Alors vous voulez construire//valider//utiliser un test ?

Marc Bourdeau [1]

The best thing about being a statistician
is that you get to play in everybody's backyard.
John Tukey [1915 – 2000]

En entrée de jeu...

- Je suis un *praticien* de l'analyse des questionnaires psychométriques
- Les questions épistémologiques y référant m'attirent un peu, mais ce n'est pas à moi de creuser ces questions
  - Mes lectures en catastrophe des dernières semaines m'ont rendu quelque peu perplexe à cet égard, enfin sur mes capacités... et beaucoup intrigué/intéressé ! Surtout au développement historique
  - Je vais me baser sur des écrits simples et ...faire des renvois à une littérature que j'espère pertinente
- Je vais passer plus de temps sur les questions de validation *statistique* des modèles
- Et surtout utiliser tout au long de l'exposé des cas réels pour illustrer quelques difficultés de la validation & de la construction des tests

Cliquer l'icone à droite pour un ensemble de références qui semblent intéressantes au second auteur. Celles qu'on utilisera dans la suite de l'exposé sont téléchargeables ; la plupart seront citées dans les références à la fin du texte.

Références

1. Professeur associé, École Polytechnique de Montréal. Ce texte, avant d'être développé considérablement, a d'abord servi pour un séminaire du Cdame le 25 novembre 2013 (*C*ollectif pour le *d*éveloppement des *a*pplications en *m*esure & *é*valuation de l'Uqàm). Il a été préparé dans sa toute première mouture en collaboration avec Sébastien Béland, chargé d'enseignement à la faculté des Sciences de l'éducation de l'Université de Sherbrooke. Les textes en bleu sont des hyperliens. Les références en rouge sont des inter-liens : ≪ Alt-Flèche à gauche ≫ revient au texte courant.

Mesurer

There is no true value of anything.
W.E. Deming [2] [1900 – 1993]

Épistémologie...

- La philosophie de la connaissance se pose des questions depuis toujours sur la mesurabilité de quoi que soit
- Il n'est de science que quantitative... mais ce n'est pas le seul mode de connaissances
- Rien n'est donné, tout est *construit*
- On admet la plupart du temps sans se poser des questions qu'on peut mesurer ce qu'on veut mesurer et que les tests qu'on utilise, éventuellement qu'on a construits (!), sont légitimes : on utilise les échelles données, éventuellement on se choisit des échelles sans trop penser aux fondements, on procède.
- Il convient d'avoir quelque notions de métrologie, surtout en sciences humaines et sociales (SHS), où les questions se posent de façon récurrente : « Est-ce que je peux mesurer ce que je voudrais mesurer ? »

En tant que consultant statistique, j'arrive le plus souvent en aval, je n'ai jamais eu de questions à cet égard, je prends *ce qui est*.... La même chose pour tout le monde ...sauf ceux intéressés pas l'épistémologie ?

En sciences humaines et sociales (SHS), on devrait tout de même se rafraîchir les connaissances à cet égard.

On pourra consulter l'introduction (Blais & Raîche) et le premier chapitre (Blais) dans Blais & Raîche (2003) qui présentent et traitent de la question de façon très synthétique, et très succinctement.

On peut aussi se rapporter aux références citées pas ces auteurs, ou encore à la bibliographie hyper-référencée plus haut.

2. Tiré de sa préface au livre de Walter Andrew Shewhart [1891 – 1967], *Statistical Method from the Viewpoint of Quality Control.* (1938), réédité (1986) avec cette nouvelle préface par Dover Publications, New York NY.

- En SHS, les construits sont la plupart du temps très complexes : ils reposent sur des théories psychologiques, sociologiques etc, souvent à base philosophique ; elles ne sont pas toutes concordantes, il n'y a pas d'*expérimentation* et de *falsifiabilité* possible...
- On ne se pose pas souvent de questions sur la façon (la légitimité) d'associer des valeurs numériques à des caractéristiques, des attributs des sujets
- Tukey [a] notait à juste titre qu'on négligeait le processus de génération des données et de la mesure
  - La question de l'échantillonnage et de l'expérimentation est cruciale en statistique, et particulièrement délicate en SHS ! Comment en effet *randomiser* ?..

a. John Wilder Tukey (1961), The statistical and quantitative methodology, *Trends in the social sciences*, New York NY : Philosophical Library.

Allons voir quelques détails des construits de HARDIESSE et de COPING qui nous servent d'illustrations dans cet exposé :

La Hardiesse

Le Coping

Tests sur des aptitudes et tests sur des attitudes

Distinguons deux types de tests

- Fishbein et Ajzen (1975) ont écrit, à la suite de Summers (1970), ce qui est devenu un classique : il a fallu plus de 500 pages pour cerner et distinguer, expliquer comment mesurer les *Belief, Attitude, Intention, Behavior*...
- Disons qu'on distingue provisoirement les aptitudes où on a une réponse dite bonne et d'autres qui ne le sont pas
- alors que pour les attitudes, que ce soit devant des 'objets' externes ou internes, i.e. psychologiques, il n'y a pas de bonne ou de mauvaise réponse.

- Ainsi les construits de hardiesse et de coping ont donné lieu à des tests d'attitudes : aucune réponses n'est la bonne, les échelles de mesure sont le plus souvent des échelles de Likert.
- Les méthodologies des TRI (issues des modèles logistiques) ont été dévoppées pour tester des aptitudes
- Alors que les méthodologies classiques (ou TCT) sont appropriées pour mesurer les attitudes
- les frontières sont maintenant assez brouillées : les méthodes de TRI sont très pratiquées pour les tests d'attitudes avec des échelles de Likert pour les réponses.

Nous ne présenterons que des cas de tests sur des attitudes

Du prêt à porter

- Il est toujours préférable de traduire un test plutôt que de se donner la grande peine d'en construire un...
- Il faut comme toujours être très circonspect : attention surtout aux niveaux de langage, aux capactités mentales des sujets à qui on s'adresse pour formuler des items//questions.
- Une procédure courante consiste à valider une traduction par une rétro-traduction : services professionnels coûteux...

De toute façon, traduire ou construire sont des processus longs et coûteux. Cliquer ci-dessous pour quelques ressources sur les tests existants, ainsi que les ressources psychométriques du logiciel R :

Ressources

Du sur-mesure

Construire du neuf pour des besoins neufs. Procédure longue et coûteuse, et pas souvent pratiquée.

On a vu des tests faire du chemin après avoir été validés sur moins de 50 sujets... Difficile d'obtenir des renseignements des auteurs !

On a vu un cas de test validé par un envoi postal à de vieilles personnes dans la campagne profonde, avec un taux de retour inférieur à 10%, et qu'on a voulu utiliser sur un groupe d'infirmières jeunes à Montréal... Les données furent recueillies avant d'examiner le test...

Prudence : aller aux sources, utiliser des tests déjà reconnus, bétonnés par une littérature de qualité

autrement re-valider//construire *from scratch*...

Voici une procédure qui se veut générale de construction de tests sur des attitudes comme celles qu'on a déjà présenté :

Construire

Enfin, quelques lignes directrices pour la rédaction d'items en deux fichiers, dont un *check list* de Gilles Raîche :

Sensibilité, Fiabilité & validité

L'intelligence ? C'est ce que mesure mon test...
Alfred Binet [1857 – 1911]

Est-ce qu'un test remplit sa fonction ? Comment répondre à cette question alors qu'on ne sait pas vraiment ce qu'on mesure, pire qu'on cherche à comprendre le concept//construit...

Pensons à l'intelligence, à l'introversion, on veut mesurer ces choses ? Allons donc ! et encore, de façon fiable, valide, qui réussit à distinguer deux individus distincts sous le rapport mesuré... Tâche impossible ?

Pensons à toutes les contraintes sociales qui s'exercent sur les individus qui ont à répondre à des dizaines de questions, alors que leur situation personnelle peut être précaire, insécure, dans une mauvaise passe, une mauvaise journée, etc.

Et pourtant il le faut !

Nous ferons simple. Commençons par retrouver quelques concepts de métrologie pratiqués en sciences appliquées.

On retrouve les concepts de *fiabilité* et *validité*. Donnons-en des définitions sommaires :

Définition. **Fiabilité :** c'est la capacité un test de redonner aux mêmes individus les mêmes valeurs numériques lors de prises de mesures répétées et indépendantes [a].

a. On dit parfois fidélité en psychométrie.

Trois techniques de vérification :

1. Test-retest
2. Utiliser des items de formes alternatives
3. La méthode par moitiés

Définition. **Validité.** Le test mesure-t-il ce qu'il est censé mesurer ? C'est la question cruciale...

Là c'est nettement plus délicat. On distingue de nombreuses sortes de validité, qu'on peut grossièrement regrouper en trois :

1. Validité de contenu : recouvre-t-on le concept ? (e.g. l'intelligence, la hardiesse, le Coping, l'introversion). Identifer le *contenu* au moins en théorie, ses dimensions, ses sous-dimensions ; recourir à des experts ; choix au hasard des items ; plusieurs tests possible à comparer...
2. Validité de construit : retour sur la théorie, la littérature... Évaluation par des experts réunis en panel ; comparaison des résultats avec des tests présumés voisins...
3. Validité par un critère : utiliser des groupes de sujets qui possèdent la qualité désirée telle qu'identifiés par des experts.
4. Validation concomitante : voisine de la précédente en ce sens qu'on utilise un *devis corrélationnel* : plusieurs tests voisins sont utilisés, et les corrélations positives entre les résultats servent de validation.

On touche ici à la nécessité de montrer la *validation discriminante*, i.e. comment décrire la proximité des tests, montrer les nuances entre plusieurs concepts voisins (e.g. la résilience & la hardiesse, souvent confondues).

Définition. **Sensibilité.** La *sensibilité* d'un test s'appelle la *résolution* d'un outil d'observation dans les sciences naturelles et appliquées. Deux sujets avec des mesures différentes sont-ils différents ?

Remarque. **on commence à comprendre ici la difficulté dans la pratique de la conception de tests et de leur « mise en marché. »**

Inutile de dire qu'on tourne souvent les coins ronds. Quantité de tests courants sont mal validés.

Remarque. **En réalité, la validation est un processus jamais terminé. Chaque nouvelle utilisation, dans des contextes différents, des échantillons autres que les échantillons princeps, demande un certain passage par la validation.**

Tout utilisateur d'un test se doit d'aller consulter toute la littérature sur celui-ci.

La *consistance interne.* Enfin, un mot sur la *consistance interne* des tests où on trouve un bon nombre d'indicateurs (voir le *package* Psych de R) dont le célèbre α de Cronbach qui n'est rien d'autre qu'un ajustement de la moyenne des corrélations entre paires d'items d'un test :

$$\alpha = \frac{N\overline{r}}{[1 + \overline{r}(N-1)]},$$

où N est le nombre d'items dans le test (ou dans une dimension du test), et $\overline{r}$ est la moyenne des corrélations inter-items du test (dimension).

Toute la question est de se donner un seuil à partir duquel on peut se dire satisfait de la consistance interne. Le α est empirique, le seuil aussi...

Remarque. On entend souvent qu'à partir de $\alpha = 0,70$ on est en affaires, et qu'avec un $\alpha \geq 0,80$ on peut être heureux. Mais on a souligné maintes fois que la valeur du α est très dépendante du nombre d'items. Ainsi pour $N = 10$, on a un $\alpha = 0,71$ (respectivement 0,87) — ces α augmentent d'ailleurs avec N — avec un $\overline{r} = 0,2$ (respec. 0,4), et qui pourra dire qu'une corrélation de 0,2 est forte ? (Carmines & Zeller, 1976, p. 46). Et des N de l'ordre de 10 sont rarissimes, on en a toujours plus ! Cela semble rédhibitoire pour l'utilisation de l'alpha (Revelle & Zinbarg, 2009 ; Schmitt 1996 ; Sijtsma, 2009). Et pourtant nul ne peut passer à côté dans la plupart des publications. DeVellis (2013) en fait encore l'apologie avec des interprétations assez poétiques des corrélations... — Voir plus loin pour un court développement sur la géométrie des corrélations.

Remarque. Il arrive qu'on place ce concept à l'intérieur de la fiabilité, qui porte mieux en ce sens le terme de *fidélité*.

Qui prend mieux son sens lorsqu'on utilise le *alpha* de Cronbach successivement sur un ensemble potentiel d'items *moins un*, dans le but d'en choisir un nombre plus limité en éliminant successivement celui qui réduit le plus ce coefficient calculé avec la technique du *moins un*.

Nous avons rencontré le cas suivant qui décrit un protocole de validation d'un test. Il peut servir un peu d'exemple.[3] On a eu à analyser des données de préférences pour lesquels les auteurs du test avaient dû construire des items adaptés à leurs besoins. Voici leur procédure très professionnelle :

Méthod-Q

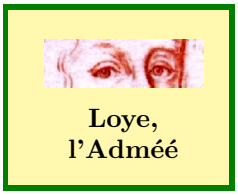
Heureusement qu'on a des moyens d'observation puissants sous la forme des analyses factorielles, qu'on peut utiliser à diverses fins dont celle de la sélection des items.

Remarque personnelle

Il règne encore dans mon esprit une certaine confusion sur le sens précis et *opérationnel* de ces termes (ainsi validation de construit et de contenu)... On entend aussi parler de *validité convergente, discriminante* etc. Les concepts en question ont aussi évolué au cours du temps (Loye, 2013).

3. Combien de fois on effectue des saisies de données pour valider des modèles complexes, basées sur des théories philosophiques profondes sans se préoccuper de savoir comment en tirer de l'information, et en désespoir de cause on appelle le 'pompier de service' qui n'y peut mais : ≪ Pensez donc un peu, une vingtaine d'observations — et le milieu n'en porte pas plus ! —, qu'est-ce qu'on peut faire avec ça ? Rien. ≫ Ce n'était pas le cas ici, la méthodologie-Q qu'on a mise en action se contente de peu de sujets. On me permettra toutefois de plaider pour une approche intégrée impliquant le spécialiste des données dès l'amorce du projet.

On ne s'attardera pas vraiment pour le moment sur ces termes ; on renvoie aux références, notamment la présentation de Nathalie Loye au Congrès de l'Adméé en 2013, entre autres pour l'histoire du concept de validation — voir le lien ci-dessous. Voir aussi Carmines & Zeller (1979) et De Vellis (2012).



Techniques d'observation

All models are wrong. Some are useful.
George E.P. Box[4] [1919 – 2013]

Heureusement qu'on a des moyens d'observation puissants sous la forme des analyses factorielles qu'on peut utiliser à diverses fins dont celle de sélection des items.

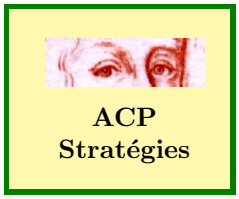
Mais aussi

d'exploration//validation//confirmation d'hypothèses//de modèles.

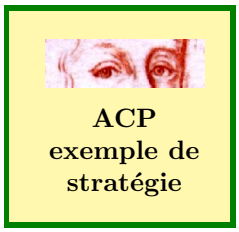
Qu'on expose ici sur un exemple...

4. Probablement le statisticien, après Fisher [1890 – 1962] dont il était le gendre, ayant eu le plus d'influence sur la science au XX^{e} siècle.

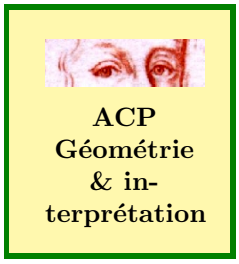
Stratégies en ACP :



Un exemple de stratégie : « rechercher un point de vue » ; les éléments actifs & illustratifs



La géométrie de l'ACP : la compréhension visuelle ; apport prépondérant de la régression et prédictions sur éléments illustratifs



Conclusions provisoires

1. Nous pouvons penser que *la seule validation essentielle* est la validation en référence à un critère : validités *prédictive* & *concomitante* (DeVellis, 2012) ; et pour cela ce sont les méthodes factorielles qui servent ; plus généralement les modèles aux équations structurelles (non décrits ici)
   - d'où l'importance de la construction de *schémas-blocs* ou *schémas structurels* (causaux) et l'utilisation de devis dits *corrélationnels* qui les explicitent
   - d'où la nécessité d'utiliser des stratégies utilisant des *variables illustratives*, i.e. des choix de points de vue pour confirmer les hypothèses de la recherche
2. Approche pragmatique : en ACP/AF ce sont les interprétations qui comptent ; il faut toujours pouvoir retourner aux items, i.e. à leurs libellés
   - d'où la nécessité d'utiliser les stratégies des *sujets illustratifs* et des parangons en ACP
   - pour confirmer des hypothèses sur les variables concomitantes, la validation ne peut se passer de modèles de régression sur composantes principales et des prédictions pour des sujets illustratifs
3. Dans la construction des tests (et même dans leur utilisation), il est essentiel de prévoir des tailles échantillonnales suffisantes pour effectuer des *validations croisées* des analyses, ou encore par des techniques de *rééchantillonnage* (non décrites ici)
4. La validation est un *processus continu...* à recommencer chaque fois que le test est utilisé ; ajouter une pierre à la construction de la validation.

Nous n'avons fait qu'effleurer ces éléments

Annexe

It ain't over 'til it's over.
Yogi Berra [1925–]

Nous avons conçu cet exposé à des fins pédagogiques. L'exemple numérique traité n'a concerné que le test sur les stratégies de Coping dont les 21 items se répartissent en trois dimensions, et nous nous sommes concentrés sur la dimension de la Recherche de soutien social qui comporte seulement 6 items.

L'avantage de la simplicité est évident. Mais la pédagogie s'est exercée au détriment de la véritable nature de la structure des données, des fins de l'enquête donc. En réalité c'est la Hardiesse qui pilote les attributs de la santé des patients interrogés, la seule variable dite *endogène* de nos données (voir le Schéma-bloc de l'expérience). Et c'est là que doit se centrer l'analyse.

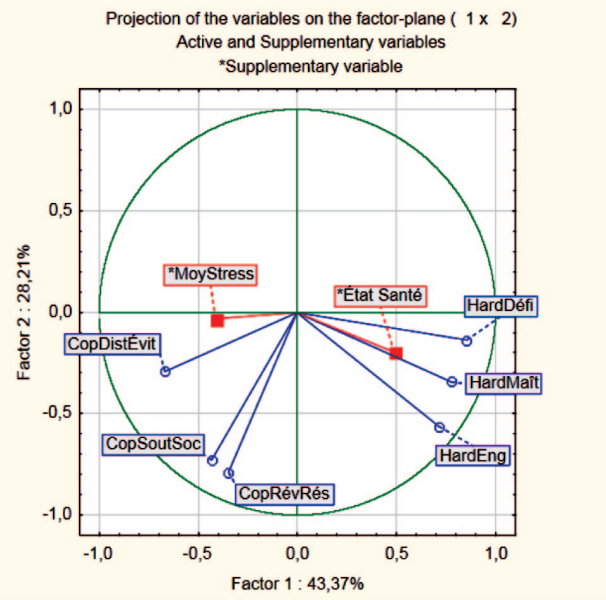


Figure 1 – Plan factoriel (1,2) dont les variable actives sont en bleu, les illustratives en rouge.

Tout de même, nous avons mis en évidence (cf. la Fig. 1, tirée d'un des fichiers d'analyse présenté) la corrélation négative entre la Hardiesse des sujets et leur stratégies de Coping (C), d'une part, et, d'autre part, positive entre la hardiesse

(H) les indicateurs de santé, qu'ils soient objectifs comme le taux de CD4 dans le sang des sujets, ou subjectifs comme leur auto-évaluation du stress (Stress), ou de leur état général de santé (Santé). Ce qui est parfaitement conforme aux attentes pour la validation concomitante, tel que nous l'avons exposé dans le texte. Les équations suivent, on peut les résumer par le schéma causal de la Fig. 2 :

$$\rho(\,\mathrm{H}\,,\,\mathrm{C}\,) < 0\;;\;\rho(\,\mathrm{H}\,,\,\text{Santé}\,) > 0\;;\;\rho(\,\mathrm{H}\,,\,\text{Stress}\,) < 0\;;$$
$$\rho(\,\mathrm{C}\,,\,\text{Santé}\,) < 0\;;\;\rho(\,\mathrm{C}\,,\,\text{Stress}\,) > 0\;;$$
$$\rho(\,\text{Santé}\,,\,\text{Stress}\,) < 0\;.$$

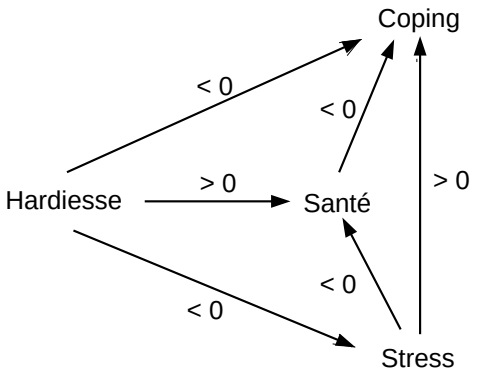


FIGURE 2 – Schéma causal des tests de nos données. La Hardiesse est la seule variable endogène. Les flèches se rapportent à des modèles linéaires simples, l'origine d'une flèche étant une variable indépendante, sa destination la dépendante. Le < 0 ou > 0 sur la flèche se rapporte au signe de la corrélation.

Mais le vrai travail de l'analyste de ces données ici est de traiter le plus profondément possible du *rôle moteur* de la hardiesse, seule variable endogène dans le schéma de la Fig. 2 dans la santé des individus, plus spécifiquement ici les sidéens pour lesquels on veut augmenter l'adhérence à leur traitement très contraignant de tri-thérapies, telle que mesurée objectivement par leur taux de cellules CD4 dans leur sang. [5]

On éludera pour le moment, la présence des autres flèches causales dans le modèle de la Fig. 2. Notons toutefois dans cette figure la présence de plusieurs schémas/variables de médiation (Baron & Kenny, 1986 ; Iacobucci, 2008).

5. On renvoie ici le lecteur à Asher (1976), un des premiers à vouloir transmettre les modèles quantitatifs de la causalité à un large public (de savants...), ainsi qu'à Morgan & Winship (2007). De même qu'à un bel ensemble de manuels récents centrés autour de la causalité parus à la *Cambridge University Press* (Freedman, 2009 ; Kleinberg, 2013), suite au texte fondamental de Judea Pearl paru en 2000 dont la deuxième édition date de 2009. Cette référence pour les plus aguerris. On trouvera ces dernières références dans la bibliographie générale dont l'hyper-lien est en début de ce texte.

Le principal objectif qui a généré cette recherche, plus que celui de vérifier si oui ou non la hardiesse est un facteur explicatif de cette adhérence — fonction descriptive & de validation prédictive du test —, est le suivant : comment pourrait-on, dans le cas positif, améliorer la hardiesse des patients et ainsi favoriser l'adhérence à leur traitement — la fonction, disons (?), prédictive avancée, peut-être (?) confirmatoire. En d'autres termes :

Peut-on définir sur la base de la hardiesse
une intervention efficace auprès des patients fragiles ?

La véritable mise à l'épreuve de ce travail est évidemment de soumettre une partie des patients interrogés, disons une moitié randomisée, à une telle intervention, de constater son effet sur les mesures des tests, bien sûr sur les critères subjectifs de l'amélioration de l'adhérence (ce qui est peu fiable), et surtout sur le critère objectif de l'augmentation du taux de CD4 dans leur sang.

Comme retombée non négligeable, ce test-retest avec groupe de contrôle constituerait, pour le psychométricien, une validation supplémentaire des tests utilisés. Peut-être la meilleure qui soit ?

Le lecteur qui a suivi cet exposé en trouvera la suite et conclusion logique en cliquant l'icone ci-dessous [a]

a. Ces pages Excel demandent encore bien des développements, elles ne sont pas très explicites, mais le déroulement en est rigoureusement logique. Notons toutefois qu'elles se lisent de gauche à droite et de haut en bas, comme un texte de livre, et seront explicitées dans un travail qui suivra.

Prédire

Re-concluons, très provisoirement...

La naissance de ce petit ouvrage est due entièrement au hasard,
et plutôt à une espèce de divertissement
qu'à un dessein sérieux.

La logique ou l'art de penser,
Antoine Arnauld & Pierre Nicole. [6]

La validation en psychométrie a fait et fait encore l'object de nombreux questionnements, le sujet semble encore en plein développement (e.g. parmi tant d'autres travaux : Kane, 2006 ; Borsboom, 2006 ; Borsboom & Markus, 2013). [7] Les épistémologues se préoccupent toujours de ce qu'on peut *vraiment* connaître. Le sujet en psychométrie semble donc particulièrement récalcitrant et pour cause. Ce qu'on veut mesurer ne serait-il pas inconnaissable (Fig. 3) ?

Les construits tels ceux qu'on aimerait mesurer ici, la hardiesse, le *coping*, ne sont pas définis autrement que par des mots, compte tenu de toutes les importantes théories mises en action pour y arriver... Les ≪ choses ≫ ? On n'en voit que des ombres, comme dans la caverne de Platon. L'essence des choses est inaccessible. La mesure est une ombre.

Il n'y a pas d'étalons de mesure grâce auxquels on peut évaluer, i.e. valider les instruments de mesure, tel qu'on en trouve en sciences naturelles. [8] La validation discriminante pose de même façon des problèmes insurmontables.

On se trouve alors en plein dans la citation de Binet plus haut, sorte de pétition de principe : ce que je veux mesurer, mon test le mesure. Mais la question nul doute est beaucoup plus complexe épistologiquement parlant.

Ne resterait peut-être que l'argument pragmatique pour justifier l'utilisation de tests psychométriques du genre de celui qu'on a traité longuement ici : que veut-on faire avec les mesures ? Si on y arrive l'instrument serait validé. Du moins dans ce sens fort limité. Et surtout sans vouloir inférer autre chose, notamment sur la nature réelle des choses mesurées.

6. Il n'y a pas de hasard, dit-on parfois. L'auteur de ces pages est tombé opportunément et pas tout à fait ...par hasard sur la série *Port-Royal* de l'émission Les nouveaux chemins de la connaissance de France-culture où sont citées ces lignes tirées de l'Avis liminaire de cet important ouvrage du XVI^{e} siècle qu'il est convenu d'appeler *Logique de Port-Royal*. On comprendra que l'auteur ait une certaines fascination pour Blaise Pascal sans du tout en partager la pensée janséniste, ce qui n'est pas un hasard. Témoin ses écrits sur ce fondateur de la probabilité. Le lecteur pourra trouver quelque intérêt au dernier ≪ Les hexagrammes mystiques de Blaise Pascal ≫ qui n'a rien à voir avec la probabilité... Le hasard bien sûr, encore et toujours, l'aurait amené sur ces chemins de traverse.

7. On pardonnera j'espère ce petit essai qu'on mettra sur le compte de l'ignorance de l'auteur, ignorance qui n'est plus à footnoter ! On cherchera à l'éclairer, les suggestions et commentaires sont bienvenus.

8. Et encore... Rappelons la citation de E. Deming en exergue du texte : *There is no true value of anything.*

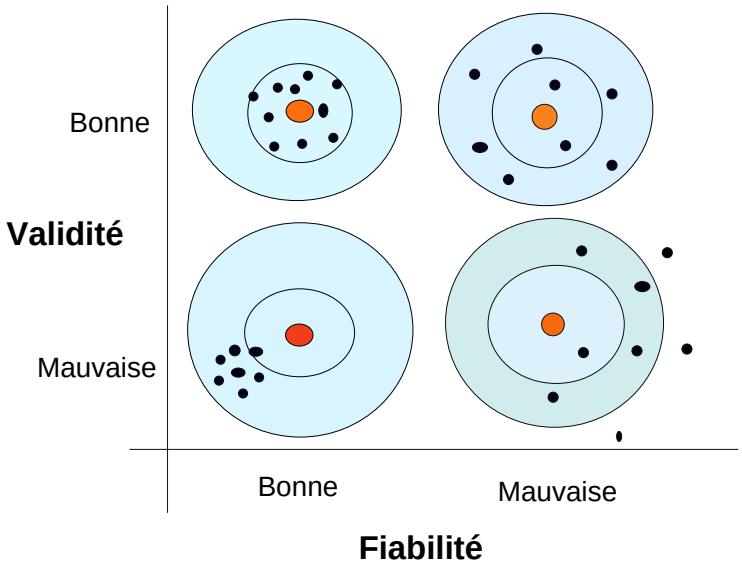


FIGURE 3 – Fiabilité & validité. L'analogie de la cible avec le construit au centre. Le problème est qu'il n'y a pas de centre à la cible en psychométrie. En fait, la cible elle-même n'est pas définie...

C'est la validation prédictive qui serait ainsi, telle que la chose nous apparaît après ce long parcours, le seul critère. Et cette validation ne pourrait se faire qu'en présence d'une 'expérimentation' statistique telle qu'on l'a préconisé plus haut. Mais là aussi, la tâche est délicate. Et peut-être insurmontable...

Revenons aux technologies statistiques.[9] On ne peut que regretter que les chercheurs en sciences de la santé qui ont défini la recherche dont on a fait l'analyse brièvement ici, n'aient pas du tout pensé en termes expérimentaux (enfin presque, mais plutôt comme le bourgeois gentilhomme fait de la prose sans trop le savoir) : ils ont construit une intervention auprès des sidéens pour augmenter leur hardiesse, intervention générale importée d'ailleurs, sans exploiter ce qui aurait pu l'être suite à une analyse des données déjà recueillies, sous l'hypothèse non vérifiée alors que la hardiesse jouait un rôle moteur dans l'adhérence des patients à leur tri-thérapie.[10]

On l'a supposé, mesuré les caractéristiques des sujets, on est intervenu sur une moitié randomisée des sujets, on les a tous remesurés. Et cela a marché...

9. Le lecteur intéressé pourra lire avec profit le charmant texte suivant : David Salsburg (2002), *The lady tasting tea. How Statistics revolutionized Science in the twentieth century.* New York NY : Henry Holt & Co.

10. D'ailleurs l'a-t-on vérifiée, cette hypothèse ? D'autres schémas causaux (cf. Fig. 2) sont possibles, d'autres hypothèses de motivation.

On a fait une thèse.[11] Fort complexe, et fort bien notée.[12]

Mais dans d'autres études, on a montré qu'agissait avant tout un effet placebo. Et il n'y a pas eu de thèse, puisque, intervention ou pas, l'effet était tout aussi marqué : montrer aux sujets qu'on s'intéressait à eux suffisait à induire un effet.[13]

La validation que nous suggérons ici n'a même pas été évoquée à l'époque. Et pour cause, les analyses rapportées ici n'avaient pas encore été conçues !

Une expérience planifiée (Spector, 1990 ; Brown & Melamed, 1981), eût-elle été menée, aurait pu valider la question de recherche dûment formulée à la suite de notre analyse[14] : « Peut-on définir un intervention *ciblée* sur un certain nombre bien limité d'items du test de la hardiesse, qui aurait amélioré l'adhérence des sidéens à leur traitement de tri-thérapie. »

Valider éventuellement cette réponse, soit. Mais valider le test de hardiesse utilisé, pas du tout. Le concept, comme tous ces construits, serait resté évanescent, tout autant que maintenant.

11. Si l'hypothèse hardie n'avait pas été avérée, on aurait quand même fait une thèse.
12. À juste titre, grâce à un appareil quantitatif bétonné, inédit à Montréal à l'époque.
13. Dans ce cas, on n'avait que très peu de sujets, ce qui invalidait même toute possibilité d'analyse statistique compétente !
14. On n'ose pas trop référer ici au manuel classique de Box, Hunter & Hunter (2006), vu la technicité du texte, bien qu'élémentaire car destiné aux expérimentateurs. La genèse de la première édition de ce texte fondamental est décrite dans l'autobiographie de George E. P. Box (2013). On a aussi l'indispensable Kutner *& al.* (5e édition, 2005) qui regorge de sagesse statisticienne.

Ramassons-nous...

1. La construction sur mesure d'un test psychométrique sur un construit n'est pas donnée. Mieux vaut utiliser du prêt-à-porter. Dans les cas où cela est absolument nécessaire
   (a) il est indispensable de travailler en équipe, de faire appel à des groupes témoins (*panels*), d'itérer plusieurs fois un protocole précis
   (b) la première sélection, très préliminaire, des items doit faire appel à un réseau conceptuel conséquent : philosophie, psychologie, sociologie, etc. Elle doit comporter un grand nombre d'items. C'est là que se pratique la validation de concept (contenu & construit) peu intensive en techniques quantitatives
   (c) la formulation des items (syntaxe, choix des mots, sans mode passif et banissant l'usage du négatif) doit tenir compte du contexte de l'utilisation prévue
   (d) une construction//validation d'un test dépend fortement du contexte de son utilisation ; on a évoqué plusieurs techniques de sélection finale des items parmi les dizaines suggérées dans la littérature, processus itératif aussi
   (e) traduire un test demande plusieurs traducteurs indépendants et une rétro-traduction de validation
2. Une validation compétente de fidélité & validité d'un test (utilisation ou construction) commence par établir un réseau de concepts voisins associés à un schéma causal, utilisant des tests déjà dûment validés, qui servira à une validation par critère (prédictive, concomitante...), la seule qui semble avoir un quelconque intérêt, une fois la validation de concept provisoirement établie (protocole du point 1)
3. Dans l'étape de la validation princeps on doit utiliser des allers-retours entre les étapes 1 & 2.
4. Les technologies statistiques pertinentes à cet égard sont d'un emploi essentiel, elles sont intensives en connaissances & jugements quantitatifs délicats
   (a) au moins dans la validation princeps, il faut prévoir des échantillons de validation (et même d'utilisation) assez grands pour permettre des validations croisées & l'utilisation de techniques de rééchantillonnage ;
   (b) pour les utilisations courantes, toujours avoir à l'esprit les techniques de planification d'expériences statistiques
5. Les principales technologies statistiques impliquées pour la construction//validation des tests :
   (a) la modélisation linéaire : simple, multiple, généralisée, *anova, ancova, manova, mancova*, à mesures répétées.
   (b) les analyses factorielles : AF, ACP, AC, ACM, AFM
   (c) les modèles linéaires structurels : demandent cependant des tailles échantillonnales importantes
   (d) les technologies précédentes appliquées en cascade

Références

Asher, H. B. (1976). *Causal modeling.* QASS n° 3. Sage Publications, Beverly Hills CA.

Baron, R. M. et Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research : conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6): 1173 – 1182.

Blais, J.-G. et Raîche, G., éditeurs (2003). *Regards sur la modélisation de la mesure en éducation et en sciences sociales*, Québec Q^{c}. Presses de l'Université Laval.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, 71(3): 425 – 440.

Borsboom, D. et Markus, K. A. (2013). Truth and evidence in validity theory. *Journal of Educational Measurement*, 50(1): 110 – 114.

Bourdeau, M., Delmas, P. et Sylvain, H. (2012). Using partial components to restore and use the concurrent validity of the Index of readiness. *Tutorials in Quantitative Methods for Psychology*, 8(2): 70 – 87.

Box, G. E. P. (2013). *An accidental statistician. The life and memories of George E. P. Box.* John Wiley & Sons, New York NY.

Box, G. E. P., Hunter, J. S. et Hunter, W. G. (2006). *Statistics for experimenters : design, innovation, and discovery.* Wiley-Interscience, Hoboken NJ, 2^{e} édition. La première édition, 1976, a marqué la statistique appliquée.

Brown, S. R. et Melamed, L. E. (1990). *Experimental design and analysis.* QASS n° 74. Sage Publications, Newbury Park CA.

Carmines, E. G. et Zeller, R. A. (1979). *Reliability and validity assessment.* QASS n° 17. Sage Publications, Beverly Hills CA.

DeVellis, R. F. (2012). *Scale development. Theory and applications.* Applied social research methods series, n° 26. Sage, Thousand Oaks CA, 3^{e} édition.

Dunteman, G. H. (1989). *Principal components analysis.* QASS n° 69. Sage Publications, Beverly Hills CA.

Fishbein, M. et Ajzen, I. (1975). *Belief, attitude, intention and behavior : An introduction to theory and research.* Addison-Wesley Publishing Company, Reading, MA.

Freedman, D. A. (2009). *Statistical models. Theory and practice.* Cambridge University Press, New York NY, édition révisée édition.

Iacobbuci, D. (2008). *Mediation analysis.* QASS n° 156. Sage Publications, Thousand Oaks CA.

Kane, M. T. (2006). Validation. *In* Brennen, R. L., éditeur : *Educational measurement.* Praeger, Westbury CN, 4^{e} édition.

Kim, J.-O. et Mueller, C. W. (1978). *Introduction to factor analysis. What it is and how to do it.* QASS n° 13. Sage Publications, Beverly Hills CA.

Kleinberg, S. (2013). *Causality, probability, and time.* Cambridge University Press, New York NY.

Morgan, S. L. et Winship, C. (2007). *Counterfactuals and causal inference. Methods and principles for social research*. Cambridge University Press, New York NY.

Neter, J., Kutner, M. H., Nachtsheim, C. J. et Wasserman, W. (1996). *Applied linear statistical models*. Irwin, Boston MA, 4e édition.

Pearl, J. (2009). *Causality. Models, reasoning and inference*. Cambridge University Press, New York NY, 2e édition.

Revelle, W. et Zinbarg, R. E. (2009). Coefficient alpha, beta, omega and the GLB : comments on Sijtsma. *Psychometrika*, 74(1) : 145 – 154.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, 8(4) : 350 – 353.

Scott Long, J. (1983). *Confirmatory factor analysis. A preface to Lisrel*. QASS no 33. Sage Publications, Beverly Hills CA.

Sijtsma, K. (2009). Reliability beyond theory into practice. *Psychometrika*, 74(1) : 169 – 173.

Spector, P. E. (1981). *Research designs*. QASS no 23. Sage Publications, Beverly Hills CA.

Summers, G. F., éditeur (1970). *Attitude Measurement*. Rand McNally & Co., Chicago IL. Tous les fondateurs de la psychométrie y sont.