant les concepts de mesure, unité de mesure, étalon, grandeur et échelle. Ils posent quelques balises qui permettent de juger, sans passion ni préconception, la légitimité des mesures, que ce soit pour des tests psychométriques, des examens scolaires ou d'autres types d'indices socioculturels. La perspective générale, bien que critique, reste favorable et optimiste.

KEY WORDS : Measurement, unit of measurement, measurement scales, topological scale, item response theory, scale unit

Measurement in the social sciences is widespread in our society, but it has a poor reputation compared with its siblings, instrumental measures in the physical and life sciences. The authors take up again the debate, in re-editing the question and scrutinizing the notions of measurement and measure, measurement unit, standard, measure and scale. They put up some landmarks by which one may decide upon the legitimacy of a measurement context, whether for psychometric tests, academic performance, or other types of socio-cultural measures. Though cautious, the outlook is rather positive and optimistic.

Note de l'auteur : Toute correspondance peut être adressée comme suit : Département des sciences de l'activité physique, Université du Québec à Trois-Rivières, Case postale 500, Trois-Rivières (Québec) G9A 5H7. Téléphone : 819-376-5011, poste 3794. Télécopieur : 819-376-5092. Adresse électronique : Louis_Laurencelle@UQTR.CA

Depuis le milieu du xx^e siècle, la mesure a envahi toutes les dimensions et tous les recoins de notre espace socioculturel. Les tests psychologiques ont proliféré, imités bientôt par les tests et les examens scolaires, les tests d'aptitude et les épreuves de sélection pour des emplois en industrie, ainsi de suite. Tout se mesure ou, à tout le moins, tout semble pouvoir être mesuré. Toutefois, plusieurs penseurs, d'aucuns étant eux-mêmes des concepteurs et des utilisateurs de tests, croient que ces mesures sont illusoires, factices, et que l'interprétation et les opérations faites sur celles-ci sont infondées et abusives. En quelques mots, la question qui se pose est la suivante : lorsqu'on administre à quelqu'un un test, un examen, une épreuve et qu'on en obtient le score X, ce résultat numérique a-t-il un sens, et y a-t-il quelque chose, une " grandeur » quelconque, dont on puisse dire qu'elle a été *mesurée* par cette opération ?

Dans les paragraphes qui suivent, nous abordons la question de la mesure en sciences humaines sous différents angles, en nous référant au concept fondamental de l' « unité de mesure ». *Le Robert* (Rey, 2001) rapporte la définition suivante de A. Lalande :

> *Unité de mesure* : grandeur finie servant de base à la mesure des autres grandeurs de même espèce.

Le concept de « grandeur », quelque peu ambigu, fait référence à la fois au degré d'extension d'un caractère — phénomène ou objet mesuré dans une dimension donnée — et au caractère lui-même. *Le Robert (op. cit.)*, encore, définit :

> *Grandeur* : ce qui est susceptible de variation et peut être calculé, évalué ou mesuré.

Bien sûr, la mesure est, la plupart du temps, rapportée à un « étalon », soit, d'après *Le Robert (op. cit.)* :

> *Étalon* : modèle légal de définition d'une unité de mesure; représentation matérielle d'une unité de mesure.

L'étalon ancre la mesure et l'unité de mesure dans les dimensions du monde physique, telles qu'elles sont exprimées, par exemple, dans les composantes du système métrique. Cependant, le rapport de la mesure à un étalon du monde physique n'est peut-être pas essentiel, et il existe, dans l'espace socioculturel et le monde humain, des « grandeurs », des phénomènes à intensité variable, qui ont en nous et sur nous une présence sensible, peuvent être « mesurés » et ne se réfèrent au mieux qu'accidentellement aux unités du monde physique. Mais n'anticipons pas.

Que reproche-t-on aux mesures et aux instruments de mesure en psychométrie, qui les désavantagerait par rapport aux mesures généralement reconnues? Quelles propriétés nous faudrait-il observer dans une «mesure psychométrique» pour qu'elle apparaisse véritable, à l'instar d'une mesure de température ou de pression? De quelle manière pouvons-nous comparer, par exemple, la mesure de longueur d'une table et le QI d'une personne, avec leurs «unités de mesure» respectives?

Les questions fondamentales

Ayant passé un test de QI, de moyenne 100 et d'écart-type 15, Robert obtient le score 115. À l'examen de mathématiques, qui comportait 20 items semblables de calcul, il en a réussi 16, ce qui lui donne la note de 80%. Il a aussi reçu la cote $T = 65$ à un test de motivation scolaire: rappelons que les cotes T ont une moyenne normative de 50 et un écart-type de 10. Sont-ce là des « mesures» véritables, ou bien des simulacres de mesure, qu'on ne devrait interpréter et utiliser qu'avec réserve?

La mesure

Si tous les objets, tous les phénomènes, toutes les personnes étaient pareils et de grandeurs égales, rien ne servirait de les mesurer: mesurer, sert à comparer, ordonner, juger les personnes ou les objets selon le degré auquel ils possèdent une caractéristique, la grandeur d'une dimension de leur être. *Le Robert (op. cit.)*, qui sous-entend vraisemblablement les mesures et les grandeurs du monde physique, propose les définitions suivantes:

> *Mesurer*: déterminer la valeur de (une grandeur mesurable), lui attribuer un nombre qui fixe son intensité ou son état (par rapport à une grandeur de la même espèce).

et:

> *Mesure*: action de déterminer la valeur de certaines grandeurs par comparaison avec une grandeur constante de même espèce, prise comme terme de référence.

S. S. Stevens, dans son classique *Psychophysics* (1975), y allait plus largement, avec:

> *Mesurer*: assigner des nombres à des objets selon certaines règles.

Cette définition, comme le note Duncan (1984), couvre toutes sortes de mesures de façon générale, en incluant notamment les soi-disant « mesures nominales» ou mesures qualitatives, lesquelles ne se réfèrent aucunement au concept de «grandeur» déjà mentionné. Les étiquettes attribuées dans ces mesures sont totalement arbitraires et elles ne constituent pas un système ordonné, tel qu'on s'attend à le retrouver dans une mesure proprement dite. La définition de Laurencelle (1998), qui n'invoque pas expressément les concepts d'unité de mesure ni de grandeur-étalon, constitue un compromis :

> *Mesure* : une fonction d'information qui exprime symboliquement la grandeur d'une dimension d'un objet.

Le cas type du QI

Donc, par la mesure de QI obtenue par Robert, peut-on croire et affirmer que l'objet « Robert » a une dimension « capacités intellectuelles », dont la grandeur serait indiquée par le score de $X = 115$? Nous croyons que si. Il semble peu douteux que, au contraire des roches ou du vent, les objets « Personnes humaines » possèdent des capacités intellectuelles; l'expérience montre, en outre, que ces capacités ne sont pas de même nature ni surtout de même force d'une personne à l'autre, donc qu'elles constituent une dimension sans doute complexe et à grandeur variable. Le test employé, disons le test *Mon_QI*, peut être ou non un instrument de mesure valide de la grandeur de ces capacités intellectuelles. En admettant que le test soit valide, comme c'est le cas notoire de plusieurs tests, quelles sont les propriétés métrologiques du score produit, $X = 115$, comparativement aux propriétés attendues d'une mesure, en sciences ?

Supposons que, au même test de QI, André obtienne $X = 100$ et Sylvie, $X = 130$. Contrairement à des planches de longueurs 100, 115 et 130 cm, pour lesquelles il est patent que l'écart de longueurs « 115 vs 100 » est de même grandeur que l'écart « 130 vs 115 », on ne peut obtenir de conclusions de cette sorte à partir des QI, et la *soustraction* elle-même, qui fournirait par exemple la différence 115–100 = 15, n'a pas de portée vérifiable ou utile pour les QI. Toutefois, il en est de même pour les températures : une augmentation ou une différence de 5 degrés sur l'échelle Celsius n'ont pas une valeur ou une portée universelle, ne correspondent immédiatement à rien de concret. D'autre part, tout comme les températures de 20, 25 et 30 °C sont de plus en plus chaudes, ont des effets correspondants et sont ressenties comme telles, les personnes ayant des QI de 100, 115 et 130 réussissent de mieux en mieux dans leurs interactions à caractère intellectuel et sont reconnues (généralement) pour être de plus en plus intelligentes. Est-ce à dire que les degrés de température lus au thermomètre et les QI résultant d'un test d'intelligence ont des propriétés similaires, les mêmes lacunes et la même structure conventionnelle ? Duncan (1984), qui scrute l'histoire et les fondements de la notion de température et de sa mesure, n'y trouverait pas de différence essentielle : y en a-t-il une ?

Échelle ordinale ou linéaire?

Ainsi, les propriétés dites linéaires (dans le jargon classique de la psychométrie)[1] sont absentes ou non démontrables de la mesure en sciences humaines. Toutefois, cela n'interdit pas certaines opérations, tels l'addition et le calcul de moyenne ou d'écart-type, qui sont pratiquées et qui, à notre avis, portent une signification valide. En fait, les mesures, qui sont élaborées au moyen de procédés relativement arbitraires, sans référence à une unité de mesure prédéterminée, reflètent, quand elles sont valides, la variation d'une grandeur dans une dimension du monde socioculturel. Dans cette dimension, les objets sont ordonnés en grandeur, selon:

$$O_1 < O_2 < O_3 < \ ... \ < O_n , \quad \text{(éq. 1)}$$

et les mesures prises, qui répercutent l'arrangement des grandeurs avec plus ou moins de précision, selon:

$$X_1 < X_2 < X_3 < \ ... \ < X_{\mathrm{n}} , \quad \text{(éq. 2)}$$

constituent ce que Ramsay (1996) appelle une «topologie régulière» (parce que les groupes de mesures X_i respectent les relations de voisinage caractérisant les grandeurs observées). Dans le langage classique de Stevens, un tel type de mesures était appelé " ordinal», c'est-à-dire reflétant purement le rang de la grandeur observée dans un ensemble. De gré avec Ramsay (1996), nous jugeons cette caractérisation indûment restrictive, d'abord parce que nous acceptons la réalité des «grandeurs» reflétées par la mesure (et que seule l'absence d'une unité de mesure idoine nous empêche de convertir en mesure linéaire), ensuite parce que, au contraire d'une mesure qui se baserait seulement sur la comparaison de grandeurs voisines dans un ensemble donné, il est possible et facile d'augmenter la précision avec laquelle la mesure reflète la grandeur visée (en réduisant l'erreur de mesure, en ajoutant des paliers de grandeur intercalaires) et ainsi, à la limite, de convertir la série de rangs en un axe métrique qui serait élastique (une topologie régulière).

En conclusion

En fin de compte, les mesures, que nous pratiquons à toutes fins utiles en sciences humaines comme en psychométrie, sont-elles valables ou, comme certains le prétendent, tout ça doit-il aller au caniveau et devons-nous réévaluer les phénomènes de la socioculture par une approche purement qualitative? Le débat, pour un aspect ou l'autre de la question, remonte à l'origine même de la mesure sociale, à Quetelet (1796-1874), Galton (1822-1911) et Binet (1857-1911). Notre position, ici, est pragmatique et carrément optimiste. De très nombreuses mesures sont, depuis longtemps, pratiquées en éducation, en psychologie, en *counseling*, et leur usage a été validé et entériné par des générations de professionnels et de praticiens. En un mot, cela fonctionne, est démontré utile et donne lieu à des modèles explicatifs et prédictifs acceptés : voilà notre argument pragmatique. Quant à l'argument théorique, nous posons comme évident et axiomatique que les grandeurs visées par la mesure, par exemple l'aptitude pour l'addition à deux chiffres, existent; que les instruments de mesure en question donnent, si valides, des résultats reflétant l'ordonnancement de ces grandeurs; que ces résultats constituent une topologie régulière, une échelle sublinéaire, une métrique élastique, possédant des propriétés encore à définir, au-delà des propriétés des échelles nominales et ordinales conventionnelles.

La discussion reste ouverte.

X, T ou θ: quel type de score reflète le mieux la capacité, la grandeur mesurée ?

Aux fins de la discussion, posons que, pour une dimension donnée de la réalité socioculturelle, il existe un système ordonné de grandeurs tel que (éq. 1) et on a pu mettre au point un instrument qui fournit le système de mesures X tel que (éq. 2), avec les imperfections usuelles. Cela étant admis, tout se passe comme si, dans le monde physique, nous avions recensé tous les objets ayant la dimension « longueur » (dans l'axe le plus étendu de leur volume spatial), nous avions collecté toutes leurs mesures, disons Y_i, basées sur le mètre-étalon, puis, abandonnant ce dernier, nous avions « cuisiné » l'axe de mesure pour l'assouplir, le rendre élastique et en altérer les segments. Les mesures résultantes, disons X_i, sont en corrélation ordinale parfaite avec les mesures originales et respectent intégralement le couple (éq. 1, éq. 2). Bien sûr, nous avons perdu quelque chose en abandonnant l'unité de mesure : une grandeur fixe, la même partout sur l'échelle de mesure, et permettant par exemple des transpositions impossibles sans elle. Mais nous n'avons pas tout perdu, et quelles qualités pouvons-nous rattraper ?

Désignons le système des mesures X_i respectant les contraintes (éq. 1) et (éq. 2) « système brut » : un tel système constitue une topologie régulière à la Ramsay. Or, il est d'usage, en psychométrie comme en docimologie, de proposer, au-delà de la mesure brute X_i, un système de mesure à valeurs normées. Les manuels pertinents font état de deux grandes classes de systèmes normés : ceux dont seules les propriétés globales de moyenne et écart-type sont imposées (systèmes à moyenne normée), puis ceux dont les propriétés de distribution au complet sont définies (systèmes à distribution normée). Les systèmes à moyenne et écart-type imposés n'altèrent pas les intervalles sur l'axe de mesure; pour cette raison, leur usage pose parfois des difficultés d'interprétation[2].

Quant aux systèmes à distribution normée, qu'il s'agisse des échelles centiles ou des scores normalisés (*i.e.* à distribution normale imposée), les intervalles entre les valeurs mesurées sont altérés par la transformation appliquée. En d'autres mots, le processus de standardisation, qui exploite la « population » des mesures telle que représentée dans le « groupe normatif », opère une transposition d'un axe (disons X) à un autre axe (disons Y) en respectant les inégalités brutes, selon :

$$Y_i > Y_j \text{ si } X_i > X_j$$

mais sans respecter l'unité de mesure (d'ailleurs non spécifiée) ni les intervalles de l'axe original. Cela a pour effet, par exemple, que l'égalité $X_4 - X_3 = X_2 - X_1$, sur l'échelle d'origine, n'entraîne pas l'égalité $Y_4 - Y_3 = Y_2 - Y_1$ sur l'échelle convertie. Cette voie pour standardiser l'échelle de mesure place donc, entre les mains du concepteur, une série ordonnée de segments qu'il peut, en principe, étirer, raccourcir, accommoder, dans le but d'obtenir un axe à intervalles égaux qui posséderait davantage de propriétés linéaires. Comment *prouver*, nous objectera-t-on, que de telles propriétés sont présentes, que les intervalles obtenus sont réellement égaux ? Une telle preuve, si on la trouve, restera toujours conventionnelle, voire litigieuse, puisqu'elle aura pour contexte et pour champ de validation les réalités et les jugements de la socioculture, plutôt que les dimensions du monde physique et les grandeurs-étalons qui les accompagnent. Cela dit, le comportement des données psychométriques dans plusieurs modèles prédictifs et leur insertion réussie dans des systèmes d'équations linéaires sont des arguments en faveur de notre thèse, à savoir que ces données, dérivées de la mesure en sciences humaines et en psychométrie, manifestent en pratique des propriétés proches des systèmes linéaires.

Prenons d'abord le système X_i (éq. 2) et un système correspondant de données converties Y_i. La corrélation de rangs entre les deux ensembles de mesures est évidemment parfaite, ce qui indique que la corrélation linéaire entre eux est aussi, généralement, très bonne[3]. Ensuite, supposons qu'on a en mains une mesure critère C, par rapport à laquelle il est loisible d'apprécier la validité concomitante ou prédictive de notre mesure X. Alors, dans la plupart des cas, l'équation linéaire prédictive « $\hat{C}_i = f_{\text{Linéaire}}(Y_i)$ » fonctionne assez bien. On peut aussi, légitimement, améliorer le modèle prédictif ci-dessus, notamment son cœfficient de détermination R^2, en procédant à des ajustements fins du système de mesure Y, c'est-à-dire en accroissant et réduisant les intervalles entre les valeurs observées successives en Y, tout en respectant bien sûr l'arrangement ordinal initial entre celles-ci.

Il reste une troisième option, à partir du système de mesure X, pour élaborer un nouveau système de mesure qui possède des propriétés métriques intéressantes : c'est la *modélisation*. L'idée consiste à imaginer puis à construire un modèle formel de l'objet mesuré, ici la personne humaine, de doter le modèle d'un ou de quelques paramètres abstraits rigoureusement définis, enfin de trouver un moyen de raccorder le modèle et ses paramètres aux performances observées chez la personne. La « théorie des réponses aux items », structurée par Lord en 1953, met en œuvre cette approche ; le lecteur trouvera un bref aperçu de la TRI dans Laurencelle (1998).

Dans le contexte d'une mesure basée sur la somme d'items, tels un test scolaire ou une épreuve psychologique normée, l'approche TRI postule un paramètre de capacité chez le répondant, soit θ, et une fonction-type qui indique a posteriori la probabilité qu'un répondant doté d'une capacité θ opte pour la (bonne) réponse désignée. L'approche consiste donc à substituer à X, qui ordinairement est le nombre brut d'items réussis, la valeur estimative du paramètre θ pour ce répondant. Le paramètre ainsi que la fonction de probabilité qui l'utilise définissent donc un système de mesure. L'axe de définition de θ est l'axe réel, de $-\infty$ à ∞; il est plausible que les auteurs qui en ont disserté les premiers avaient en tête la loi normale, elle-même définie sur l'axe réel, comme modèle de répartition dans la population. Or, comme le montre Ramsay (1996), la métrique de θ n'est imposée que par la forme de la fonction de probabilité qui l'utilise : toute transformation monotone de θ peut être compensée par une contre-transformation de la fonction de probabilité. Ainsi, en modifiant la fonction afin qu'elle fournisse la même probabilité requise, Ramsay passe de θ, qui varie de $-\infty$ à ∞, à $\Psi = 1/[1 + \exp(-\theta)]$, qui varie maintenant de 0 à 1, soit comme une capacité normée. Quels que soient les mérites du modèle TRI (ou de ses nombreux avatars), celui de définir une mesure (θ) ou une métrique indiscutable n'en est pas un. En outre, ni la mesure θ ni aucun de ses avatars ne possèdent de propriété métrologique nouvelle par rapport au score brut X, et il est quasi certain que le score X possède en puissance les mêmes propriétés que celles découlant de la modélisation TRI.

Ramsay (1996), en prolongement de l'argument évoqué ci-dessus, développe deux autres idées. Il pose d'abord l'hypercube de probabilité $[P_j(\theta)]^n$, établi à partir du paramètre θ (ou d'un paramètre équivalent) et dont les axes sont donnés par les fonctions de probabilité d'item $P_j(\theta)$ pour n items. Dans un tel espace, chaque sujet, à capacité θ_0, définit un hyperpoint, $[P_1(\theta_0), P_2(\theta_0), ..., P_n(\theta_0)]$, et l'ensemble des N sujets constitue un essaim. La projection de chaque hyperpoint sur la diagonale principale de l'hypercube définit un axe de mesure, de domaine $(0, \sqrt{n})$, relié de façon monotone au paramètre θ; cette mesure, de plus, constitue un estimateur de la valeur vraie de X, $\varepsilon(X_0|\theta_0)$, associée à la valeur $\theta = \theta_0$. Ramsay, en outre, propose la longueur d'arc, $s(\theta_0)$, une mesure de la distance nécessaire sur la «trajectoire» depuis son origine jusqu'au point correspondant à θ_0. Une telle mesure a pour avantage de faire référence explicitement à la structure géométrique des données réelles[4], à l'essaim des hyperpoints. Cependant, la diagonale principale, définie dans un système d'axes de probabilité, est tout aussi réelle et elle entretient un lien plus simple avec les données brutes, mais surtout la trajectoire menant, d'un hyperpoint à l'autre, jusqu'au point $[P_j(\theta)]^n$, est non seulement non rectiligne, mais elle est généralement non monotone[5], ce qui pose un problème de validité pour cette mesure.

Enfin, les corrélations *linéaires* ordinairement très fortes entre ces divers paramètres, soit X, θ, $\varepsilon(X_0|\theta_0)$ et $s(\theta)$, laissent à penser que les avantages prétendus des paramètres dérivés sont plus idéaux que réels. Toutes choses étant presque égales (ou linéairement équivalentes) par ailleurs, pourquoi ne préférerions-nous pas X, la somme d'items toute simple, qui est immédiatement basée sur les réponses (physiquement) observables du sujet?

L'unité de mesure appliquée et la précision de la mesure en psychométrie

Il existe une «unité de mesure» concrète qui, comme en physique, permet de construire des grandeurs, soit en additionnant les unités, soit en les subdivisant. Ce type de mesure et d'unité de mesure n'existe pas en sciences humaines. Il n'est pas universel non plus en physique ni en général, comme le montrent les exemples de la mesure au thermomètre ou de l'indice de consommation d'une voiture automobile. Néanmoins, dès qu'on utilise un système numérique pour noter une mesure, on fait affaire avec l'unité de l'instrument utilisé. La règle graduée au centimètre a le centimètre pour unité, le thermomètre est noté habituellement au degré Celsius près, etc. Ce type d'unité, que nous désignons u, existe partout, sert à la lecture sur instrument et impose certaines contraintes de précision à la mesure.

Tout instrument de mesure a son unité de référence, c'est-à-dire la plus petite subdivision significative qu'il transmet sur la grandeur de l'objet mesuré. Ainsi, que ce soit pour le rendement d'un élève à un test de mathématique ou le QI d'une personne, le score produit varie de manière discontinue, le plus souvent d'un point à la fois. Soit Y, la valeur réelle de la grandeur d'un objet et X_0, une mesure possible. Alors, la mesure attribuée à la grandeur Y sera $X_0 + \Delta_u$, où Δ est un entier positif ou négatif, défini par $\Delta = \lfloor (Y - X_0) / u \rfloor$.

Laurencelle (1998) énonce quelques théorèmes relatifs à l'effet de la discontinuité présente sur la précision de la mesure, l'un notamment sur la variance d'erreur et un autre sur le cœfficient de fidélité. Le concept de *capacité discriminante* (Laurencelle, 1997), dérivé du concept de pouvoir classificatoire de Ferguson (1949), fonde une interprétation probabiliste originale de l'unité de mesure u telle que présentée ici[6]. Eisenhart (1947) aborde l'effet de cette granularité de l'axe sur la valeur et la distribution des indices statistiques.

La valeur de la mesure en sciences humaines déborde bien sûr le seul aspect des propriétés de son axe numérique. La mesure est *appliquée* dans des contextes qui donnent lieu à des décisions, un classement des personnes, une sanction d'admission dans un programme ou de diagnostic médical ou psychosocial. Comme les «grandeurs» mesurées sont de l'ordre de la socioculture et n'ont pas de référence physique indépendante, leur seule référence pratique, sur laquelle appuyer ces décisions, est la répartition des mesures dans la population. Or, outre la linéarité discutable des mesures, leur propriété de précision joue encore davantage pour influencer la qualité des décisions, et notre sentiment est que les auteurs et praticiens en psychométrie ont fait bon marché de cette propriété. Laurencelle (1998, 2000, soumis) aborde la question de la précision des normes et des seuils normatifs *du point de vue de l'incertitude du classement ordinal* dans la population. Finalement, à notre connaissance, personne jusqu'à présent ne s'est intéressé aux effets combinés de l'erreur de mesure et, disons, de l'erreur ordinale sur la précision des seuils.

La légitimité des opérations statistiques

Les mesures X circulant en sciences humaines n'ont généralement pas d'unité de mesure physiquement ou indépendamment définie, comme c'est le cas pour les mesures du monde physique, avec le mètre, le gramme, la seconde. Cette absence d'étalon, qui soit fixe et surtout égal d'une zone à l'autre de l'axe de mesure, fait que l'axe métrique utilisé est élastique et n'autorise pas entièrement les opérations linéaires. Cependant, ce qu'on pourrait nommer les «irrégularités» élastiques de la mesure constituent généralement un phénomène restreint, local, de peu d'envergure, et qui ne suffit pas à annuler la valeur expressive des opérateurs statistiques usuels. En fait, les dites « irrégularités » de progression dans l'axe métrique sont facilement masquées par l'erreur de mesure[7]. De plus, dans les cas tout de même fréquents où l'ensemble des données de population se répartit convenablement, c'est-à-dire d'une manière quasi normale ou non fortement opposée au modèle normal, les indices statistiques usuels (moyenne, écart-type, corrélation) n'en subissent nul contrecoup. Par exemple, la médiane de population est un indice de position adéquat même pour des mesures d'un axe élastique, *et* la moyenne arithmétique en est un estimateur efficace, peu biaisé si la répartition des données n'accuse pas d'asymétrie trop forte. Autre exemple : un groupe à écart-type plus élevé a des caractéristiques plus hétérogènes qu'un autre à écart-type plus faible, que l'axe de mesure soit élastique ou non.

La pratique de traitements statistiques de tous ordres, depuis le simple calcul d'une moyenne de classe en mathématique jusqu'à l'utilisation de modèles factoriels alambiqués par L. L. Thurstone (1959) et J. P. Guilford (1971), montre, avec tous les vices qu'on peut lui reconnaître et les reproches qu'on lui a adressés, que la mesure en sciences humaines véhicule une réalité, qu'elle a des propriétés parentes de celles reconnues en sciences physiques, et qu'elle peut servir d'ingrédient dans l'édification de modèles prédictifs ou explicatifs de notre socioculture.

NOTES

1. Les échelles (de mesure) linéaires incluent les échelles dites à intervalles égaux et celles à rapports égaux.
2. Le modèle sous-entendu de distribution normale sert généralement de guide aux interprétations. Or, comme on peut le constater en observant soit la « population », soit un échantillon normatif de taille suffisante, souvent la distribution empirique des mesures brutes X_i et, conséquemment, celle de leurs équivalents normés (par une transformation linéaire) sont à peu près normales et confortent la validité des interprétations faites.
3. Dans le cas de deux variables aléatoires à distributions normales, la corrélation entre elles (ρ) est très proche de leur corrélation de rangs (ρ_{Spearman}). En fait, asymptotiquement sur n, « $\rho = 2 \sin (\pi\rho_{\text{Spearman}}/6)$ » (Yule & Kendall, 1946), l'écart le plus grand entre les deux cœfficients ρ et ρ_{Spearman} étant d'environ 0,02 (lorsque $\rho_{\text{Spearman}} = 0{,}58$).
4. Il s'agit en fait des données calculées, sur la base d'un modèle TRI, en vertu duquel le vecteur de réponses du sujet, soit $(r_1, r_2, ..., r_n)$, où $r_j = 0|1$, est transformé en un vecteur des probabilités associées à θ_0, soit $(P_1, P_2, ..., P_n|\theta_0)$.
5. Dans le sens où sa monotonicité n'est pas assurée, sauf dans des cas de cohérence dimensionnelle extrême, lesquels sont inhabituels en sciences humaines.

6. L'auteur (voir aussi Laurencelle, 1998) propose de fixer comme unité de mesure *u* la grandeur *D* telle que toute personne de valeur vraie V_i ait une probabilité d'au moins γ d'obtenir un score X_i dans l'intervalle $\{V_i - {}^1/_2 D;\ V_i + {}^1/_2 D\}$. Raisonnement et calcul sont fondés sur le modèle de la théorie des tests et la loi normale de l'erreur.
7. Nous pensons ici à l'erreur de mesure classique (résultant d'abord des imprécisions de l'instrument de mesure ou de son opération) plutôt qu'à l'« erreur ordinale » évoquée à la section précédente.

RÉFÉRENCES

Duncan, O.D. (1984). *Notes on social measurement*. New York : Russel Sage Foundation.

Eisenhart, C. (1947). Effects of rounding or grouping data. In C. Eisenhart, M.W. Hastay & W.A. Wallis (dir.), *Techniques of statistical analysis* (pp. 187-223). New York : McGraw-Hill.

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

Guilford, J.P. (1971). *The analysis of intelligence*. New York :McGraw-Hill.

Laurencelle, L. (1997). La capacité discriminante d'un instrument de mesure. *Mesure et évaluation en éducation*, *20*, 25-39.

Laurencelle, L. (1998). *Théorie et techniques de la mesure instrumentale*. Sainte-Foy : Presses de l'Université du Québec.

Laurencelle, L. (2000). L'incertitude des seuils statistiques et les limites de tolérance, avec des applications en psychométrie. *Lettres Statistiques*, *11*, 1-29.

Laurencelle, L. (accepté). L'incertitude des seuils statistiques et l'établissement d'une norme de qualification sûre. *Mesure et évaluation en éducation*.

Lord, F.M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, *13*, 517-548.

Ramsay, J.O. (1996). A geometrical approach to item response theory. *Behaviormetrika*, *23*, 3-16.

Rey, A. (2001). *Le Grand Robert de la langue française* (2e édition). Paris : Dictionnaires Le Robert.

Stevens, S.S. (1975). *Psychophysics*. New York : Wiley.

Thurstone, L.L. (1959). *Multiple-factor analysis*. Chicago : Chicago University Press.

Yule, G.U., & Kendall, M.G. (1946). *An introduction to the theory of statistics*. London : Griffin.