**Mesure et évaluation en éducation**

# Questions et réponses sur la précision des mesures et la variance d'erreur

Louis Laurencelle et Denis Allaire

Volume 18, numéro 3, 1996

URI : https://id.erudit.org/iderudit/1092253ar
DOI : https://doi.org/10.7202/1092253ar

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. & Allaire, D. (1996). Questions et réponses sur la précision des mesures et la variance d'erreur. *Mesure et évaluation en éducation*, *18*(3), 27–42. https://doi.org/10.7202/1092253ar

Résumé de l'article

La variation des mesures d'une grandeur physique et la variation des mesures de force, d'anxiété ou d'habileté en mathématique ont-elles la même signification psychométrique, les mêmes attributs statistiques ? Y a-t-il plus d'une façon dinterpréter la différence (X-Y) provenant de deux mesures distinctes d'une même personne ? Les estimateurs de l'erreur-type de mesure sont-ils tous équivalents ? Les auteurs jettent un regard inquisiteur sur ces questions d'apparence anodine.

*Mesure et évaluation en éducation*
*Vol. 18 n° 3, pages 27 à 42.*

# Questions et réponses sur la précision des mesures et la variance d'erreur

**Louis Laurencelle**
*Université du Québec à Trois-Rivières*

**Denis Allaire**
*Université de Sherbrooke*

*La variation des mesures d'une grandeur physique et la variation des mesures de force, d'anxiété ou d'habileté en mathématique ont-elles la même signification psychométrique, les mêmes attributs statistiques? Y a-t-il plus d'une façon d'interpréter la différence (X-Y) provenant de deux mesures distinctes d'une même personne? Les estimateurs de l'erreur-type de mesure sont-ils tous équivalents? Les auteurs jettent un regard inquisiteur sur ces questions d'apparence anodine.*

***(variance d'erreur, erreur de mesure, estimation, erreur-type d'une différence, écart-type d'une différence)***

*Do variations in measures of any physical dimension and variations in measures of the muscular strength, anxiety or mathematical ability of a person stem from the same psychometric construct, and do they have the same statistical characteristics? Is there more than one way of assessing the difference (X-Y), between two distinct measures in the same person? Are different estimating functions of the standard error of measurement all equivalent? The authors inquire into these seemingly innocuous questions.*

***(error variance, measurement error, estimation, standard error of a difference, standard deviation of a difference)***

## Préambule

Robert obtient 58,7 % à l'examen de mathématique de fin d'année: la note de passage étant de 60 %, doit-on le classer comme «échec en mathématique»? Le score requis pour l'admission dans un cours de «créativité abstraite» est 130 au quotient intellectuel (Q.I.): refuse-t-on Sylvie, qui a obtenu 128 au test? Pierre et Paul concourent tous les deux pour un poste de «réviseur linguistique» dans un journal hebdomadaire, et ils obtiennent respectivement 69 points et 64 points à une épreuve de compétence en

français: lequel faut-il retenir? Marc-André obtient des scores de 107 et 95 respectivement aux Q.I. verbal et non verbal: que peut-on en conclure?

D'un point de vue naïf, les questions énoncées ci-haut appellent chacune une réponse simple et évidente, et cette réponse simple est le plus souvent la réponse retenue en pratique. Toutefois, une analyse plus sérieuse nous conduit à une attitude circonspecte et à une réponse réservée, comme c'est la coutume en science. En effet, l'examen de mathématique administré à Robert, par exemple, comporte un *échantillon* seulement des questions possibles et des habiletés visées par le programme. Robert aurait obtenu un autre résultat, peut-être un résultat meilleur, à partir d'un échantillon différent. De plus, Robert est lui-même d'un jour à l'autre dans des dispositions changeantes pour répondre à l'examen. Pour des examens qui seraient strictement équivalents, Robert obtiendrait néanmoins des résultats un peu variables d'une fois à l'autre. Ce résultat-ci de 58,7 % nous permet-il équitablement de refuser à Robert sa promotion en mathématique? Les autres situations que nous avons dépeintes donnent naissance à un ensemble de réflexions semblables.

Les instruments de mesure ne sont pas absolument parfaits. Certains instruments fournissent même des mesures biaisées, par exemple des mesures trop fortes par rapport à la quantité réelle à évaluer. Même pour des instruments justes, c'est-à-dire sans biais, les mesures produites d'une fois à l'autre pour le même objet évalué varient, *fluctuent*, de façon apparemment aléatoire. Cette fluctuation est (au moins) de deux ordres:

- il y a d'abord une *fluctuation de mesure*, imputable à l'opération de mesure et aux inconsistances de l'instrument de mesure. L'*unité de mesure* utilisée avec l'instrument ajoute une autre imprécision, puisque la position réelle de l'individu évalué est ramenée au multiple le plus proche de l'unité de mesure;

- de plus, la plupart des instruments qui mesurent une caractéristique d'un être vivant sont sensibles aux *fluctuations propres* du caractère évalué, que ce soient la capacité de répondre à un test de Q.I., l'habileté à l'examen de mathématique, la force musculaire du bras, *et cætera*. Cette fluctuation propre est réelle et elle émane de la complexité du système vivant qu'on tente de mesurer.

Les gens de sciences (*e.g.* physique, génie) tout comme les gens de sciences humaines ont cristallisé dans un concept générique cette

imperfection, cette imprécision d'un instrument de mesure: le concept d'*erreur de mesure* ($\varepsilon_M$). L'erreur de mesure est cette différence qui sépare la *mesure* du caractère d'un objet donné de la valeur réelle, ou *valeur vraie*, de ce caractère, soit

(1) $$\varepsilon_M(\text{Objet}) = X(\text{Objet}) - V(\text{Objet}).$$

Il est ordinairement impossible de déterminer précisément l'erreur ($\varepsilon_M$) attachée à une mesure X donnée. Cependant on peut estimer l'amplitude habituelle de cette erreur, de cette différence. Le moyen classique pour ce faire est d'estimer la *variance d'erreur* ou sa racine carrée, l'*erreur-type de mesure* (ou erreur-type), associées à un instrument de mesure. Les relations entre erreurs de mesure ($_1\varepsilon_M$, $_2\varepsilon_M$, $_3\varepsilon_M$, ···), variance d'erreur ($\sigma_M^2$) et erreur-type ($\sigma_M$) s'expriment par

(2) $$\text{Variance d'erreur} = \sigma_M^2 = f(_1\varepsilon_M, _2\varepsilon_M, _3\varepsilon_M, \cdots),$$

où $f(\cdot)$ est une fonction d'estimation de la variance, et

(3) $$\text{Erreur-type} = \sigma_M = \sqrt{\sigma_M^2}\ .$$

Dans la théorie classique des tests (Gulliksen, 1950; Lord & Novick, 1968), l'imprécision ou, plutôt, la précision de l'instrument de mesure sont conceptualisées aussi par la *fidélité*. La fidélité, symbolisée par $\rho_{xx}$ (ou $r_{xx}$), admet plusieurs définitions mathématiquement équivalentes. Le coefficient de fidélité est une mesure relative, variant de 0 à 1 et indiquant le degré de précision des mesures produites. La relation avec la variance d'erreur est

Fidélité = Perfection - Proportion de variance d'erreur

(4) $$\rho_{xx} = 1 - \frac{\sigma_M^2}{\sigma_X^2}\ ;$$

dans l'expression ci-haut, $\sigma_X^2$ dénote la *variance observée*, c'est-à-dire un indice de la dispersion des mesures dans la population des objets mesurables. Un simple remaniement algébrique permet d'exprimer l'erreur-type de mesure à partir du coefficient de fidélité, soit

(5) $$\sigma_M = \sigma_X\sqrt{(1-\rho_{xx})}\ .$$

Les concepts et les idées de base portant sur la précision des mesures ayant été rappelés, avons-nous des questions dans la salle...? Les quelques questions suivantes et leur traitement nous permettront d'examiner de plus près les concepts mentionnés. Elles nous paraissent mériter l'attention parce que, pour les unes, la réponse n'est pas si simple et, pour les autres, il n'y a peut-être pas de bonne réponse.

## Question 1

*Doit-on classer Robert «échec en mathématique» avec sa note de 58,7%?*

**La réponse simple**

La note de passage étant fixée à 60 %, Robert ne l'a pas atteinte. La règle de passage (*i.e.* «obtenir une note égale ou supérieure à 60 %») étant connue et appliquée à tous, il n'est qu'équitable de classer Robert comme ayant un «échec en mathématique».

**La réponse du livre**

Supposons qu'on a pu estimer la variance d'erreur pour l'examen de mathématique à environ 6 (%), d'où une erreur-type de $\sqrt{6} \approx 2{,}45$ %. Qu'est-ce que cela nous apprend?

- Que, de fait, la mesure effectuée par l'examen n'est pas parfaite, qu'elle n'est pas absolument précise et qu'il y a lieu de l'interpréter avec prudence;

- Que le score obtenu par Robert fluctue en plus ou en moins, selon une amplitude d'environ 2,45 %.

Le 58,7 % obtenu pourrait être situé à la marge supérieure de cette fluctuation, indiquant une position possible de (58,7 - 2,45) ≈ 56,25 %, voire à la marge inférieure, indiquant (58,7 + 2,45) ≈ 61,15 %. Les auteurs invoquent aussi le modèle de la distribution normale, et ils supposent que l'erreur de mesure $\varepsilon_M$ varie (d'une occasion de mesure à l'autre) avec une moyenne de zéro et une variance $\sigma^2_M$. Utilisant ce modèle et la valeur $\sigma_M \approx 2{,}45$, on pourrait affirmer par exemple que

(6) $\Pr( X_{Robert} \in \{ {}_{i}X_{Robert}-1{,}960\sigma_M ; {}_{i}X_{Robert}+1{,}960\sigma_M \} ) \approx 0{,}95$ ,

soit, avec ${}_{i}X = 58{,}7$,

(6a) $\Pr( X_{Robert} \in \{53{,}90; 63{,}50\} ) \approx 0{,}95$ .

En mots, si on mesurait Robert un très grand nombre de fois, 95 pour cent des intervalles obtenus, tel l'intervalle {53,90; 63,50}, enfermeraient la valeur réelle de Robert. Ainsi, considérant l'erreur de mesure et selon un coefficient de confiance de 0,95, on peut donc affirmer que, pour Robert, il est plausible que sa valeur réelle atteigne au moins le seuil requis de 60 %.

**La réponse casse-tête**

Qu'est-ce qui, d'une fois à l'autre, fait varier le résultat de Robert *même alors qu'il s'agirait d'un test sans erreur de mesure ni effet d'apprentissage ou de fatigue*? Mettons les choses au clair, en distinguant deux classes de mesures: les mesures d'un caractère inanimé, et les mesures «vivantes», représentant un phénomène vivant et animé. Ainsi, soit ${}_{1}T$, ${}_{2}T$, ${}_{3}T$, *etc.*, une série de mesures de la taille (T) de Robert, en centimètres (cm). Prise dans des conditions convenables, la longueur de taille peut être réputée *fixe* (*i.e.* inanimée) et, si les mesures varient quelque peu d'une fois à la l'autre, cette variation est toute assignable à une erreur instrumentale (due à l'instrument lui-même ou à sa procédure d'utilisation[1]), disons $\varepsilon_M$. On voit ainsi que, pour la taille,

(7) $${}_{i}T_{Robert} = {}_{T}V_{Robert} + {}_{T}\varepsilon_{M,i} ,$$

la *mesure* de taille étant l'addition d'une valeur fixe (caractéristique de Robert), en cm, et d'une petite quantité fluctuante ($\varepsilon_M$), dépendant seulement de l'instrument de mesure. En corollaire, on peut écrire

(8) $${}_{i}T_{Robert} \to {}_{T}V_{Robert} \quad \text{si} \quad {}_{T}\varepsilon_{M,i} \to 0 ;$$

dans ce cas, si l'erreur de l'instrument de mesure était nulle, les mesures prises seraient toutes égales les unes aux autres et coïncideraient avec la «valeur vraie» de Robert, sa taille réelle.

Qu'arrive-t-il de différent avec l'examen de mathématique? Il y a ici aussi de l'erreur instrumentale (au sens large), de sources aussi diverses que l'échantillon de questions ou de problèmes choisis, la pondération des objectifs retenus dans l'examen, voire le nombre de questions. Mais, au-delà de l'imprécision qui découle certainement de l'examen administré, le

répondant lui-même est de disposition changeante et il ne peut pas, comme pour la taille corporelle, être caractérisé par une valeur unique, ou par une valeur fixe. En d'autres mots, même si l'erreur de mesure ($\varepsilon_M$) était abrogée ou nulle, Robert (le répondant) produirait quand même tout un répertoire de résultats, une *distribution* en langage statistique. Dans ce cas, et en dénotant par ${}_iX_{Robert}$ une $i^e$ mesure (hypothétique) de Robert en mathématique,

$$(9) \qquad {}_iX_{Robert} = {}_XD(Robert) + {}_X\varepsilon_{M,i} ,$$

c'est-à-dire qu'à chaque fois que Robert est mesuré, un résultat émane de la distribution (${}_XD$), ou répertoire de valeurs X, caractéristique de Robert, ce résultat étant contaminé par une erreur de mesure ($\varepsilon_M$) proprement dite.

> La différence fondamentale entre ces deux concepts, d'une part la grandeur (fixe) d'un caractère inanimé, d'autre part la distribution des grandeurs du répertoire d'un caractère vivant, tient peut-être à la dimensionnalité élevée de ce dernier, à l'abondance de ses déterminants. Dans une occasion de mesure donnée, la grandeur mesurée reflète, chez le caractère vivant, un équilibre momentané des déterminants, une homéostasie au sens large, grâce à quoi l'organisme évalué tente, à travers tous ses caractères, de maintenir son adaptation aux environnements interne et externe.

Les auteurs classiques en théorie des tests identifient et reconnaissent explicitement les sources de variation mentionnées plus haut. De plus, ils prennent une *option pratique* qui consiste à réduire ce deuxième cas (concernant la mesure des caractères animés) au premier, comme suit. Si l'on prend (théoriquement) un très grand nombre de mesures de Robert et que l'on en calcule la moyenne arithmétique, il se produit deux résultats intéressants: (1) la portion de mesure assignable à l'erreur $\varepsilon_M$, qui fluctue autour de zéro avec une amplitude $\sigma_M$, [car $\sigma(\bar{\varepsilon}_M) = \sigma_M/\sqrt{n} \to 0$ si $n \to \infty$] et vient à disparaître; (2) la moyenne des mesures tend en même temps vers la *moyenne de la distribution*, la moyenne des résultats du répertoire de Robert, disons ${}_X\bar{V}_{Robert}$. On pourrait donc écrire

$$(10) \qquad {}_iX_{Robert} = {}_X\bar{V}_{Robert} + {}_X\varepsilon_{P,i} + {}_X\varepsilon_{M,i} ;$$

en mots, la $i^e$ mesure de Robert nous renseigne sur la moyenne $\bar{V}$ de son répertoire de réponses, et elle reflète aussi une variation propre ($\varepsilon_P$) typique du caractère mesuré en plus d'une erreur de mesure $\varepsilon_M$ typique de l'instrument de mesure. Le caractère mesuré (ici, l'habileté en mathématique, ou bien la force musculaire, la tension artérielle, l'intelligence, etc.) étant vivant,

il fluctue; ces fluctuations lui sont *propres*, et elles ne peuvent être éliminées par un procédé statistique ou autre. Dans une autre approche, celle de la *théorie de la généralisabilité* (Cardinet & Tourneur, 1985; Cronbach, Gleser, Nanda & Rajaratnam, 1972), la fluctuation observée dans les mesures est théoriquement imputée à un ou plusieurs facteurs, des facteurs statistiquement (ou expérimentalement) contrôlables et à chacun desquels on peut assigner une composante spécifique de la variance d'erreur. Cette approche, foncièrement déterministe, a néanmoins des points d'affinité avec celle que nous avançons: les deux approches *séparent* la fluctuation observée en composantes différentes, et elles traitent semblablement l'accumulation des mesures (comme dans notre équation (17)).

Nonobstant les considérations précédentes, les auteurs classiques confondent généralement erreur instrumentale et fluctuation propre (*i.e.* $\varepsilon = \varepsilon_M + \varepsilon_P$) et font *comme si* l'agrégat jouissait des propriétés statistiques de l'erreur de mesure. Notre opinion est qu'ils commettent là une simplification hâtive, et nous avançons le double exemple suivant pour nourrir la polémique.

Cas 1. Robert est-il plus grand qu'Antoine? Supposons tout de suite que la mesure est prise en cm, que l'erreur-type du procédé de mesure est dénotée $\sigma_M$ et que nous pouvons mesurer Robert et Antoine un nombre quelconque de fois. Si nous disposons d'une seule mesure de la taille de Robert ($T_R$) et d'Antoine ($T_A$), la théorie, qui tient compte de l'erreur de mesure, nous dicte le test

$$(11) \qquad [T_R - T_A] \,/\, [\sigma_M\sqrt{2}] \,,$$

qu'on peut interpréter approximativement comme une variable normale centrée réduite (*i.e.* significative à 5 % si sa valeur absolue excède 1,960). Avec $n_R$ mesures de la taille de Robert et leur moyenne $\bar{T}_R$, et avec $n_A$ et $\bar{T}_A$ pour Antoine, la comparaison des tailles sera plus fine (si $n_R > 1$ ou $n_A > 1$). En effet, les erreurs dans chaque mesure, qui brouillent la différence réelle entre les longueurs de Robert et d'Antoine, sont ici «moyennées» et perdent de l'amplitude. En un mot, le test (en général) devient

$$(12) \qquad [\bar{T}_R - \bar{T}_A] \,/\, [\sigma_M\sqrt{(1/n_R + 1/n_A)}] \,;$$

s'il est statistiquement significatif, on l'interprète en affirmant que la différence des tailles de Robert et Antoine est vraiment non nulle et n'est pas assignable à l'effet des erreurs de mesure.

Cas 2. Robert mérite-t-il de passer en mathématique? En admettant que la note d'examen soit susceptible d'une fluctuation quelconque, d'amplitude (ou erreur-type) $\sigma$, le test

(13) $$[X_R - 60] / \sigma$$

permet de vérifier si, les fluctuations aidant, Robert s'écarte sérieusement ou non de la note de passage 60 %. Qu'arrive-t-il maintenant si nous pouvons mesurer Robert *n* fois? Imaginons un enseignant très consciencieux qui, dans le but d'évaluer plus précisément ses élèves, fabriquerait *n* (= 5 ou 20, par exemple) examens équivalents. Supposons encore qu'on puisse déterminer indépendamment la variance d'erreur instrumentale ($\sigma_M^2$) et la variance propre ($\sigma_P^2$). Cette détermination est très faisable dans plusieurs cas mais elle est difficile ici. Quoi qu'il en soit, nous pouvons administrer à Robert les *n* examens (dans des conditions de passation équivalentes), et nous en obtenons la moyenne $\bar{X}_R$, disons $\bar{X}_R = 58{,}7$. Les auteurs classiques (qui confondent erreur instrumentale et fluctuation propre) utilisent une variance unique,

(14) $$\sigma^2 = \sigma_M^2 + \sigma_P^2$$

avec le test

(15) $$[\bar{X}_R - 60] / [\sigma/\sqrt{n}] .$$

S'il est significatif (et, pour Robert, il le deviendrait si on accroît suffisamment le nombre *n* d'examens équivalents), ce test indique que la moyenne de la distribution caractéristique de Robert est inférieure au seuil de 60 %. D'un autre côté et selon nous, l'évaluation répétée de Robert a deux fonctions distinctes: elle permet de réduire en moyenne l'impact de l'erreur de mesure ($\varepsilon_M$) *et* elle permet de mieux délimiter le répertoire (et l'amplitude) des réponses possibles de Robert, en particulier $\sigma_P^2$. On peut en effet, à partir des *n* mesures de Robert, calculer aussi une variance échantillonnale $s_X^2$, laquelle combine, comme dans l'équation (15), la variance d'erreur et la variance propre. Si nous possédons un bon estimé de la variance d'erreur, obtenu indépendamment, alors nous pouvons effectuer

(16) $$\hat{\sigma}_P^2 = s_X^2 - \sigma_M^2 .$$

Quoi qu'il en soit, le test approprié serait plutôt

(17) $$[\bar{X}_R - 60] / [\hat{\sigma}_P + \sigma_M/\sqrt{n}] ;$$

ce test, tout différent de (16), sera dit significatif seulement si le *répertoire* (des réponses) de Robert n'atteint pas (en proportion suffisante) la marque des 60 %. Avec un nombre ($n$) infini d'évaluations et si $\bar{X}_R$ = 58,7 %, seule l'erreur de mesure (moyenne) vient à s'évanouir, et il y a peut-être lieu d'admettre le résultat de Robert, qui par contre serait refusé en appliquant la règle indiquée en (15).

En un mot, si comme en science la «valeur vraie» d'un caractère est ce qui reste après qu'on ait éliminé l'erreur de mesure, cette «valeur vraie» (pour un objet donné) est un *nombre* lorsque le caractère mesuré est inanimé, ou une *distribution statistique* lorsque le caractère fluctue de lui-même (dans des conditions stables).

## Question 2

*Marc-André obtient des scores de 107 et 95 respectivement au Q.I. verbal et au Q.I. non verbal: que peut-on conclure de la différence entre ces scores?*

### La réponse simple

Il faut supposer, disons-le tout de suite, que nous disposons d'un test de Q.I. de qualité et récemment standardisé[3]: chaque type (verbal, non verbal) aurait pour moyenne imposée $\mu$ = 100 et pour écart-type $\sigma$ = 15 (par exemple). Dans la population ou, mieux encore, dans la catégorie de personnes appropriée à Marc-André, les indices de Q.I. verbal et non verbal *moyens* sont chacun de 100. L'écart de 107 à 95, pour Marc-André, est surprenant. Allouant un jeu de ±1 point pour chaque type de Q.I., on pourrait s'inquiéter de la faiblesse relative suggérée par l'indice non verbal du test.

### Les réponses du livre

Malgré la surprise qu'il pourrait éprouver devant les résultats disparates de Marc-André, le psychologue (ou l'intervenant en général) en a vu d'autres et il sait que les mesures sont fluctuantes. Il y a donc ici un contexte de *fluctuations* à considérer et la réponse qu'on obtiendra dépend du contexte, de la question spécifique qu'on se pose.

*Dans un premier temps*, le spécialiste est conscient que le test utilisé n'est pas absolument précis, que la fidélité de chaque sous-quotient n'est pas parfaite. Dénotant par $\sigma_X$ et $\rho_{XX}$ l'écart-type et le coefficient de fidélité du Q.I. verbal, par $\sigma_Y$ et $\rho_{YY}$ ceux du non verbal, alors, les erreurs de mesure étant indépendantes d'une partie à l'autre du test, la variance d'erreur de la différence (X - Y) égale la somme des variances d'erreur des composantes, d'où on obtient l'erreur-type

$$(18) \qquad \sigma_e(X-Y) = \sqrt{[\sigma_e^2(X) + \sigma_e^2(Y)]} = \sqrt{[\sigma_X^2(1-\rho_{XX}) + \sigma_Y^2(1-\rho_{YY})]} .$$

Si, comme dans notre exemple, $\sigma_X = \sigma_Y = \sigma$, alors

$$(19) \qquad \sigma_e(X-Y) = \sigma\sqrt{(2 - \rho_{XX} - \rho_{YY})} .$$

Ainsi, le test

$$(20) \qquad [X - Y] / \sigma_e(X-Y) ,$$

s'il est significatif (*i.e.* s'il déborde l'intervalle ±1,960), nous indique si la différence (X - Y) est solidement plus grande que zéro, en d'autres mots si notre ami Marc-André risque de conserver ce décalage d'une évaluation à l'autre.

*Dans un second temps*, quand le spécialiste prend la peine de calculer la différence (X - Y) entre les deux types de Q.I., quelle interprétation spécifique entend-il donner à cette différence? Nous croyons que, outre le fait de s'assurer que cette différence n'est pas seulement attribuable à de l'erreur de mesure, le spécialiste veut aussi savoir si une telle différence (X - Y) est *exceptionnellement grande*, plus grande que celles qu'on retrouve dans la population. Reprenant les termes d'un article récent (Laurencelle, 1995), «L'univers de généralisation n'est plus constitué seulement des *occasions de mesure*, mais plutôt des *personnes*, et l'erreur-type doit alors tenir compte de la corrélation (dans la population) entre X et Y» (p. 130). Il s'agit en fait de l'écart-type des différences (X - Y) obtenues dans la population,

$$(21) \qquad \sigma(X-Y) = \sqrt{[\sigma_X^2 + \sigma_Y^2 - 2\rho_{XY}\sigma_X\sigma_Y]}$$

ou, comme pour notre exemple avec $\sigma_X = \sigma_Y = \sigma$,

$$(22) \qquad \sigma(X-Y) = \sigma\sqrt{[2 - 2\rho_{XY}]} .$$

Cette mesure de variabilité, $\sigma(X-Y)$, sera généralement plus grande que l'erreur-type de tantôt: elle lui serait à peu près égale seulement si X et Y mesuraient exactement le même caractère chez les personnes[4], ce qui est moins que vraisemblable pour les types de Q.I. Ainsi, le test

$$(23) \qquad [X - Y] / \sigma(X-Y)$$

nous indiquera s'il y a lieu de s'alarmer dans le cas de Marc-André ou si une différence telle que (107-95) entre les indices de Q.I. se retrouve habituellement dans la population.

Comme quoi l'exploitation judicieuse des *erreurs-types* (reflétant la variabilité d'une occasion de mesure à l'autre) ne doit pas nous faire oublier les *écarts-types* (qui reflètent la variation de mesure entre les membres d'une population de référence).

## Question 3

*Comment, dans une situation pratique, obtenir un bon estimé de la variance d'erreur?*

**La réponse simple**

Un physicien, un technicien en sciences appliquées qui effectue une mesure au moyen d'un instrument calibré et affichant l'unité de mesure $u$ prendra comme champ d'erreur l'intervalle $(X \pm \frac{1}{2}u)$. En général, toutefois, et en sciences humaines en particulier, l'amplitude de l'erreur est à la fois plus grande que $\frac{1}{2}u$ et plus difficile à estimer. Restreignant notre discussion aux applications de tests et d'examens (typiques surtout de la psychométrie et de la docimologie), nous retrouvons quelques méthodes d'*estimation* de la variance d'erreur et de l'erreur-type de mesure. Si le spécialiste peut administrer deux fois le même test, ou deux tests exactement comparables, dans des conditions semblables pour les $n$ sujets, alors il disposera des données sommatives suivantes (en dénotant ${}_1X$ la première mesure d'un sujet et ${}_2X$ la seconde):

$$\{ n, {}_1\bar{X}, {}_1S, {}_2\bar{X}, {}_2S, r_{12} \} ;$$

${}_1S$ et ${}_2S$ sont les écarts-types, et $r_{12}$ est le coefficient de corrélation entre les deux séries de mesures. Le théorie des tests nous informe que cette

corrélation $r_{12}$ est ici un estimateur du coefficient de fidélité $\rho_{XX}$. Considérant l'équation (5) plus haut, un estimateur évident s'impose pour l'erreur-type, estimateur que l'on retrouve dans tous les textes classiques,

$$\hat{\sigma}_M = {}_1S\sqrt{(1-r_{12})} \,. \qquad \text{Est. A} \tag{24}$$

Parmi d'autres estimateurs à partir des mêmes données, celui qui est basé sur les différences $({}_1X - {}_2X)$,

$$\hat{\sigma}_M = S_{({}_1X-{}_2X)}/\sqrt{2} \qquad \text{Est. B} \tag{25}$$

mérite d'être signalé. Mentionnons aussi les estimateurs *internes*, basés sur la décomposition du test X en parties pseudo-équivalentes (sous forme $X = y_1 + y_2 + \cdots + y_k$), et donnant lieu à l'estimation de fidélité par la «méthode des moitiés» ou par le coefficient «Alpha» de Cronbach (voir par exemple Bernier, 1985).

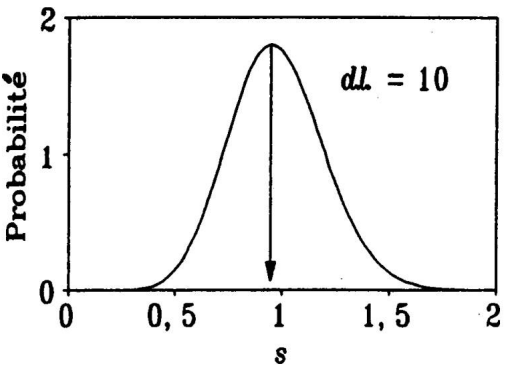
**La réponse du statisticien**

Spécialistes et praticiens du *testing* et de la mesure sont, pour la plupart, conscients que les mesures obtenues fluctuent d'une fois à l'autre pour une même personne ou pour un même objet. Les cours de statistique, de plus, nous ont sensibilisés au problème plus subtil de la variabilité des moyennes, d'où l'utilité du test *t* pour décider d'une différence entre moyennes ou l'analyse de variance. Mais il faut aussi être clairement conscient que tous les indices statistiques fluctuent, y inclus l'écart-type et le coefficient de corrélation, ceux qui sont impliqués dans nos techniques d'estimation de l'erreur-type de mesure.

Le problème de fluctuation, si on peut dire, se pose donc pour les estimateurs de l'erreur-type, qui reposent eux aussi sur des données fluctuantes. Les formes que prend la fluctuation des différents estimateurs d'erreur-type dépassent pour la plupart les possibilités d'une description mathématique rigoureuse. Fait exception, par exemple, l'estimateur B, en formule (25) ci-haut, qui a une distribution issue de la «loi du Khi-deux»[5], avec le paramètre des *degrés de liberté* (*d.l.*). La figure suivante illustre la distribution des valeurs possibles d'un écart-type basé sur *d.l.* = 10 (en supposant ici une valeur de référence $\sigma = 1$); cet écart-type (ou erreur-type (25)), pourrait avoir été calculé à partir d'une série de $n = 11$ différences (X - Y).

Une figure vaut mille mots; néanmoins, transcrivons en quelques mots les principaux éléments que nous avons à vue. Premièrement, nous observons

toute une série de valeurs à probabilités plus ou moins grandes, non pas une seule valeur ni forcément *la* bonne valeur (qui serait ici $\sigma = 1$). Deuxièmement, la distribution n'est pas tout à fait symétrique ni centrée sur sa moyenne. En fait, avec $d$ degrés de liberté, la moyenne est à peu près égale à $1 - 1/(4d)$ et le mode exactement égal à $\sqrt{[(d-1)/d]}$; il s'agit donc d'une statistique *biaisée*. Ces valeurs de moyenne et de mode, en plus de la médiane, se rapprochent de 1 ($= \sigma$) lorsque $d$ augmente. Troisièmement, l'intervalle de fluctuation de l'écart-type $s$ diminue lui aussi lorsque $d$, c'est-à-dire, le nombre de données utilisées, augmente. Cette imprécision de l'écart-type (ou de l'erreur-type) peut en fait être évaluée par le *coefficient de variation* (CV), soit le quotient de l'écart-type d'une statistique (telle que $s$ ou $\hat{\sigma}_e$) sur sa moyenne. En ce qui concerne l'écart-type avec *d.l.* degrés de liberté, on a CV $\approx 1/\sqrt{(2d.l.)}$.[6] Avec *d.l.* = 10 comme dans la figure ci-haut, CV $\approx 1/\sqrt{20} \approx$ 0,224; c'est dire que la valeur obtenue avec 10 *d.l.* (ou $n$ = 11 différences X-Y) fluctue selon une marge typique d'environ 22 % de la valeur véritable. Il faudrait 50 *d.l.* pour réduire cette marge de variation à 10 %, et 5000 *d.l.* pour la réduire à 1 %.




**La réponse pratique**

L'étude des propriétés statistiques des différents estimateurs de l'erreur-type en théorie des tests échappe aux méthodes mathématiques classiques. Pour pallier cette difficulté, nous avons procédé par une étude *Monte Carlo* à l'évaluation de ces propriétés, c'est-à-dire par programmation informatique et en créant des populations de données fictives au moyen de nombres pseudo-aléatoires. De cette étude (Allaire & Laurencelle, en préparation), qui comparait six estimateurs différents de la variance d'erreur (et de l'erreur-type), voici l'essentiel.

En plus des estimateurs A (formule 24) et B (formule 25) déjà présentés, il vaut la peine de mentionner l'estimateur C suivant,

(26) $\hat{\sigma}_M = \sqrt{[{}_1S_2S(1 - r_{12})]}$ Est. C

qui exploite le même ensemble de données que l'estimateur A, en «récupérant» le second écart-type ${}_2S$. Il s'avère donc, et ce sont nos recommandations conclusives,

1) qu'il y a plusieurs estimateurs différents de la variance d'erreur et que pratiquement tous sont biaisés: la plupart *sous-estiment* la valeur cible. L'estimateur B n'est presque pas biaisé. L'estimateur C affiche un biais de sous-estimation encore plus fort que l'estimateur A;

2) que, parmi tous les estimateurs considérés, les estimateurs B et C sont dans les plus précis, c'est-à-dire qu'en moyenne, ils s'éloignent le moins de la valeur cible. L'estimateur A, celui qu'on retrouve habituellement dans les manuels et les «recettes de calcul» et qui exploite un seul des deux écarts-types (${}_1S$, ${}_2S$) disponibles, se classe défavorablement.

Y a-t-il d'autres questions dans la salle?...

## NOTES

1. Parmi les sources de l'erreur instrumentale, mentionnons l'universelle erreur de lecture, ou erreur due à l'*unité de mesure* «*u*», les mesures rapportées par un instrument (ou par un nombre) l'étant à des multiples de $u$: par exemple $u$ = 1 cm, ou $u$ = 0,1 % (un dixième pour cent), ou $u$ = 1 point dans un questionnaire. On peut montrer que la portion de variance (d'erreur) attribuable à cette source est de $u^2/12$.

2. Le symbole $\sigma_M$ est introduit ici pour désigner l'écart-type des fluctuations numériques attribuables entièrement à l'opération de mesure, par contraste au symbole $\sigma_P$, évoqué plus loin, qui caractérise l'ampleur (la dispersion) du répertoire des grandeurs possibles d'un caractère vivant.

3. Le premier auteur, dans une étude complémentaire réalisée en 1974-1975 auprès des clientèles de prématernelle, maternelle et primaire de la Commission des écoles catholiques de Montréal (Laurencelle & Joyal, 1975), a démontré des décalages normatifs importants en Q.I. verbal et Q.I. non verbal du (premier)

WISC (normes américaines), allant jusqu'à un écart de 15 points pour certain groupe d'âges.

4. C'est-à-dire si la corrélation entre les «valeurs vraies» de X et Y était parfaite. En fait, la corrélation mesurable entre X et Y est *atténuée* par leurs fidélités imparfaites, et nous avons (en dénotant par $_{V}\rho_{XY}$ la corrélation entre les valeurs vraies, dépourvues d'erreur) $\rho_{XY} = {}_{V}\rho_{XY}\sqrt{(\rho_{XX}\rho_{YY})}$. En prenant approximativement $\sqrt{(\rho_{XX}\rho_{YY})} \approx \frac{1}{2}(\rho_{XX}+\rho_{YY})$ et selon $\sigma_X = \sigma_Y = \sigma$, nous obtenons $\sigma(X\text{-}Y) \approx \sigma\sqrt{[2-{}_{V}\rho_{XY}(\rho_{XX} + \rho_{YY})]}$. Enfin, faisant l'hypothèse (peu plausible) que ${}_{V}\rho_{XY} \to 1$, alors $\sigma(X\text{-}Y) \to \sigma\sqrt{[2-\rho_{XX}-\rho_{YY}]} = \sigma_e(X\text{-}Y)$, comme en (19).

5. Il s'agit en fait de la loi «Khi», correspondant à la racine carrée du «Khi-deux». Ces deux lois stipulent que l'erreur de mesure a une distribution normale.

6. Une approximation plus serrée est $CV \approx (8d.l. - 1)/[\sqrt{(8d.l.)(4d.l.-1)}]$.

## RÉFÉRENCES


Allaire, D. & Laurencelle, L. (en préparation). La variance d'erreur et ses estimateurs en théorie des tests: une étude Monte Carlo.

Anastasi, A. (1982). Psychological *testing* (5e édition). New York: Macmillan.

Bernier, J.-J. (1985). Théorie des tests (2e édition). Chicoutimi: Gaëtan Morin.

Cardinet, J. & Tourneur, Y. (1985). Assurer la mesure. Berne: Peter Lang.

Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). The dependability of behavioral measurements. New York: Wiley.

Gulliksen, H. (1950). Theory of mental tests. New York: Wiley.

Johnson, N.L., Kotz, S. & Balakrishnan, N. (1995). Continuous univariate distributions (2 volumes)(2e édition). New York: Wiley.

Laurencelle, L. & Joyal, J.-P. (1975). Normes-CECM du WISC pour les âges de 5 ans 8 mois à 7 ans 7 mois. Rapport d'étude déposé à la Commission des écoles catholiques de Montréal.

Laurencelle, L. (1995). Recension de «Introduction à la psychométrie» (traduction Françoys Gagné) d'Anne Anastasi, 1994. Mesure et évaluation en éducation, 17(3), 127-133.

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

*Les auteurs tiennent à remercier les évaluateurs anonymes pour leurs commentaires et suggestions. Leur contribution a permis de clarifier et, dans un cas, d'élargir le propos de l'article.*