Cours de psychométrie 2

Introduction

- REUCHLIN 1974
- Un test est une technique permettant une **description quantitative et contrôlable** du comportement d'un individu placé dans une situation définie par **référence aux comportements des individus d'un groupe défini, placés dans la même situation.**
- Le test est donc une épreuve impliquant une tâche à remplir, identique pour tous les sujets, avec une technique précise pour l'appréciation du succès ou de l'échec. C'est aussi un instrument standardisé qui permet de situer un individu dans un groupe biologiquement et socialement déterminé

Les différentes étapes de construction d'un test :

1) Echantillonnage
2) Choix et sélection des items
3) Etude de l'unidimensionnalité et de la sensibilité de l'épreuve
4) Etude de la fidélité
5) Etude de la validité
6) Etalonnage

1-L’ECHANTILLONAGE

L'échantillonnage est l'opération de sélection d'un échantillon. De façon formelle on peut dire "**Échantillonner c'est prendre correctement la partie d'un tout pour que l'on puisse faire une estimation sur ce tout à partir de cette partie** "

- Le test est construit pour **différencier les individus** d’une **population** donnée, et sa mise au point commence nécessairement par la définition d’une **population de référence.**
- Lors de la construction d'un test on va extraire un ou plusieurs échantillons représentatifs (échantillonnage) afin de mettre au point le test, étudier ses qualités métrologiques puis l'étalonner.

-Population :

Ensemble d'individus sur lesquels porte l'étude statistique .

Chaque individu est appelé « unité statistiques »

Une fois la pop définie, on utilise un échantillon représentatif.

-Echantillonnage :

Opération de sélection d'un groupe d'individus présentant les caractéristiques de la population parente à laquelle le test est destiné.

- Echantillon :

Un sous ensemble visé par l'étude statistique

- Les trois règles de l'échantillonnage :

Equiprobabilité : tous les individus ont la même probabilité d'apparaitre dans l'échantillon

Indépendance : la présence d'un individu ne modifie pas les chances des autres individus de la population d'apparaitre dans l'échantillon

Normalité : pas d'obligation de normalité dans la population

a)L’échantillonnage probabiliste

Les méthodes probabilistes sont des méthodes d'échantillonnage dans lesquelles chaque individu de la population est tiré au sort et à donc la même probabilité de faire partie de l'échantillon. Ces méthodes (à l'exception de l'échantillonnage en grappe) nécessitent toujours une liste exhaustive de la population parente. Cette contrainte rend souvent ces échantillonnages probabilistes difficiles à réaliser.

Caractéristiques principales :

- Basés sur des lois de probabilité
- Chaque élément de la pop a des chances d’être choisi
- Habituellement le plus représentatif

Les différents types :

- Aléatoire simple
- Systématique
- Stratifié
- En grappe
- A plusieurs phases

1.ALEATOIR SIMPLE

Choix des individus dans une pop

Chaque membres a les même chances d'être choisi

Avantages :

-Simplicité

Inconvénients :

-Difficultés à établir une liste exhaustive de la population

-Coûteux

2.ECHANTILLONNAGE SYSTEMATIQUE

Méthode qui exige une liste de la population où chaque individu est numéroté de I jusqu'à N

- **Calcul du rapport :**

r = N／ taille de l'échantillon

-Si on veut un échantillon de 200 personnes sur 10000 :

10000 /200

- On choisi un nombre entre 1 et r (ici 50)
- Si on choisi d=4 alors on choisi la 54e sujet et le 104e et ainsi de suite jusqu'au 200

Avantages :

-Répartir l'échantillon sur toute la population

-Facilite la sélection car seul le premier sujet est aléatoire

Inconvénients :

-Echantillon pas toujours représentatif de la population

-Avoir la liste des différents éléments de la pop

3.ECHANTILLONNAGE STRATIFIE

Pour éviter les biais d'échantillonnage et s'il est possible de savoir comment se distribuent certains caractères de la population on peut stratifier l'échantillon selon ses critères Il s'agit de reproduire au niveau de l'échantillon à une plus petite échelle, les caractéristiques distinctives de la pop de référence.

Ex 52% femme et 48% de femme

Avantages :

-Simple à utiliser

-Représentativité facile à observer

-Caractéristiques maitrisées = conclusions plus fiables

Inconvénients :

-Avoir la liste des différents éléments de la pop

-Connaitre la répartition de la pop dans les différentes caractéristique

-Il faut un grand échantillon pour respecter les proportions dans les caractéristiques

4.ECHANTILLONNAGE EN GRAPPES

La technique entraine la division de la population en groupe ou grappes alors qu'avant l'unité statistique était choisie individuellement.

On sélectionne au hasards des grappes pour représenter la pop puis on sélectionne les individus des grappes choisies Il vaut mieux choisir un grand nombre de petites grappes que l'inverse

Avantages :

-Simple à utiliser car grappes et pas individuel

-Pas besoin de liste globale de la pop

Inconvénients :

-Risque d'imprécision des résultats (les unités voisines ou grappes ont tendance à se ressembler)

-On ne connait pas à l'avance la taille de l'échantillon (grappe et pas unité statistique)

5-ECHANTILLONNAGE A PLUSIEURS PHASES

Les données de base sont collectées auprès d'un échantillon d'unité de grande taille, ensuite pour un sous-échantillon de ces unités Les plus souvent on utilise deux phases

Ex : on interroge toute exploitations agricoles (phase 1) puis seuls les éleveurs de volailles (phase 2

Avantages :

-Rejet de certaines données

-Rejet d'une partie de l'échantillon

Inconvénients :

-Réduction de la taille de l'échantillon

b)L’échantillonnage NON probabiliste

Ces échantillons sont construits à partir des caractéristiques d’une pop risque d’erreur

Les différents types :

- A l’aveuglette
- Des volontaires.

1- A L’AVEUGLETTE

Eléments choisis au fur et à mesure qu’il se présentent Simple, rapide, peu coûteux

Ex : prendre les 10 premières personnes qui sortent d’un immeuble

2. DE VOLONTAIRES

Eléments choisis sur une base volontaire Meilleure représentativité si on sélectionne dans les volontaires

Ex : recrutement par affichage ou annonce sur le net

3-PAR QUOTAS

Il s'agit de l'une des formes les plus courantes d'échantillonnage non probabiliste. L'échantillonnage est effectué jusqu'à ce qu'un nombre déterminé d'unités (quotas) pour diverses sous-populations soient sélectionnées.

L'échantillonnage par quotas est un moyen de satisfaire les objectifs de taille d'échantillon pour les sous-populations.

Les quotas peuvent être basés sur les proportions de la population. Par exemple, si la population compte 100 hommes et 100 femmes, et il faut tirer un échantillon de 20 personnes, 10 hommes et 10 femmes peuvent être interviewés.

4-PAR RESEAU

éléments choisis à travers des réseaux sociaux ou autres types de réseaux

2) Choix et sélection des items

Le choix et sélection des items en réalité sont la quatrième étape dans la démarche de la construction des tests, mais avant d'approfondir dans cette opération, on doit exposer en détaille les principales étapes de la démarche de la construction d'un test :

1- Définir précisément les objectifs : outil de recherche ou d'application (diagnostique ou pronostique) ? La plupart des tests ont été construits pour valider des théories puis ont été remaniés en outils d'application.

Ex: Inventaire de Dépression de Beck ou Beck Depression Inventory (IDB)

- L'objectif de cette échelle est d'évaluer le degré de la dépression, Il met en avant des **pensées automatiques** et des **dialogues internes** chez le dépressif qui amènent et amplifient le phénomène. Il montre que travailler sur ces pensées avec les patients peut les amener à les remettre en cause et à guérir. **C'est une perspective cognitive**

2- Définir précisément le domaine à mesure : on doit spécifier ce qu'est le trait à étudier en précisant (le plus exhaustivement possible) et hiérarchisant les différentes caractéristiques de ce domaine, ce qui amène à échantillonner les items (créer un groupe de question représentatif du domaine à mesurer).

Ex: Sous leur perspective cognitive, les troubles dépressifs se caractérisent par une vision négative handicapante de: l'estime de soi , des expériences faites dans la vie (et du monde en général), et de l'avenir. Ces trois points définissent la triade cognitive.

- Les patients dépressifs se perçoivent eux-mêmes souvent comme désespérés, inefficaces, mal-aimés, et tendent à se mésestimer à cause de problèmes physiques, mentaux et/ou moraux. Ils culpabilisent excessivement, et se croient inutiles et incapables d'être aimés par autrui. Ils dévaluent systématiquement leurs expériences actuelles et antérieures. Ils se perçoivent difficilement comme des individus acceptés, ou même ressentir un bien-être émotionnel. Ils perçoivent leur vie comme dénuée de sens et de plaisir, et ne pensent pas réussir à surmonter les obstacles de buts qu'ils se sont fixés. Pour eux, tout leur semble « trop dur à supporter ». Ils pensent également que leurs problèmes continueront indéfiniment et que leur avenir ne sera constitué de frustration et de privation. Les tentatives de suicide sont une expression extrême du désir d'échapper aux problèmes qui leur paraissent incontrôlables, interminables et insupportables

3-Définir précisément la population à laquelle s'adresse le test pour faire un échantillonnage des sujets.

4- Construction des items : choix de la forme et du contenu des items ; vérification de la validité de contenu (avec des experts du domaine) concernant toutes les caractéristiques du domaine étudié et dans les même proportions (ce qui est parfois difficile voire impossible pour des domaines non-explicitables).

EX: l'échelle de dépression

- Cognition sur soi :

  « Je ne vaux rien, je ne suis pas à la hauteur » (estime de soi faible).

- Cognitions sur l'environnement :

  « Ce monde est pourri, les gens sont égoïstes » (monde injuste).

- Cognitions sur l'avenir :

  « Rien ne s'améliorera jamais, c'est sans espoir » (avenir sans espoir)

5-Standardisation : consignes, matériel, notation (pour la fidélité inter-correcteurs).

EX : consignes échelle de beck

Ce questionnaire comporte 13 séries de quatre propositions. Pour chaque série, lisez les quatre propositions, puis choisissez celle qui décrit le mieux votre état actuel de ces deux dernières semaines. Entourez le numéro qui correspond à la proposition choisie.

EX : cotation et interprétation

Analyse des scores :

- 0–9 : indique une dépression mineure ;
- 10–18 : indique une légère dépression ;
- 19–29 : indique une dépression modérée ;
- 30–63 : indique une dépression sévère.

6- Pré-test évaluation des items et de la standardisation.

7- Construction d'échelle de validité interne en particulier pour les mesures conatives.

8- Étude des qualités psychométriques : analyse des items et sensibilité de l'épreuve ; fidélité dans le temps et homogénéité ; validité (empirique pour estimer l'efficacité diagnostique ou pronostique du test ; théorique pour voir si les données sont conformes au construit de départ).

9- Travail de reconstruction et d'amélioration à chaque étape de l'étude des qualités psychométriques et éventuelle révision de la théorie de départ.

10- Construction de normes et d'étalonnages pour donner une interprétation : données recueillies sur de nouveaux échantillons.

11- Contre-validation : vérification des données sur de nouveaux échantillons.

12- Publication éventuelle : rédaction précise d'un manuel selon les normes internationales, protection du test et vente.

13- Révision du test en fonction des changements culturels, linguistiques etc

Elaboration des items d'un test

- Un test est constitué d'un ensemble d'items (questions simples ou complexes) pouvant avoir différents formats. Ces items sont construits (inventés) par le psychologue (le plus souvent une équipe de chercheurs et psychologues) et peuvent être totalement nouveaux ou s'inspirer de tests récents ou anciens. Ils font toujours (nécessairement) référence à un corpus de connaissances existant à un moment donné.

La nature et le format des questions sont très vastes et dépendent aussi bien de **"l'objet mesuré",** que de la cible (personnes à interroger), **des modalités de passation souhaitées** (collectif, individuel, informatisé) ou encore de **contraintes temporelles**. Lors de l'élaboration de ces questions **on doit fixer des règles générales de passation** comme **l'ordre des items ou encore le nombre des questions**. Pour fixer l'ordre des items par exemple on peut choisir un ordre de **difficulté croissant ou aléatoire**, on peut s'autoriser ou non le **mélange** de questions appartenant à des sous-dimensions différentes.

- Dans ce qui suit, nous présentons quelques éléments importants soulignant la complexité de la construction et de la sélection des items d'un test.

1. **Format des questions:**

   Habituellement, on distingue différents formats. Tous présentent des avantages et des inconvénients. Les règles de constructions des items peuvent cependant être différentes selon que l'on élabore un test cognitif ou un questionnaire de personnalité. Habituellement on distingue cependant les formats suivants :

- **Le type traditionnel** (questions totalement ouvertes) : ce format est plus rarement utilisé car il pose des problèmes de standardisation de la cotation.
- **Formes à corrections objectives** qui font appel à la mémoire ou un traitement particulier (ex. : « 8 représente quel pourcentage de 64 »), un jugement, une évaluation, etc. Parmi les formes à corrections objectives on distingue :

  Le type "ouvert" : il existe une réponse juste (avec variante). Ces items peuvent être plus difficiles à coter pour un débutant. Dans les échelles de Wechsler, ce type correspond aux sous-tests "**vocabulaire" ou "similitude**" par exemple. Des questions faisant intervenir peu le langage comme des puzzles sont classées dans cette catégorie (elles sont parfois appelées questions de performances mais ce sont des questions ouvertes avec une réponse juste).

- **Les questions "VRAI - FAUX"** avec une réponse parmi deux. On peut distinguer deux mode d'utilisation :

· **"VRAI-FAUX" ou "OUI-NON"** utilisé dans les épreuves cognitives. On peut ne pas répondre (ce qui les distingue des items dichotomiques ci-dessous).

Les items dichotomiques (équivalent d'un "VRAI-FAUX") mais utilisés dans les questionnaires, ils obligent le sujet à exprimer un avis. Par exemple :

Je suis anxieux à l'approche des examens ▢OUI ▢ NON

- **Les questions à choix multiples (QCM*) : proches des VRAI-FAUX,** ce sont des questions avec une ou plusieurs bonnes réponses parmi un ensemble de propositions alternatives accompagnant, une question, une affirmation .. Dans ces épreuves les fausses réponses sont appelées « les distracteurs » ou les "**leurres**".
- **Les questions d'appariement :** ce sont des questions proches du format à choix multiples dans lesquelles on demande de mettre en relation (apparier) des énoncés (phrases, mots, expressions) qui sont souvent présentés sur 2 colonnes. Exemple :

BINET	QI
STERN	TEST MENTAL
CATTELL	AGE MENTAL
SPEARMAN	FACTEUR g

Les échelles de Likert (du nom du psychologue Rensis Likert) sont très souvent utilisées dans les questionnaires (personnalité, opinions, valeurs, etc.). Dans ces échelles la personne interrogée doit exprimer son degré d'accord ou de désaccord vis-à-vis d'une affirmation. La réponse est exprimée sous la forme d'une échelle qui permet de nuancer son degré d'accord. Par exemple :

- *Je suis souvent en colère :*
  - ☐ *Pas du tout d'accord*
  - ☐ *Pas d'accord*
  - ☐ *D'accord*
  - ☐ *Tout à fait d'accord*

- **Analyse et sélection des items:**

  Un test est constitué d'un ensemble d'items (questions) devant différencier les sujets le plus correctement et le plus efficacement possible. Lorsque l'on construit une épreuve, la subjectivité et/ou les connaissances antérieures des chercheurs ou psychologues à l'origine de l'épreuve jouent un rôle important.

- Plusieurs étapes sont nécessaires pour la construction d'une épreuve. Un premier ensemble d'items constitue une version provisoire du test qui est administrée à un échantillon de personne. Cet ensemble d'items est remanié en fonction des premiers résultats observés (analyse des items) ou de l'avis d'experts du domaine (lorsqu'il s'agit de questionnaires). On ne se contentera pas de supprimer des items, le plus souvent il peut être nécessaire soit d'en revoir certains soit même, d'en construire de nouveaux. **Par exemple**, pour les items à choix multiples une analyse des distracteurs (réponses fausses proposées) peut conduire à modifier un ou plusieurs distracteurs.

- Lors de la sélection des items de nombreux critères rentrent en ligne de compte (longueur de l'épreuve, homogénéité de l'épreuve, difficulté souhaitée de l'épreuve, etc.). Pour les épreuves d'évaluation de « performances » cognitives, on prend en compte la difficulté (**comme le p-index**) des items et leur discriminabilité (**comme le d-index**).

a. Indice de puissance (p-index)

- L'indice de difficulté ou p-index (power en anglais) est aussi appelé en français indice de puissance. Ce p-index (*p) est simplement le rapport entre le nombre de personnes qui réussissent l'item et le* nombre de personnes qui l'ont passé (*p x 100 donne donc directement le pourcentage de réussite à* l'item). Cet indice de difficulté varie entre 0 et 1 (0 signifiant qu'un item est systématiquement échoué [0% de réussite] et à l'inverse 1 [100% de réussite] signifie qu'il est systématiquement réussi).

Exemple:

N	item1	item2	item3	item4	item5	item6	item7
1	3	1	2	1	2	5	1
2	4	1	5	1	3	5	2
3	1	2	5	2	4	3	3
4	2	3	5	1	3	5	3
5	3	1	4	1	1	1	2
∑	13	8	21	6	13	19	11
	p-index 5/13=0.38	**p-index 5/8=0.62**	**p-index 5/21=0.23**	**p-index 5/6=0.83**	**p-index 5/13=0.38**	**p-index 5/19=0.26**	**p-index 5/11=0.45**

- Observation:

L’item 4 a un p- index 0.83 le score le plus élevé,

L’item 3, a un p-index 0.23 le score le plus faible.

- Si l'on ne prend que des items d'indice **p élevé** (items difficiles) l'épreuve trop difficile ne discriminera que les très bons (les autres échoueront à tous les items). A l'inverse si l'on ne prend que des items difficiles (**à l'indice p trop faible)**, l'épreuve trop facile l'épreuve ne permettra de différencier que les sujets les plus en difficulté (les autres réussiront). Sachant que l'objectif est de maximiser la sensibilité de l'épreuve lors de la sélection des items, on choisit une majorité d'items dont le p-index est proche de .50 et on en prend de moins en moins au fur et à mesure que l'on s'éloigne de cette valeur vers **1 ou 0**. On fait l'hypothèse que la majorité des personnes se trouvent dans une **zone centrale** (on maximise à ce niveau la sensibilité du test) et on a besoin de moins d'items lorsque l'on s'éloigne de cette moyenne, car les personnes seraient moins nombreuses. On peut aussi, en manipulant cet index, construire des tests plus sensibles pour les personnes ayant des scores élevés ou inversement plus sensibles pour les personnes ayant des difficultés.

b. Indices de discrimination (d-index)

- Un bon item est un item qui doit distinguer les sujets en fonction de leur position sur la dimension évaluée. Un item de difficulté moyenne doit, par exemple, être réussi par toutes les personnes dont les compétences sont supérieures à ce niveau moyen et être échoué par les personnes dont les compétences sont inférieures à ce niveau moyen.
- Dans l'exemple précédant: Le (d-index) pour tout les items, est égale le (**p-index élevé**) moins le (**p-index faible**).
- d-index = 0,83-0,26
- **d-index =0,57**

3- Sensibilité et mesure d'une dimension:

- Lorsque l'on cherche à évaluer une dimension (exemples : une aptitude, un trait de personnalité). le test doit permettre de différencier le plus possible les personnes. **La sensibilité est alors le pouvoir séparateur, différenciateur d'un test. La sensibilité est donc la capacité d'un test à détecter une** variation du score vrai sur le trait mesuré (dans la théorie classique des tests). La méthode de sélection des items permet normalement de s'assurer de la sensibilité des tests .
- Pour étudier **la sensibilité** d'un test, une première méthode consiste à établir la distribution des résultats et d'examiner sa forme via le calcul d'indices de dispersion (écart-type ou autre), d'asymétrie ou d'aplatissement. Si l'épreuve est trop facile ou trop difficile, on observe une distribution asymétrique (effet plancher = trop difficile ou effet plafond = trop facile). On préfère en général une distribution plutôt normale, symétrique, au mieux légèrement aplatie qui présente une dispersion et un pourvoir différenciateur plus important.

- Si la distribution n'est pas une distribution normale, la sélection des questions étaient probablement incorrecte et le choix des questions doit être revue et/ou les questions remaniées. Lorsque l'o sélectionne les items on cherche à rendre la courbe « **plus normale** » d'une part et, d'autre part, à maximiser la dispersion de l'épreuve. Ce remaniement de l'épreuve s'effectue souvent en augmentant le nombre d'items de difficulté moyenne.

4-Etude de la fidélité:

- Une question importante lors de l'élaboration d'un test mesurant une dimension est de se demander si les différences observées entre les personnes correspondent à des différences réelles ou si ces différences observées sont fortuites (dues au hasard, entachées d'erreur et donc non répétables). C'est ce que l'on étudie avec la fidélité (un test fidèle est un test avec une erreur de mesure faible). La fidélité est donc un indicateur de la précision et de l constance des scores. Plus un instrument est fidèle, plus le score observé sera proche du score vrai.

4.1-Méthodes pour évaluer la fidélité:

a) test-retest:

- La méthode du **test-retes**t consiste à faire passer deux fois l'épreuve aux mêmes personnes avec un intervalle de temps souvent fixé aux alentours de 1 à 3 mois et de calculer la corrélation entre les performances observées lors de la première puis de la seconde passation. Le coefficient de fidélité est parfois appelé dans ce cas **« coefficient de constance » ou de « stabilité ».**
- *Inconvénient de cette méthode* :

  Il est difficile de fixer le temps optimal entre deux passations. Si le délai est trop long, la personnalité des individus, le niveau de compétence, etc. peuvent avoir changé, l'individu étant susceptible d'évolution. Si le délai est trop court, les résultats peuvent être faussés par un phénomène d'apprentissage ou de mémorisation.

b) les formes parallèles:

La méthode des tests parallèles permet d'éviter les inconvénients de la méthode du test-retest. Le principe consiste à construire deux versions semblables d'un test, deux formes équivalentes, dont seul le détail des items varie. Les deux versions sont alors passées le même jour ou avec un délai très court entre les deux passations. Ce coefficient de fidélité est appelé aussi le **coefficient d'équivalence et la méthode, méthode d'équivalence.**

- *Inconvénient de cette méthode* :

l'équivalence n'est jamais parfaite entre les formes parallèles et, à la limite, deux épreuves ne sont vraiment équivalentes que si elles comportent les mêmes items.

c) méthode du partage o des deux moitié:

La méthode du partage ("split-half" ou encore méthode de bissection) est d'une certaine façon similaire à celle du test parallèle. Les sujets passent l'épreuve une seule fois mais le test est ensuite subdivisé en deux moitiés en utilisant une des 3 procédures suivantes de bissection : (i) la partition aléatoire (random split) ; (ii) la séparation des items pairs et impairs ; (iii) la réalisation d'une partition appariée (en fonction du contenu et de la difficulté = matched split).

- ***Inconvénients de cette méthode :***

Le coefficient obtenu va être différent selon la méthode de bissection utilisée et le nombre de bissection* explose très rapidement avec le nombre des items.

5-Etude de la validité :

De façon générale, le concept de validité renvoie à la relation qui existe entre les éléments théoriques (modèles, définitions, concepts, hypothèses, etc.) et la réalité empirique supposée les représenter. Cette notion essentielle en psychologie scientifique (quelle est la validité de l'opérationnalisation que l'on propose ?) a été particulièrement étudiée en psychologie différentielle. Concernant les tests, si l'étude de la fidélité permet de répondre à la question : « le test mesure-t-il quelque chose ? », la validation d'un test suppose que l'on se pose une seconde question : « ***le test mesure-t-il ce qu'il est censé mesurer ? », ou encore « le test fournit-il bien l'information qui correspond à ce dont a besoin celui qui voudrait l'utiliser ? ».***

- La validité réfère donc à l'ensemble des éléments (preuves) qui doit conduire à nous assurer que l'interprétation des scores par les utilisateurs sera correcte. C'est un processus essentiel (fondamental) dans l'élaboration des tests. La validité d'un test est sous la responsabilité du concepteur de test (qui doit fournir des preuves de validité).

5.1-Méthodes pour évaluer la validité:

a) Validité de contenu

La notion de validité représentative ou de contenu (content-validity) porte sur la façon dont le test couvre, à partir de l'ensemble des questions posées, le domaine que l'on veut évaluer. On cherche donc à savoir dans quelle mesure les items du test constituent un échantillon représentatif du ou des comportements que l'on veut évaluer (intelligence, aptitude, trait de personnalité, etc.).

La validité de contenu suppose que des experts jugent si une mesure représente pleinement la définition de ce que l'on veut mesurer. Par conséquent, cela implique une définition théorique (du concept) acceptée par les pairs, et une sélection des indicateurs (questions) qui couvrent de manière exhaustive l'ensemble du "concept" qui veut être mesuré. La validité du contenu est une technique qualitative permettant de s'assurer que la mesure correspond au concept tel qu'il a été défini par le chercheur

- **b) Validité empirique**

  Selon **Piéron** (Vocabulaire de la Psychologie, 1951), la validité empirique s'évalue par le degré de liaison entre le rendement du sujet dans un test et son rendement dans une autre activité que le test est censé prévoir. Dans cette perspective, le test est considéré comme un instrument qui sert à prédire un comportement qu'on appelle le critère. La validation est l'étude de la relation entre le test et ce critère.

- On distingue deux types de validation empirique :
- **Validité concourante** (concurrent validity) : la mesure en question (test) et le critère ou les critères sont étudiés simultanément. Une corrélation forte entre le test et ces critères permettra d'affirmer qu'il existe une validité concourante (convergente ou concomitante sont des termes aussi utilisé).
- **Validité prédictive** : elle concerne un critère futur qui peut être corrélé avec la mesure. Il existe donc un délai entre la mesure effectuée avec une épreuve (un test) et l'évaluation sur le critère. Le test sert à pronostiquer (prédire) le critère qui sera évalué ultérieurement sur le plan empirique (par exemple, la réussite scolaire un an plus tard).

c) Validité de construit

Validité de construit (validité conceptuelle, validité théorique). Ce type de validité est utilisé lorsque l'on cherche à savoir si une mesure donnée est associée à d'autres mesures selon des hypothèses théoriques concernant les concepts qui sont mesurés. Cette démarche n'est pas spécifique à la méthode des tests, mais est une des méthodes générales de construction et de vérification d'une hypothèse en science expérimentale. Il s'agit d'étudier et de vérifier les liaisons constatées entre les variables et les hypothèses qui ont guidé les modalités de détermination de la dimension psychologique que l'on veut évaluer.