2014 ■ vol. 10 ■ no. 1

The discriminating capacity of a measuring instrument: Revisiting Bloom (1942)'s theory and formula

Louis Laurencelle [a]

[a] Université du Québec à Trois-Rivières, Canada

Abstract ■ "Discriminating 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, anticipated by Bloom and derived here 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, discriminating capacity appears as a new tool in the conceptual apparatus of classical test theory.

Keywords ■ Discriminating capacity, Classificatory power, Classical test theory, Reliability.

louis.laurencelle@uqtr.ca

Introduction

A 1-meter measuring stick graduated in cm allows one to categorize all possible one-dimensional objects according to their lengths in 101 different lots: those having less than 0.01 m, those ranging from 0.01 to less than 0.02, etc., up to those having 1.00 or more. Similarly, a bathroom scale in kilograms, graduated in ½ kg and ranging up to 150 kg, can separate people or objects according to their masses in 301 different categories, In how many categories can we allocate pupils from their scores in a math exam? How many truly different intensity levels can we obtain from a psychological scale of suicide propensity? How many distinct categories of cognitive ability can produce some particular IQ test?

The "discriminating capacity" of a measuring instrument or test is the number of categories used by the test or instrument among which it can classify objects. For a purely physical instrument, such as the measuring stick or bathroom scale, discriminating capacity, tentatively noted D′, is easily seen to be :

$$D' = R/u, \tag{1}$$

i.e. the quotient of the range (R) of magnitudes covered over the unit of measurement (u)[1]. However, this simple definition faces two drawbacks. As a first problem, in most measuring instruments and objects to be measured, the effective range (R) is both indefinite and ambiguous:

- indefinite, because the measuring stick, for instance, can be used end to end so that the virtual range is infinite and, more to the point, in psychological, cognitive or biological phenomena, there are no real, i.e. definite, minimum or maximum and determination of the range is practically unfeasible;
- ambiguous, because, for a given value of R, the distribution of objects or persons in the target population varies across the range, or from one category to another, the central zone usually being more densely occupied than the left and right tails, whereas discriminating capacity should indicate the number of *effective* categories, or the number of categories *effectively occupied*, throughout the measuring instrument.

A second, twofold difficulty pertains to "u", the unit of measurement. It is generally undefined in educational and psychological measurement[2], and the allocation of a person to a particular unit score, e.g. IQ = 112, is largely unreliable, on account of a significant "measurement error" typical in most of our procedures.

The arguments outlined above lead us to tackle in a

[1] More precisely, one should write: D′ = $\lfloor$R / $u\rfloor$ + 1, where $\lfloor x \rfloor$ denotes the integer part of *x*. Thus, for a 1-meter measuring stick in 0,01 graduation marks, D′ =$\lfloor$1 / 0.01$\rfloor$ + 1 = 101.

[2] The measurement unit is only stipulated to 1 (e.g. number of correct responses, total count of item scores), as in the raw scores in IQ evaluation, in scholastic exams, in personality scales, etc.

more realistic manner the concept of discriminating capacity, by taking into account the instrument's measurement error, or its reliability (noted ρ_{XX}) and by stipulating the normal, Gaussian, law as a distribution template for population's scores. With these conditions in mind, we define the discriminating capacity of a measuring instrument as *the number of efficacious value intervals or categories, among which a measuring instrument having reliability ρ_{XX} can allocate a normal distributed population of people so that a measured person have a probability at least ½ of being put in his/her proper interval or category*.

The discriminating capacity, that we shall note D, is given by :

$$D \approx \frac{2.67}{\sqrt{1-\rho_{\mathrm{XX}}}}. \tag{2}$$

In the following, we develop and justify the mathematical bases of the concept and formula, first paying tribute to the authors who earlier have addressed the same issue.

Historical notes

No published article but one do refer to a concept similar to our concept of discriminating capacity: we will quote B. S. Bloom's work later. However, related ideas can be found in the earlier literature, particularly Ferguson (1949)'s "classificatory power", on which our concept is partly based.

In their treatise "Métrologie générale" (1966), Bassière and Gaignebet refer only verbally to an instrument's *information capacity*, defined as the number of different states it can take and transmit. "Information capacity of an instrument depends both on its resolution power and its response time which limits the number of measurements per time unit" (pp. 140-141). Mention of time, or of an information rate per time unit, connotes the concept of "channel capacity" in information theory (e.g. see Ralston & Reilley, 1983), albeit in our context it stirs up a notional mix-up.

In other respects, specialists of psychological and educational measurements have since long ago acknowledged the effect of the success rates of items on the spread or variance of the test's scores, hence its discriminating capacity. This precise issue, which regards the analysis and selection of items to be included in a simple, one-dimensional scale with the aim of producing a fair distribution of total scores, appears for instance in Davis (1951) and Anastasi (1997); we may summarize it as follows. If the inter-correlations among items are null or weak, one must select those items having a median success level, e.g. a level of ½ for dichotomous items. On the opposite, if items are highly (and positively) correlated, thus discriminating the same examinees, one should choose an array of items with well spread, stepped-up difficulty levels. Although they are highly relevant to our purpose, these considerations on item selection and test construction take place logically before, or under, the concept of discriminating capacity, a macroscopic property of the measuring instrument.

The term "discriminating power" comes up in some classical textbooks on psychometrics, in a quite specialized meaning, referring to an item's effectiveness in discriminating the "best" from the "worst" respondents, according to the measured attribute (Henrysson, 1971). This property of an item is translated in a number of different indices, such as the so-called "homogeneity index", the biserial correlation coefficient (between the item and the total score), etc. (see also Guilford, 1954). It also appears in item response theory (Hambleton, Swaminathan & Rogers, 1991) in the guise of parameter "*a*", the multiplicative component of the item response function.

Ferguson's 1949 paper, "On the theory of test discrimination", suggests to consider and quantify the minute spread of scores produced by a test or of data from any measuring instrument; Ferguson proposes also a "discrimination coefficient" (see also Guilford, 1954). He argues that, if n examinees' scores are scattered across k value categories, with f_1, f_2, …, f_k scores per category, the resultant discriminations equal the number of paired non-null differences among examinees allowed by the test. This number, computed as:

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

thus reflects the instrument's effectiveness to discriminate the measured individuals or objects one from the other. This number can be obtained more simply by:

$$N_d = \left(\begin{array}{c} n \\ 2 \end{array}\right) - \sum\nolimits_{j=1}^{k} \left(\begin{array}{c} f_j \\ 2 \end{array}\right). \tag{4}$$

Quantity N_d depends basically on three factors : the total number (n) of measurements, the number (k) of different values (or measurement categories) available, and the distribution of individual scores among value categories (f_1, f_2, ...). Holding n constant, the discriminating power may grow with k, but it revolves essentially around the distribution of frequencies f_j. Indeed, the sole adjunction of one or more value categories, above k, will not induce additional discriminations except if these new categories are effectively occupied. Moreover, the maximum number of discriminations allowed with a k-category system occurs when the categories are evenly occupied, i.e. when $f_1 = f_2 = \ldots = f_k$. This maximum number is, approximately:

$$\begin{aligned} maxN_d &\approx \binom{n}{2} - k\binom{n/k}{2} \\ &= \frac{n(n-1)}{2} - \frac{n(n-k)}{2} \end{aligned} \tag{5}$$

Ferguson's (1949) "discrimination coefficient", or index of classificatory power, is then the quotient of the effective number of discriminations allowed by the test (N_d) on the virtual maximum number (max N_d), i.e.:

$$\delta = N_d / maxN_d. \tag{6}$$

Thurlow, in a long article published later (in 1950), claims co-authorship of Ferguson's concept. Interestingly, Thurlow puts in the notion of "stable discriminations", that is, differences that keep on re-test, by contrast to unstable or reverting discriminations. He links up this notion to the test or instrument's reliability. However, Thurlow does not pursue this idea further.

Bloom's proposal. Finally, in a paper published in 1942, Bloom propounded his concept of discriminating capacity, without lending it a name, and gave the formula:

$$D_{Bloom} = \frac{\text{Range of scores}}{3\sigma_X\sqrt{1-\rho_{XX}}}, \tag{7}$$

where σ_X and ρ_{XX} are the test's standard deviation and reliability coefficient respectively. In his words, "This ratio indicates the number of categories which may be obtained from this range of test scores so that the chances of a point in one category overlapping with the corresponding point in the next category is about one in one thousand." (p. 521). Bloom also gives, for illustration, the case of a normal distribution of scores, where "the range of scores on a test is six times the standard deviation" (p. 521), simplifying the above formula to $2/\sqrt{1-\rho_{XX}}$. This article of Bloom (1942) anticipated Laurencelle (1997)'s own proposal[3].

Bloom's unnamed concept is indeed, by its description and planned utilization, a "capacity", referring to the number of value categories, or set cardinality of values, conveyed by the test. However, Bloom's 1942 paper evades three significant issues, on top of not deriving explicitly his proposed formula. He does not link his "number of categories" with the idea of discriminating among examinees, as Ferguson (1949) does. He does not explain how he obtains his probability statement ("one in one thousand") nor how this probability links with the other parameters of his formula. Finally, he supposes known the "range of scores" from a test: is it a virtual range, i.e. from the absolute minimum to the absolute maximum possible score, an empirical range derived from actual measurements, etc.? And, in his given example, he puts the range of a normal distribution as 6 times its standard deviation, an unjustified assertion.

Development of the concept of discriminating capacity

Our concept of discriminating capacity is essentially based on Ferguson's index of classificatory power. However, we generalized Ferguson's concept, which referred to a frequency distribution, to apply it to any probability distribution of value intervals and then *invert* it in order to determine the corresponding number k^* of efficacious intervals or value categories. Furthermore, discriminating capacity is also based on the measuring instrument's reliability, embodied by its reliability coefficient ρ_{XX}, through the specification of a probability (noted γ) of categorizing each measured object in its proper category or interval. The stages of our development of the concept appear in the following paragraphs.

Correct categorization of a person, or object. Let's suppose a measurement context wherein some attribute is measured in a person or an object, through a measuring device or test. The test, applied to the person, renders a measure, X_i; the unit of measurement is not specified, and the test's reliability coefficient ρ_{XX} is known.

For a given object or person i, the precise magnitude, or true value T_i, exists[4], and each measurement X_i is a valid estimate of it. In fact, X_i may deviate more or less from T_i depending on whether the

[3] At the time of publication (1997), the author was not aware of Bloom's 1942 paper nor of any other reference to it, whether direct or indirect.

[4] A constructive definition of T_i is $[\Sigma X_{i,o}] \,/\, n_o \rightarrow T_i$ when $n_o \rightarrow \infty$, object i being repeatedly measured on an indefinite number of occasions o (Lord & Novick, 1968 ; Laurencelle, 1998).

test's precision, or reliability, is low or high. The greater the value interval encompassing T_i , the higher the probability that X_i lies within its limits. Let L be the length of this interval and γ, the probability of the said interval containing X_i, we have:

$$\gamma(L) = Pr X_i \in (T_i–1/2L, T_i + 1/2L). \qquad (8)$$

Probability γ that an object be classified within its proper interval, i.e. in the immediate vicinity of its true value T_i, is a direct function of L, the interval length. In the limit, $L = 0$ would crush to zero the probability that X_i be in the vicinity of T_i : indeed, it would be fanciful to think that, with an instrument having an infinitely divisible scale, the measured value X_i be equal to the true T_i unto its last decimal digit. In order to numerically categorize a measured object with some degree of plausibility, one needs that the probability of a correct categorization be established and sufficient, which entails in turn the determination of a sufficient interval length.

Determining a sufficient interval length. The difference between the observed X_i and hypothesized true T_i value is usually dubbed "measurement error" and noted ε (Lord & Novick, 1968 ; Laurencelle, 1998). Postulated to be a random variable, the expected value of ε if 0, its variance σ_ε^2, and its distribution symmetrical. It is expressly for this difference ε between estimate and true value that the normal random model of distribution was reinvented by Gauss in 1809 (Stigler, 1986). Thus, we may legitimately relate the ε variable to the normal (or Gaussian) model and tag its distribution as $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. In this context, the measurement of object i at occasion o is expressed in the model:

$$X_{i,o} = T_i + \varepsilon_o \qquad (9)$$

Moreover, denoting a standard normal variable by Z, i.e. $Z \sim N(0, 1)$, we can rewrite (8) more explicitly, as:

$$\begin{aligned}\gamma(L) &= Pr\{T_i + \varepsilon_o \in (T_i–1/2L, T_i + 1/2L)\}\\ &= Pr\{\varepsilon_o \in (–1/2L, +1/2L)\}\\ &= Pr\{Z \in (–1/2L/\sigma_\varepsilon, +1/2L/\sigma_\varepsilon)\}\\ &= 1–2 \times Pr\{Z > 1/2L/\sigma_\varepsilon\}\end{aligned} \qquad (10)$$

Fixing probability γ to some predetermined value, we may invert (10) and find the interval length L needed so that an observed X_i measurement be rightly categorized with probability at least γ. This inversion is simply:

$$L(\gamma) = 2\sigma_\varepsilon z_{[1/2(1+\gamma)]}; \qquad (11)$$

in the above expression, $z[½(1+\gamma)]$ is the 100×½(1+γ) percentile of the standard normal distribution.

Finally, we can transform the obtained sufficient interval length L to a standard scale, with mean 0 and variance 1, by dividing both parts of equation (11) by the instrument's standard deviation σ_X. As the reliability coefficient ρ_{XX} is equally defined by:

$$\rho_{XX} = \sigma_T^2/\sigma_X^2 = 1 - \sigma_\varepsilon^2/\sigma_X^2, \qquad (12)$$

quantity σ_ε may be written as $\sigma_X\sqrt{1-\rho_{XX}}$. Thus, the standardized sufficient interval length, λ(γ), becomes:

$$\lambda(\gamma) = 2\sqrt{1-\rho_{XX}}\, z_{1/2(1+\gamma)}. \qquad (13)$$

With a measurement scale X categorized, or cut up, in value intervals of common length $\lambda(\gamma)\times\sigma_X$, the X_i measurement of some object would be assigned to its proper category or interval with probability γ or better.

The next question to answer is how many such sufficient intervals does a measurement scale contain.

The number of intervals in a probability distribution and the efficacious intervals. Let's take up an ideal measurement process, rendering an X variable with perfect reliability (ρ_{XX} = 1) and having in some population a given probability distribution $f(X)$ with distribution function $F(X) = \int_{-\infty}^{X} f(y)dy$. The probability that X be exactly equal to some value x is null, by definition. On the other hand, the probability that X falls in some value interval, for example $p_{a,b}$ = Pr{ $X \in (a, b)$ }, $b > a$, is easily computed as:

$$p\{a, b\} = Pr\{X \in (a, b)\} = F(b)–F(a); \qquad (14)$$

this probability can be calculated for every interval in X.

Now, Ferguson's measure of discrimination or classificatory power, more precisely his N_d quantity (4), is calculated from the actual frequencies of observations in the various intervals of values, rather than from their probabilities. Let's imagine a data sample of size n, the n observations X_i being distributed with respective frequency f_j in interval j. For the jth interval, the expected value $\hat{f}_j$, or the average of f_j across all possible samples of size n in the population, we have :

$$\hat{f}_j = E\{F_j\} = n \times p_j. \qquad (15)$$

With k value intervals, Ferguson's expression (4) becomes, asymptotically and for an hypothetical sample of size n :

$$\begin{aligned} E\{N_d\} &= \binom{n}{2} - \sum\nolimits_{j=1}^{k} \binom{n\cdot p_j}{2} \\ &= \frac{n}{2}\left(1-\sum_{j=1}^{k} p_j^2\right) \end{aligned} \tag{16}$$

As sample size n is arbitrary and constant "2" of no import, we may simplify formula (16) so that:

$$C_k = 1 - \sum\nolimits_{j=1}^{k} p_j^2, \tag{17}$$

this quantity C_k being proportional (instead of equal) to the number of discriminations allowed by the probability distribution for the actual set of value intervals. Following Ferguson (1949) and Thurlow (1950), it is easy to show that the maximum number of discriminations, the maximum possible value of (17), occurs when all probabilities are equal[5], i.e. when p_j = 1/k for every j, this maximum being :

$$maxC_k = 1 - 1/k\,. \tag{18}$$

Thus, the maximum number of discriminations is obtained when all frequencies or, equivalently, all probabilities, are set equal: we shall designate such intervals having equal probabilities "efficacious intervals", and denote the number of efficacious intervals of a measuring system by k^*. The following rule enables us to find k^* in a given situation. Let C_k defined by (17), the number of discriminations allowed by a measuring system with frequency distribution { f_j}. Then, equalling C_k = max C_{k^*} and inverting (18), we get:

$$k^* = (1 - C_k)^{-1}; \tag{19}$$

index k^* indicates the number of *efficacious intervals*, i.e. virtual value intervals with equal probability content such that they produce the actual number of discriminations allowed by the measuring system. In other words, if our system were cut up in k^* value intervals, each with an occupancy value proportional to 1/k^*, it would permit C_k discriminations among objects, C_k being the observed parameter.

It is important to note that the calculation of k^* depends only marginally on the number k of original intervals in the measuring system: this number k may even keep undetermined. We may then rewrite definitions (17) and (19) by generalizing them so that they apply to unbounded measuring systems, having an indeterminate number of intervals: such is the normal probability distribution, which extends to both infinites. The generalized definitions are simply:

$$C = 1 - \sum\nolimits_{-\infty}^{\infty} p_j^2 \tag{17′}$$

and:

$$k^* = (1 - C)^{-1} = {}^{1}\!\Big/\!\sum\nolimits_{-\infty}^{\infty} p_j^2 \tag{19′}$$

The number of efficacious intervals with correct categorization. The preceding discussion, on the number k^* of efficacious intervals, revolved around an ideal measuring system, in which reliability is perfect and there is no "measurement error". In such a system, the value intervals can be subdivided *ad libitum* and be made indefinitely fine, and the categorized objects will still be correctly placed. However, actual measurements are very rarely "pure", and the reliability value (ρ_{XX}) which characterizes them is generally less than 1.

Consequently, in order that the measured values X_i taken from a test or measuring system with reliability ρ_{XX} give rise to trustworthy categorizations, one must take into account the measurement error in each case. We have shown earlier that it is feasible to fix an interval length, $L_X(\gamma)$ = $\lambda(\gamma)\times\sigma_X$, such that the probability of categorizing a measured object in its proper interval is at least γ. We may then segment in one way or another the X axis into a sequence of bordering value intervals of length $L_X(\gamma)$. Taking up the hypothetical distribution already mentioned, we can find with (14) the probability of occurrence in each interval j, then compute (17′) and finally (19′) ; this last calculation gives us index k^*, the number of efficacious intervals typical of this measuring system, co-determined by the prescribed γ parameter (the probability of correct categorizations) and the system's reliability coefficient ρ_{XX}.

Normal probability model and "discriminating capacity". The ideas and formulas outlined above allow the determination of k^* for each specific measuring process, based on its reliability ρ_{XX}, the distribution of values in the "population", whether empirical or

[5] A simple demonstration of this theorem is the following. Let var(p_j), the variance among the p_j's, and k×var(p_j) = Σ p_j^2 – [Σ p_j]^2/k. Because Σ p_j = 1, we have k×var(p_j) + 1/k = Σ p_j^2. Now, by definition of variance, var(p_j) ≥ 0. The minimum value of Σ p_j^2 corresponds to var(p_j) =0 ; with all p_j values equal and their sum adding to 1, we have p_j = 1/k for every j.

Table 1 ■ Probabilities for categorizing in the "correct" and in bordering value intervals, for two probability levels (γ) (standardized normal error model)

	-4	-3	-2	-1	Correct (0)	1	2	3	4
γ = ⅓	.0013	.0143	.0825	.2352	.3333	.2352	.0825	.0143	.0013
γ = ½	.0000	.0004	.0211	.2285	.5000	.2285	.0211	.0004	.0000

theoretical, and for a given probability γ of right categorization. Yet, in order to arrive at a sharper and better framed concept, we shall put forward two major simplifications, one on the distribution of *X*, and the other on the γ parameter.

In the first place, it is handy, and often justifiable, to posit a reference model for the distribution of values *X* in some population, such a model substituting for the real, usually unknown, distribution and somewhat "idealizing" it. We refer of course to the normal probability model, which we shall adopt and which is generally used in the psychometric domain and elsewhere and is taken to represent the distribution of psychological, biological and other quantities.

We must note that, as regards its discriminating capacity, the normal model does not stand out over other possible models. Here is how we may characterize the "discriminating power" of probability model *f* (where $f(x) \geq 0$ for all *x* and $\int f(x)\,\mathrm{d}x = 1$). For a measurement system where $\rho_{XX} = 1$, the value intervals, say of length *u*, may be indefinitely subdivided, so that $k^*(f,\ u)$ increases indefinitely. Let's standardize this number by multiplying it by length *u*, and obtain:

$$K_f = u \times k^*(f,u). \qquad (20)$$

Following (19′), the value of $k^*(f,\ u)$ can be estimated by:

$$k^*(f,u) = \left[\sum_{(u\cdot\sigma)=-\infty}^{+\infty} (\sigma \cdot u \cdot f(x))^2\right]^{-1}, \qquad (21)$$

where σ is the distribution model *f* 's standard deviation, so that:

$$K_f \to \left[u \cdot \sigma \sum_{(u\cdot\sigma)=-\infty}^{+\infty} f^2(x)\right]^{-1} \text{ when } u \to 0\,. \qquad (22)$$

With large values of *u*, i.e. $u / \sigma \sim O(1)$, a more precise evaluation of $k^*(f,u)$ will obtain with:

$$k^*(f,u) = \left\{\sum_{(u\cdot\sigma)=-\infty}^{+\infty} [F(x+u\cdot\sigma) - F(x)]^2\right\}^{-1}. \qquad (23)$$

The (asymptotic) value of K_f in the case of the normal probability model is $2\sqrt{\pi}$ = 3.5449. Student's *t* distribution with parameter ν = 3 ("degrees of freedom") has 2.5133, and 3.1094 with ν = 5; symmetrical *Beta* β(3,3) distribution gets $K = \sqrt{12}$= 3.7041, whereas the lopsided β(1,5) gets 2.5559. The "optimal" uniform distribution, aliased as β(1,1), obtains $K = 3.4641$ ($= \sqrt{12}$), a value slightly less than the normal's (and one that we may explain away by the fact that, in spite of its optimality due to equal density intervals, the distribution of doubly bounded, contrarily to the normal density).

The second decision concerns parameter γ, the probability of correctly placing a given measurement in its proper value category. Two candidate values come to mind, γ = ⅓ and γ = ½. The choice of ⅓ could be justified in that there would be an equal chance of a datum being categorized in its own category, or in some higher-valued or some lower-valued category; on the other hand, with this choice, there would be twice more chance that the datum be thrown in a category other than its own. For γ = ½, the chance of placing the datum correctly comes even, the remaining ½ covering the bordering intervals on either side. Table 1 indicates the layout of probabilities in the vicinity of the proper category (labelled "0"), for both values of γ : recall that the generating function for the error variable responsible for the fluctuation in categorizing is stipulated to be the normal probability density. For obvious reasons, we chose γ = ½.

Hence, we submit our concept of discriminating capacity D_ρ in the explicit parameter setting given by:

$$D_p = k^*[normal f, \gamma = 1/2, \rho_{XX}]. \qquad (24)$$

For some value of ρ_{XX}, or equivalently of $\sigma_\varepsilon = \sqrt{1-\rho}$ in a standardized scale *X′*, the standardized interval length λ(½) is obtained with (13). We use this length in segmenting the standardized *X′* axis, to form the system of contiguous intervals such as:

Table 2 ■ Illustrative values of D_ρ ($\gamma = ½$)

ρ_{XX} :	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
D_ρ :	3.86	4.05	4.28	4.56	4.91	5.35	5.96	6.86	8.37	11.80

$$\begin{aligned}&-\infty\ldots(-5\lambda/2,-3\lambda/2),(-3\lambda/2,-\lambda/2),\\&(-\lambda/2,\lambda/2),(\lambda/2,3\lambda/2),(3\lambda/2,5\lambda/2),\ldots\infty\end{aligned} \qquad (25)$$

Probabilities p_j for each interval can then be obtained and their squared values summed up to produce index C (17′) : summation may start in the middle 0-centered interval, then proceed bilaterally, the sum converging rapidly as the interval borders reach ±3. Formula (19′) then supplies $k^* = D_\rho$, the searched for value of discriminating capacity. Some illustrative values of D_ρ are shown in Table 2.

Modeling the D_ρ function. Function D_ρ rises but lazily for lower values of parameter ρ_{XX}, and it takes up speed at about $\rho_{XX} = 0.50$, the upper half of ρ (i.e. 0.50 – 1.00) being altogether the most interesting for measurement specialists. Transformation $g(\rho_{XX}) = 1/\sqrt{1-\rho_{XX}}$ brings about almost perfect collinearity with D_ρ, with $R^2 \geq 0{,}9999$ for any $\gamma \geq ⅓$. In the instance of our D_ρ with $\gamma = ½$, we obtain (approximately[6]):

$$D_p(\gamma = 1/2) \approx 2.67/\sqrt{1-\rho_{XX}}\,. \qquad (26)$$

An alternate, imitative, function, is given by:

$$D_p \approx 2 \times \sqrt{\frac{1+\rho_{XX}}{1-\rho_{XX}}}; \qquad (27)$$

This function correlates highly ($R^2 > 0{,}999$) with the former and it covers the complete range, reaching down to $\rho_{XX} = 0$ (with the concomitant $D_0 = 2$), an asset not shared by formula (26).

Recalling the equations in (12), we see that another expression for our index D_ρ is:

$$\begin{aligned}D_p &\approx C_\gamma/\sqrt{1-\rho_{XX}}\\&= C_\gamma \times \frac{\sigma_X}{\sigma_\varepsilon}\end{aligned}\quad , \qquad (28)$$

with C_γ a slope coefficient depending upon γ and identifying the regression equation. This form (28) echoes in some way our former naïve D′ = R / *u* formula (1), the numerator of which bears on the scatter of values on the *X* axis and the denominator mirroring the precision, either structural or statistical, of the measuring device. Expression (28) suggests yet another, more profound, analogy, now with a test's *information function* (Baker & Kim, 2004; Hambleton et al., 1991):

$$\sqrt{I(\theta)} = \frac{1}{\sigma_\varepsilon(\theta)}; \qquad (29)$$

the expected value of this function (averaged over the θ domain) is surely correlated with capacity (28), be it only because of the obvious mathematical relatedness of the concepts.

A short example

As a fictitious example, let's take an IQ test of intellectual or cognitive abilities. Tests of that sort are commercially available, their scores distributed usually as $N(100, 15^2)$, and they offer a reliability value around $\rho_{XX} = 0.90$. Rewriting (11) equivalently, the interval length needed to categorize some measured person in its proper interval with probability γ is:

$$L(\gamma) = 2\sigma_\lambda\sqrt{1-\rho_{XX}}\,z_{[1/2(1+\gamma)]}. \qquad (30)$$

With our values of parameters σ_X, ρ_{XX} and γ, we get:

$$\begin{aligned}L(1/2) &= 2 \times 15\sqrt{1-0.90} \times z_{[0.75]}\\&\approx 6.40\end{aligned} \qquad (31)$$

where $z_{[0.75]} \approx 0.6745$. The conventional and arbitrary unit of the IQ scale is 1 "point". Had we re-defined this unit (e.g. via conversion tables for standard scores) in such a way that it includes 6.40 original "points", the revised IQ scale would allow one to contend that one's measured score is the right one with probability ½. Furthermore, the number of efficacious categories managed through this measuring instrument is $D \approx 2.67 / \sqrt{1-0.85} \approx 8.44$ (formula 26) or $2 \times \sqrt{(1+0.90)/(1-0.90)} \approx 8.72$ (formula 27) : thus, it can classify the whole population as effectively as if their scores were distributed among about 9 equally sized categories.

[6] The transformed functions, though they are nearly linear, can be minimized on different criteria, thus entailing some arbitrariness in the choice of a solution. Let alone the actual calculation of D_ρ for some specific value of ρ, we retained (for the set of γ values shown above) a solution that seemed to minimize the differences between actual and predicted values of D_ρ, inside the range ρ = [0,50 ; 0,95]. The reader may prefer some other solution strategy.

Concluding remarks

The ability to segregate and categorize objects or people according to their values is a fundamental property of measurement. The concept of discriminating capacity proposed here, in line with Bloom (1942)'s own proposition, puts this property in an operational form, also taking into account measurement uncertainty and error as it is understood in classical test theory.

Beyond Bloom's (1942) inceptive, and incomplete, theorization, our investigation of the concept led us to another interesting concept, characterizing too a measurement system: the sufficient interval length, $L(\gamma)$. In the social and biological realms, wherein measurements frequently present no accountable or substantive measurement unit, quantity $L(\gamma)$, a sort of yardstick for a γ-defined length on the measurement axis, could well serve as a substitute.

Discriminating capacity, as presented here, indicates the ability of a measuring instrument to distribute objects or people among a set of neatly defined categories having quasi equal sizes or capacities. The concept, which denotes a property of a continuous-valued measurement system, could be extended to refer to a discrete-valued or closed category system, nominal scale or non-numeric descriptive process, such as can be found in social investigations and so-called qualitative observational studies. This generalization, subsuming also that of Ferguson's classificatory power, would give us a first metrological tool bridging the gap between discrete- and continuous-valued observational systems and, perhaps, help in reconciling measurement specialists in the pure vs. social sciences.

References

Anastasi, A. (1994). *Psychological testing* (6th ed.). Upper Saddle River (NJ): Prentice-Hall.

Baker, F. B. & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.

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

Bloom, B. A. (1942). Test reliability for what? *Journal of educational psychology*, *33*, 517-526.

Davis, F. B. (1951). Item selection techniques, in E. F. Lindquist (ed.), *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* (2nd ed.). 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 (ed.), *Educational measurement* (2nd ed.) (pp. 130-159). Washington D. C.: American Council on Education.

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.* Québec : 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. (eds): *Encyclopedia of computer science and engineering* (2nd ed.). 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. Direct measures of discriminations among individuals performed by psychological tests. *Journal of psychology*, *29*, 281-314.

Citation

Laurencelle, L. (2014). The discriminating capacity of a measuring instrument: Revisiting Bloom (1942)'s theory and formula. *The Quantitative Methods for Psychology*, *10* (1), 5-12.

Received: 14/08/12 ~ Accepted: 14/02/13