



Savoir, Comprendre, Apprendre. Leçons de Psychométrie

(4^{ème} édition)

SUPPORT DE COURS

.....*Science de Psychologie*



Document généré automatiquement et mis à jour le 16/01/2026 08:47

Site web du QUIZ : www.psychometrie.jlroulin.fr

Jean-Luc Roulin

"Theories are to measurements what maps are to navigation.

*One can make measurements without theory,
but it is the theory that gives meaning to the results."*

Robert Lloyd Ebel (1910 – 1989)

Comments on the measurement theorist's dilemma (1977, p.107)

*"Toute science, à mesure qu'elle progresse
vers la perfection, devient mathématique dans ses idées"*

Alfred North Whitehead (1861-1947)

An Introduction to Mathematics (1911)

*Ce document est généré automatiquement à partir de sources
construites pour s'afficher via un navigateur. La mise en page peut donc être imparfaite.*

Remerciements à

Corentin Gonthier (*Université de Rennes*)

Sotta Kieng (*Université de Genève*)

*et tous-tes étudiant-e-s d'Unidistance et de l'Université Savoie Mont Blanc
pour les modifications qu'ils ont proposées*

Table des matières

A. Présentation	6
B. Pratique et connaissance des tests	9
1. Mesurer	10
1.1. <i>Mesure représentationnelle</i>	10
1.2. <i>Mesure réflexive et mesure formative</i>	11
1.3. <i>Tests (définition)</i>	13
2. Usage des tests	15
2.1. <i>Standardisation</i>	15
2.2. <i>Diffusion des tests</i>	16
2.3. <i>Qui peut utiliser les tests ?</i>	16
2.4. <i>Manuel des tests</i>	18
2.5. <i>Règles d'utilisation des tests</i>	18
2.6. <i>Quand ne pas utiliser un test ?</i>	19
2.7. <i>Rapport psychométrique</i>	20
2.8. <i>Codes, standards, directives</i>	22
3. Classification des tests	22
3.1. <i>Tests d'efficience</i>	23
3.2. <i>Tests de personnalité</i>	24
3.3. <i>Le Quotient Intellectuel</i>	25
4. Le code de déontologie	28
C. Construction d'un test et création des items	31
1. Théorie classique des tests	31
2. Elaboration des items d'un test	32
2.1. <i>Format des questions</i>	32
2.2. <i>Difficulté et validité des questions</i>	34
2.3. <i>Cotation des QCM et des VF</i>	35
2.4. <i>Les biais de réponses</i>	36
3. Analyse et sélection des items	37
3.1. <i>Indice de puissance (p-index)</i>	37
3.2. <i>Indices de discrimination</i>	38
3.3. <i>Sélectionner les bons items</i>	40
3.4. <i>Le cas des items à choix multiples</i>	42
4. MRI-TRI	44
4.1. <i>Les postulats</i>	46
4.2. <i>Coupe caractéristique d'un item (CCI)</i>	46
4.3. <i>Paramètres des CCI</i>	47
4.4. <i>Les différents modèles</i>	50
4.5. <i>Des items aux individus</i>	50
4.6. <i>Intérêts et limites</i>	53
D. Les qualités métrologiques	55
1. Sensibilité	55

1.1. Sensibilité et mesure d'une dimension	55
1.2. Sensibilité et spécificité	57
2. Homogénéité et dimensionnalité	60
2.1. Homogénéité	60
2.2. Unidimensionnalité	61
3. Fidélité	62
3.1. Définitions	62
3.2. Erreur systématique - Erreur aléatoire	64
3.3. Sources de l'erreur aléatoire de mesure	65
3.4. Méthodes pour évaluer la fidélité	66
3.5. Interprétation du coefficient	73
3.6. Propriétés	74
4. Validité et validation	76
4.1. Les preuves de la validité	77
4.2. Terminologie plus ancienne	79
5. Validité vs fidélité	83
6. Contre validation	84
E. L'étalonnage d'un test	85
1. Présentation	85
1.1. Définition	85
1.2. Types d'étalonnage	86
1.3. Tables d'étalonnage	86
2. Construction d'un étalonnage	87
2.1. Quantilages	88
2.2. Rangs percentiles	89
2.3. Echelle réduite	90
2.4. Echelle normalisée	91
2.5. Scores z	94
2.6. Scores Standards Normalisés	95
2.7. Autres Scores standards	96
2.8. Un étalonnage particulier : le QI standard	99
3. Étallonages continus et inférentiels	100
4. Correspondance entre étallonages	102
5. Détermination d'un score seuil	102
F. Intervalles de Confiance et comparaison de scores	106
1. Introduction à la notion d'IC	106
2. Calculer un IC pour un score observé	106
2.1. Erreur standard de mesure et TCT	107
2.2. Erreur standard de mesure et MRI (C-ESm)	108
2.3. Méthode classique	109
2.4. Méthode corrigée	109
2.5. Exemples de calcul	110
3. Différence entre deux scores	111
3.1. Méthode de comparaison	111

3.2. Exemple de calcul	112
4. Indice de changement fiable	113
G. Complément : l'échantillonnage	116
1. Définitions	116
1.1. Échantillon	116
1.2. Population parente	116
1.3. Modèle de la population parente	117
2. Méthodes d'échantillonnage	117
2.1. Échantillonnage probabiliste	117
2.2. Échantillonnage non probabiliste	120
3. Taille des échantillons	123
H. Complément : Statistiques	126
1. Prérequis Statistiques	126
1.1. Les échelles de mesures	126
1.2. Statistiques descriptives	128
1.3. La loi normale	147
2. Introduction à l'analyse factorielle	150
2.1. La réduction des données	151
2.2. Décomposition linéaire	151
2.3. Analyse en Composantes Principales (ACP)	152
2.4. AFE	166
2.5. En résumé (à savoir)	168
2.6. Usage - avertissements	170
3. Analyse factorielle des correspondances	170
4. Analyse factorielle confirmatoire	171
4.1. Principe général	171
4.2. Un bon modèle ?	172
4.3. Pour conclure	173
I. Brèves sur des auteurs	174
J. Glossaire	181
K. Liste des principaux acronymes utilisés	194
L. Bibliographie	196

A - Présentation

La psychométrie est une discipline de la psychologie scientifique qui a pour objectif "d'évaluer", à partir de l'observation de comportements dans des conditions standardisées, des caractéristiques psychologiques telles que les aptitudes, les attitudes, les traits de personnalité ou les processus cognitifs (non observables). La psychométrie concerne donc l'ensemble des principes à la base de la mesure en psychologie. Dans l'enseignement universitaire en langue française (niveau licence), ce terme est associé plus particulièrement à la pratique et la construction des tests. Les cours de Licence en psychométrie respectent cette tradition et restreignent l'enseignement de la psychométrie à la pratique et la construction des [tests](#). De fait, ce cours aborde aussi les règles éthiques qui accompagnent cette pratique.

Cette discipline de la psychologie demande souvent à l'étudiant d'apprendre (et comprendre) de très nombreux termes nouveaux, d'acquérir des connaissances connexes comme la notion d'échantillonnage, ou de maîtriser des notions de statistiques descriptives ou inférentielles. Un cours de psychométrie est donc "rugueux" pour l'étudiant mais indispensable car il est essentiel pour apprécier l'intérêt des tests mais aussi leurs limites. Tous les professionnels qui utilisent ou font référence à des tests devraient avoir une formation minimum en psychométrie.

Enfin, la psychométrie doit son développement aux travaux et recherches d'un nombre d'auteurs importants. Certains contribuèrent directement à ce domaine de la psychologie, certains en étaient très éloignés. La dernière partie de cet ouvrage présente les biographies de quelques uns de ces auteurs.

Concernant ce document

Cette nouvelle version, contrairement aux précédentes, est réalisée alors que je ne suis plus en position d'enseignant. Initialement, je souhaitais supprimer l'accès à ce document vieillissant mais, à la demande de certains collègues, j'ai décidé de le maintenir. Il reste de nombreuses améliorations possibles et des chapitres à compléter (peut-être pour la 5^{ème} édition). J'ai cependant profité de cette 4^{ème} édition pour :

- Ajouter de nouveaux sous chapitres (exemples : le coefficient oméga, l'indice de changement fiables, etc.).
- Restructurer le document :
 - *Le chapitre pratique et connaissance des tests* apporte à la fois des informations sur ce que signifie mesurer, sur l'usage et la classification des tests, et enfin sur la déontologie.
 - *Trois chapitres importants participant tous à la construction d'un test* : le premier concerne les constituants du test (c'est à dire les items), le second l'étude des qualité métrologiques et le troisième, l'étalonnage.
 - Deux chapitres complémentaires : le premier concerne *l'échantillonnage* et le second *la comparaison de scores et les intervalles de confiances*.
 - Puis pour terminer un chapitre "[Compléments nécessaires](#)" regroupe les anciens chapitres "*pré-requis statistiques*" et "*Analyse factorielle*".
 - On retrouve toujours en fin d'ouvrage des brèves sur les auteurs, un glossaire de la plupart des termes utilisés, une liste d'acronyme courant et la bibliographie.
- Tenir compte des évolutions actuelles concernant l'étalonnage et des nouvelles recommandations concernant les tests.

- Reformuler des paragraphes qui semblaient obscurs à certains étudiants ou étudiantes (malheureusement, il en restera, je m'en excuse).
- Reprendre des formules ou des figures (simplifications).
- Mettre à jour les liens rompus. Cependant, ces liens ne sont jamais stables. Ce sont cependant tous des liens pointant vers des sites d'organisations officielles et vous devriez rapidement retrouver le document ou une version plus récente lorsque le lien est rompu (*.i.e. error 401 : The page you are looking for is not available*) !

J'espère que ces améliorations en seront réellement.

En bref...



- Ce cours correspond essentiellement à un cours de Licence de psychologie => certaines notions sont seulement introduites et des éléments complémentaires "*pour aller plus loin*" sont parfois présentés.
- L'organisation de cette version du cours n'est cependant plus contrainte par le QUIZ et qui respecte le format d'apprentissage conseillé (1. Pré-requis statistiques ; 2. Pratique des tests ; 3. Échantillonnage ; 4. Construction des tests ; 5. Métrologie ; 6. Étalonnage ; 7 . Intervalle de confiance ; 8. Analyse factorielle).
- Tous les concepts introduits sont interdépendants. Une représentation (carte mentale) est proposée pour aider à organiser les connaissances de façon cohérente dans ce domaine (www.psychometrie.jlroulin.fr/mindmap/scalp.html).
- Un glossaire des termes techniques (non exhaustif) et une présentation brève d'auteurs (incomplète) sont disponibles en fin d'ouvrage.

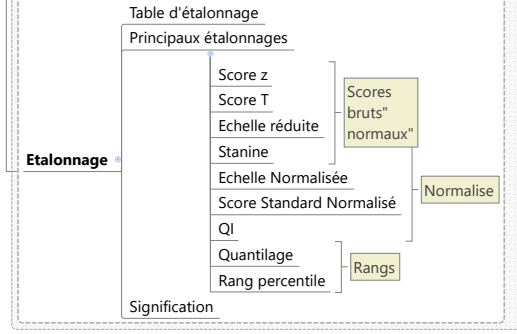
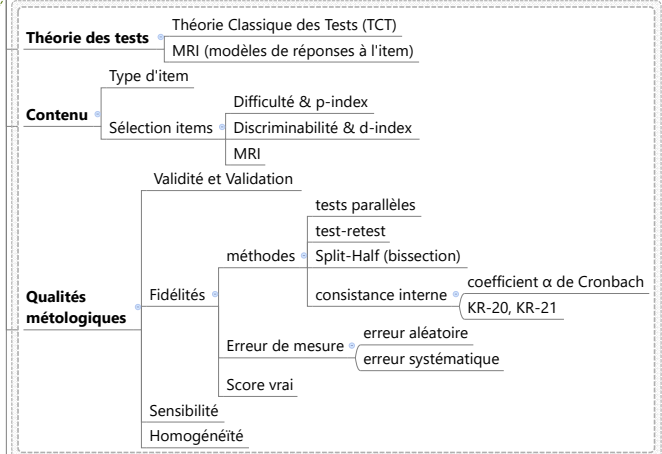
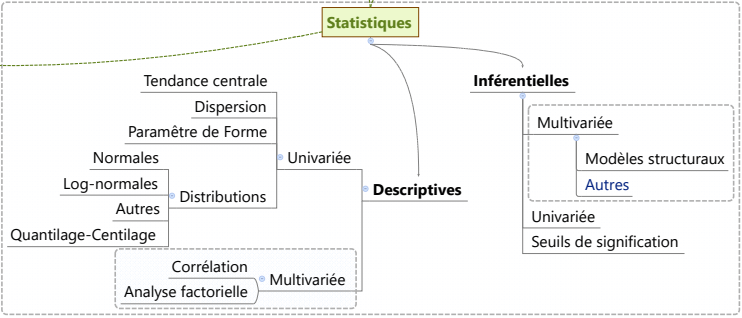
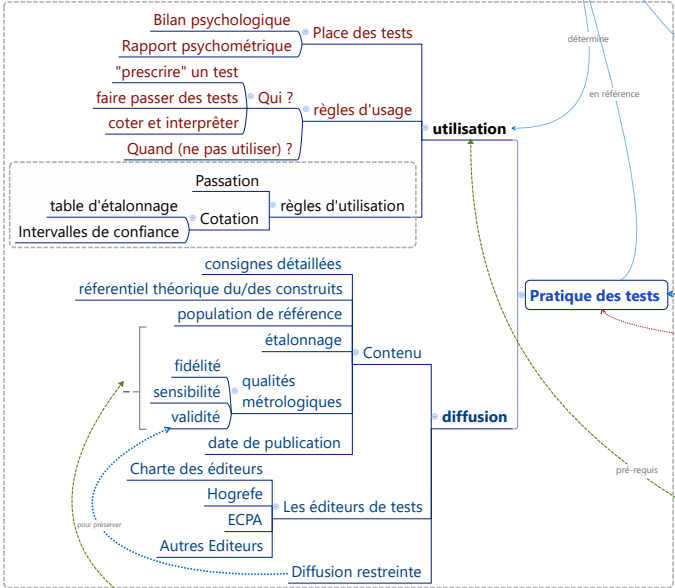
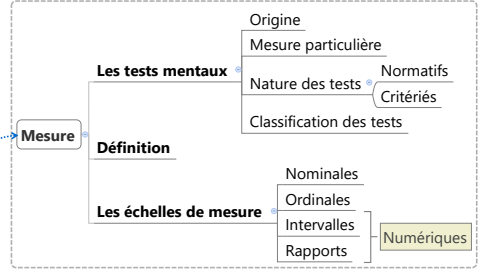
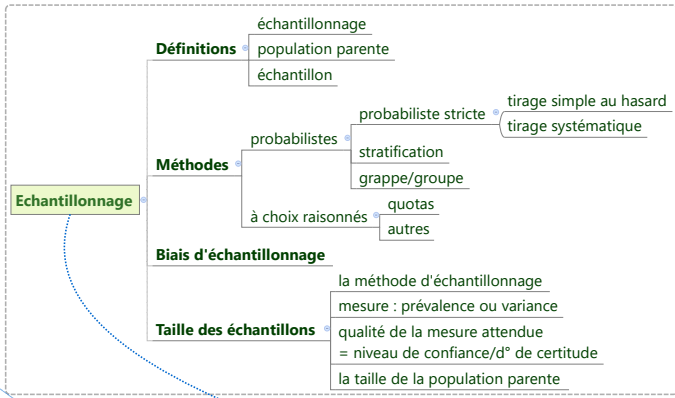
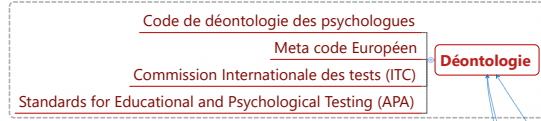


[Licence Creative Commons Attribution
Pas d'Utilisation Commerciale 4.0 International.](https://creativecommons.org/licenses/by-nc/4.0/)



Savoir et
 Comprendre
 Apprendre
 Leçons de
 Psychométrie

Cours



Psychométrie

Construction des tests

Statistiques

utilisation

Pratique des tests

diffusion

Inférentielles

Descriptives

Etalonnage

Mesure

Les tests mentaux

Définition

Les échelles de mesure

Théorie des tests

Contenu

Qualités métrologiques

Etalonnage

détermine

en référence

en référence

Nécessaire

concerne

pré-requis

pré-requis

comprendre

pour préserver

B - Pratique et connaissance des tests

Psychological tests do not measure mental processes directly; rather, they measure samples of behavior which are assumed to reflect those processes.

Anne ANASTASI (1997).

La pratique des tests n'est pas une pratique en aveugle. Dans le domaine de la santé elle s'inscrit dans une démarche clinique pour donner des éléments de réponse à des questions précises posées par le psychologue lui-même, par une institution, par un patient, etc. Le processus d'évaluation (qui comprend les tests) permet de faire une appréciation des forces, des faiblesses, et des particularités des comportements d'une personne tout en tenant compte le fait que les outils d'évaluation sont imparfaits.

En pratique, il faut savoir que certains manuels de test ne comportent pas toutes les données psychométriques utiles pour s'assurer de leurs qualités et il ne faut pas faire confiance en aveugle aux auteurs des tests ou penser qu'un test vendu est forcément fiable ([valide](#) et [fidèle](#)). Parfois, les informations sont présentes mais il est facile de voir que celles-ci sont erronées ou fausses (même pour des tests édités par des maisons d'éditions connues). Une bonne connaissance des méthodes de construction des tests est donc nécessaire (afin de porter un regard critique sur les outils que l'on utilise).

Au niveau individuel, le psychologue doit limiter les erreurs dans l'interprétation et doit intégrer les caractéristiques psychométriques des tests dans sa pratique. L'utilisation de tests obsolètes ou non valides peut ouvrir la voie à des poursuites en faute professionnelle. Par exemple, aux États-Unis plusieurs cas documentés montrent que l'utilisation de tests psychométriques non validés, obsolètes ou inadaptés, a conduit à des réponses disciplinaires ou juridiques, allant jusqu'à des sanctions, amendes, voire interdictions d'usage, notamment dans le cadre professionnel. En fait, on considère comme faute professionnelle l'usage inapproprié d'un test par méconnaissance entre autres de ces propriétés psychométriques.

Des connaissances théoriques en psychométrie sont donc un des éléments essentiels à une bonne pratique dans le respect du code de déontologie professionnelle, mais aussi et surtout, dans le respect des personnes qui font confiance aux psychologues.

Pour aller plus loin ...

Les biais de l'évaluation subjective

Spontanément nous réalisons au quotidien des évaluations de la taille d'un objet, de la température, du poids, mais aussi des évaluations concernant des caractéristiques comme les intérêts d'une personne, sa personnalité, ses compétences, etc. Toutes ces évaluations sont des évaluations subjectives.

Pour des données comme la taille, la température, nous savons que nos mesures subjectives sont entachées d'erreur et nous acceptons facilement de contrôler ou corriger ce jugement par un instrument de mesure objectif (mètre, thermomètre, etc.). Nous remplaçons même cette évaluation subjective par une évaluation "outillée" si la mesure revêt une importance professionnelle.

Qu'en est-il concernant l'évaluation subjective des processus psychologiques ? L'évaluation des processus psychologiques est plus complexe que l'évaluation de la taille d'un objet et il est donc normal de s'appuyer, lorsque l'on est un ou une professionnelle, sur des évaluations standardisées ([tests](#)). En effet, en quoi et pourquoi l'évaluation subjective des processus psychologiques serait plus simple que

l'évaluation de caractéristiques comme la taille et le poids ? Avons-nous une meilleure connaissance et expertise dans le domaine psychologique ?

Les chercheurs ont bien entendu travaillé sur l'évaluation subjective. En résumé, on peut dire que pour les processus psychologiques, notre expertise supposée est souvent mise en défaut. Les recherches dans ce domaine ont conduit à décrire plusieurs sources d'erreur de jugement ou d'évaluation, que tous psychologues devraient connaître :

1. **L'erreur fondamentale d'attribution.** On a tendance à expliquer les comportements d'autrui par des causes internes (la personne) plus que par des causes externes (environnement).
2. **Le biais de confirmation.** On a tendance à favoriser les informations qui vont dans le sens de nos attentes et de minimiser celles qui les infirment.
3. Nous sommes sous l'influence de **nos stéréotypes** (croyances relatives associées à des groupes). De nombreuses études ont montré que nos réponses, par exemple, à des questions portant sur une personne sont dépendantes des stéréotypes que l'expérimentateur va activer dans la situation d'évaluation.
4. **L'effet de halo.** De façon générale, il s'agit d'une perception sélective d'informations allant dans le sens d'une première impression que l'on cherche à confirmer. L'évaluation d'une caractéristique (par exemple : souriant) a tendance à être extrapolée à d'autres caractéristiques (gentil, sympathique, présent, motivé, etc.).
5. **L'effet de contraste.** Un jugement (une évaluation) est dépendant des jugements (évaluations) effectués auparavant. Par exemple, si on vient de voir une personne âgée présentant des difficultés cognitives majeures, on peut sous-estimer les difficultés cognitives de la personne suivante lorsque celles-ci sont moins marquées.
6. **Illusion des séries.** Biais de raisonnement consistant à percevoir à tort des coïncidences dans des données.
7. etc.

Ces mécanismes sources d'erreurs sont confirmés dans de nombreuses études (Khaneman, 2012) et montrent que les psychologues dans leur pratique professionnelle ne peuvent pas se limiter à une évaluation subjective. Ils se doivent aussi d'utiliser des outils de mesure plus objectifs (cf. aussi à ce sujet les différents [codes de déontologie](#)).

1. Mesurer

Like other tools of science, measurement extends the range of our experience and gives precision to our observations

Stevens, S. S., 1946

Avant de définir la notion de test en psychologie (ou test mental), il est important de s'arrêter quelques instants sur le concept de mesure. L'objectif n'est pas d'apporter des réponses toutes faites mais de faire prendre conscience de la complexité des postulats ou des différentes significations que peut prendre un simple terme comme mesurer. Le test mental n'étant qu'un cas particulier qui n'est pas, nous le verrons, une mesure au sens strict mais un positionnement sur une "échelle de mesure" par rapport à un groupe de référence.

1.1. Mesure représentationnelle

Dans une conception classique la mesure d'une quantité consiste à déterminer combien de fois elle

contient une quantité élémentaire (quantité de référence ou étalon) du même type. Le système international des unités de mesure identifie 7 unités fondamentales : mètre, kilogramme, seconde, ampère, kelvin, mol, candela. Ces unités de mesure correspondent respectivement à des quantités physiques de longueur, masse, temps, courant électrique, température, quantité de matière et intensité lumineuse.

La mesure représentationnelle (sans contrainte de référence à une quantité élémentaire) est une procédure précise et explicite qui attribue aussi des nombres aux "objets". La règle d'attribution définit la signification de la mesure. Potentiellement, tout peut donc être mesuré. Le problème reste cependant la signification de la mesure qui dépend toujours de cette règle d'attribution.

La mesure est donc une notion (processus) complexe qui concerne des caractéristiques d'un objet ou d'une personne (la taille, la tension, l'extraversion, l'intelligence, la température corporelle, etc.). Toutes ces mesures n'ont pas le même statut mais de façon générale, mesurer, c'est toujours **attribuer des nombres aux objets, selon des règles déterminées**. Les règles d'attribution des nombres donnent la signification de la mesure. Selon les propriétés des nombres qui sont conservés suite à cette attribution, la nature de ces échelles de mesure est différentes (cf. chapitre [échelles de mesure](#)).

Pour aller plus loin...

- *Un test psychologique n'est pas une mesure au sens classique du terme. Cela supposerait un étalon unique définissant ainsi l'unité de mesure. Un test psychologique est cependant une mesure représentationnelle dans le sens où le résultat est l'attribution d'un ou plusieurs nombres selon des règles bien définies devant respecter certaines propriétés. L'interprétation de ces nombres reste cependant complexe et l'utilisateur d'un test doit avoir compris les règles permettant ces attributions mais aussi la signification de ces règles et les limites associées à leur usage.*
- *Il existe de nombreux débats sur ce qu'est la mesure en psychologie. L'un des critiques les plus virulents en langue française est probablement Stéphane Vautier qui développe l'idée qu'en l'état actuel des connaissances en psychologie, "l'hypothèse scientifique par défaut est qu'on ne sait mesurer aucune grandeur psychologique, si mesurer signifie qu'on sait observer une certaine variabilité dans des conditions expérimentales telles que cette variabilité ne dépende que de la variabilité de la grandeur qu'on veut mesurer" (blog : [Carnet d'enseignement et de recherche de Stéphane Vautier](#), 2017). En accord partiel avec sa conception, nous considérons que les tests psychologiques dans la pratique clinique ne sont que des mesures représentationnelles et à ce titre des instruments d'observation et d'évaluation standardisés dont les résultats se doivent d'être interprétés en regard de l'ensemble des autres informations dont dispose le psychologue.*

1.2. Mesure réflexive et mesure formative

En psychologie différentielle, de façon très générale, une dimension correspond à un continuum le long duquel un trait, un processus ou une performance varie en intensité. On distingue les [dimension théorique](#) et les dimensions [dimensions opérationnelles](#).

Une [dimension théorique](#) (intelligence, extraversion, anxiété, aptitude numérique) n'est pas observable directement (par définition, c'est une "construction conceptuelle"). Elle est appréhendée à travers une [variable latente](#) (qui est un construit statistique) associée aux différences interindividuelles observées sur des variables manifestes. Ces variables manifestes sont les performances observées sur différentes épreuves ([dimensions opérationnelles](#)).

Pour définir une variable latente, on utilise ce qu'on appelle un modèle de mesure. On distingue actuellement deux types de mesures : les mesures réflexives et les mesures formatives.

Mesure réflexive. Dans la démarche classique de construction des tests, on suppose qu'il existe une dimension sous-jacente théorique (non observable) et que le résultat au test est causé essentiellement par cette dimension (la variable latente). Par exemple, si vous avez une bonne aptitude spatiale, vous avez une probabilité plus élevée d'obtenir un score élevé aux tests mesurant l'aptitude spatiale. La variable latente associée à la dimension théorique prédit la variable mesurée. Dans ce cas on parle de modèle de mesure réflexive. Dans un modèle de mesure réflexive, on s'attend à ce que les corrélations entre les indicateurs (variables manifestes) d'une même variable latente soient fortes.

La mesure formative. Dans ce type de mesure la causalité est inversée. En effet, on parle de modèle de mesure formative lorsque les variables mesurées sont la cause du "construit" mesuré, définissent ce construit. Une variable est dite formative lorsqu'elle est «formée» ou directement modifiée et influencée par les indicateurs. Ce sont donc les indicateurs qui "créent" le construit mesuré. Par exemple :

- La valeur d'une voiture est déterminée par son âge, l'état, la taille, la marque. On peut choisir d'autres facteurs pour déterminer sa valeur.
- Le niveau socio-économique est déterminé par la source du revenu, le métier, le voisinage, l'habitation (exemple : l'indice de WARNER, 1960).
- etc.

Dans ces exemples on note que pour ce modèle de mesure, il n'y a pas d'hypothèses sur les corrélations (i.e. covariances) entre les composants de la mesure. Elles pourraient être égales à zéro, être positives ou négatives. Les variables latentes formatives sont "le résultat" des variables manifestes (via une combinaison linéaire).

Pour résumer

Si on représente par des rectangles les variables observées (variables manifestes), par des ovales les variables latentes et par des flèches les relations de causalité, la distinction entre variables manifestes et latentes est représentée de la façon suivante :

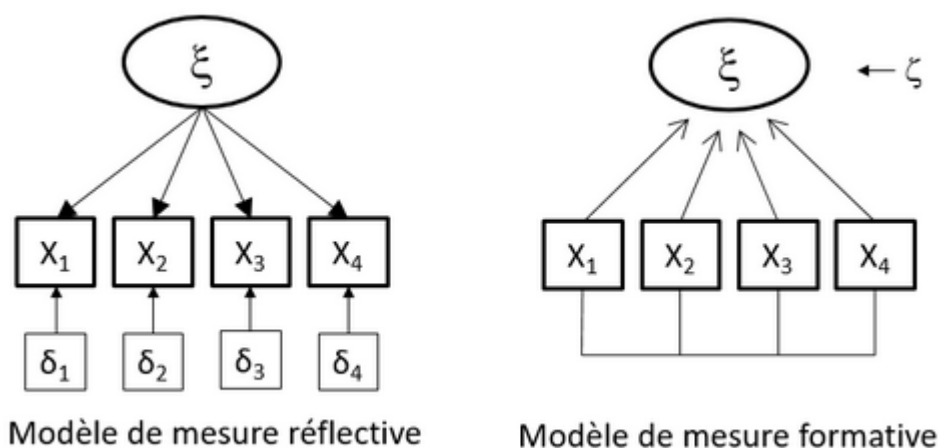


Figure B.1 : Représentation des relations entre variables manifestes (X) et variable latente pour les modèles de mesure réflexive et mesure formative

Conséquences. Cette distinction entre mesures réflexives et formatives n'est pas une simple réflexion sur la nature ontologique des construits : dans le cadre d'une mesure formative le construit, le sens de la mesure est donné par les indicateurs que l'on choisit d'associer alors que dans une mesure réflexive

le construit existe indépendamment des indicateurs. En fait, cette distinction a des conséquences sur les modèles statistiques utilisés et sur la signification de la mesure en psychologie. De très nombreux débats ont encore cours sur la nature des construits et les méthodes. Dans la théorie [classique des tests](#), le modèle est un modèle de mesure réflexive, modèle qui correspond à la représentation implicite ou explicite des psychologues et de nombreux chercheurs en psychologie. Dans le cadre d'un cours de licence en psychologie nous n'abordons pas ces débats, même si cette réflexion sur le sens de la mesure est essentielle. Que fait le psychologue quand il "évalue" ?

Pour aller plus loin...

Les débats autour de ces deux modèles de mesures font référence aux articles de Borsboom Mellenbergh et van Heerden (2003, 2004). Par ailleurs il existe de très nombreux numéros spéciaux de journaux scientifiques sur ce sujet (Measurement: Interdisciplinary Research & Perspective, Journal of Educational Measurement).

1.3. Tests (définition)

1.3.1 Tests mentaux

Les tests mentaux ont pour objectif de "mesurer" des caractéristiques psychologiques. Ils cherchent à évaluer le plus objectivement possible les différences interindividuelles dans le domaine de la personnalité (au sens large) ou de la cognition. Il n'est pas possible cependant d'évaluer directement des différences comme l'intelligence, l'extraversion-introversion, l'aptitude spatiale, le névrosisme, etc. Ces concepts sont des concepts théoriques ([dimensions théoriques](#)) dont le psychologue propose une ou plusieurs définitions. Il ne peut observer que des manifestations de différences interindividuelles dans des [situations standardisées](#) (un inventaire de personnalité sera composé par exemple d'items portant sur les comportements habituels de la personne ; un test d'intelligence quand à lui proposera des questions qui permettront l'expression d'un ou plusieurs aspects de ce qu'on appelle intelligence). L'ensemble des réponses produites par une personne permettra d'obtenir un score brut qui sera comparé aux réponses habituellement observées dans un [échantillon de standardisation](#) (échantillon de référence ou de comparaisons). Ce score devra posséder certaines propriétés ([qualités métrologiques](#)) et sera transformé ([étalonné](#)) de façon à connaître la position d'une personne par rapport aux personnes de cette échantillon (parfois appelé [normatif](#)). Cette transformation en scores standardisés dont les propriétés sont connues de tous les psychologues facilitera l'analyse et l'interprétation des résultats.

Définition d'un test psychologique : On appelle test mental une situation expérimentale standardisée servant de stimulus à un comportement (Pichot, 1997) et qui doit répondre aux critères suivants :

- ◆ Ce comportement est comparé à d'autres individus placés dans la même situation.
- ◆ Le comportement déclenché est enregistré avec précision, objectivé et catégorisé selon des règles précises.
- ◆ Les propriétés de la mesure sont connues ([sensibilité](#), [fidélité](#), [validité](#)).
- ◆ L'utilisation/l'interprétation répond à des normes et nécessite des connaissances techniques et théoriques

Autrement dit : le test doit permettre de décrire le comportement et situer ce comportement dans un groupe biologiquement (exemple : âge) et socialement déterminé (test dit normatif). Pour que cette

comparaison ait un sens, il faut donc que le test soit identique pour tous, tant pour la passation que pour l'appréciation des réponses (standardisation).

A savoir :

- **Le test n'a pas (jamais) de valeur universelle.** Il permet de situer un individu dans un groupe.
- Le terme de test a de multiple sens au quotidien, mais le terme de test psychologique ou test mental à un sens bien spécifique (cf. ci-dessus).
- On présente souvent les anciennes épreuves de sélection des fonctionnaires* chinois comme des précurseurs des tests (probablement à tort) car s'ils présentaient certainement des caractéristiques communes avec les tests, il s'agissait cependant d'outils de sélection, d'évaluation de connaissances, plus proches de la notion de concours que de tests. Par ailleurs, ce système a largement évolué pendant 1300 ans et n'a pas toujours reposé sur une véritable évaluation des compétences.
- Aujourd'hui, les applications de la méthode des tests concernent le monde de la santé, celui de l'éducation et de la formation, et enfin le monde du travail.

(*) système keju ou mandarinisme en usage depuis plus de 1300 ans et aboli en 1905. Prend vraiment la forme d'épreuves "plus standardisée" sous la dynastie des Tang (612 après JC) (source : Wang, 2004)

Étymologie de terme test

Le mot test anglais proviendrait de l'ancien français ("test") issu lui-même du latin *testum* (récipient rond). Au début du XII^{ème} siècle, en ancien français, "test" signifie "débris de pot cassé" (devient tesson), puis l'usage de test au XIII^{ème} siècle est associé à "Pot de terre". En 1762, "têt" devient « récipient de terre dans lequel on fait l'opération de la coupellation » [source : <http://www.cnrtl.fr/>].

Le mot test prend différentes significations au cours de son histoire. En 1686, ce terme anglais n'a rien avoir avec la signification actuelle. C'est un serment, introduit par acte du parlement de 1673, par lequel on renonce à la croyance de la transsubstantiation (Angleterre). Son usage change au cours du 19^{ème} siècle et ce terme anglais "test" reprend une signification plus proche de son origine latine ou de l'ancien français* puisque qu'il désigne "une coupelle de métallurgiste servant à isoler les métaux précieux", puis signifie "ce qui permet de déterminer la qualité ou la pureté de quelque chose". [James Mc Keen Cattel](#) introduit en 1890 le syntagme "mental test" (l'usage général à cette époque signifiait "épreuve permettant de mesurer des phénomènes ou des aptitudes dans des conditions expérimentales précises". Avec [Alfred Binet](#), l'usage devient "épreuve permettant de mesurer des phénomènes ou des aptitudes dans des conditions expérimentales précises". Le Comité des termes techniques français a proposé de réserver le mot "test" à la psychologie et de remplacer le mot test par Essai, essai témoin, épreuve dans les autres domaines (source : Le Centre National des Ressources Textuelles et Lexicales : <http://www.cnrtl.fr>). Cette recommandation ancienne (1959) n'est pas appliquée.

1.3.2 Tests normatifs et critériés

La définition précédente concerne uniquement ce qu'on appelle les tests normatifs. L'utilisation du terme de test dans un cadre plus large que la psychologie (comme la sélection de personnel ou l'évaluation) conduit Glaser (1963) à distinguer :

- **Les tests normatifs** : il s'agit de situer un individu par rapport à un groupe de référence. Ils répondent totalement à la définition précédente des tests mentaux ou à la définition des tests en psychologie.
- **Les tests "critériés" (ou à « référence critérielle »)** : l'objectif est de situer une personne par

rapport à un univers de contenu ou en référence à un critère. Par exemple, un test de niveau en mathématique est dit à référence "critérié" si l'objectif est de situer la performance d'une personne en référence à un nombre de connaissances acquises.

Les tests critériés restent toujours, selon nous, des mesures représentationnelles. Par exemple, pour un test de connaissances, le nombre (même pondéré) de connaissances acquises ne peut pas être considéré comme la référence à une quantité élémentaire (étalon) qui serait "une connaissance particulière". On compare les connaissances acquises à un ensemble de connaissances (souvent partiellement hiérarchisées) pour savoir où se situe la personne par rapport à cet ensemble de connaissances.

Selon la nature du test (normatif ou critérié), la logique d'interprétation varie. On distingue habituellement :

- **Logique critériée** : on compare à un critère externe (cf. tests critériés ci-dessus).
- **Logique normative** : on compare la performance à la performance à un groupe de références (cf. test normatif). Avec un test critérié, on peut utiliser cette logique si on a des informations sur les performances habituelles d'un groupe de référence.
- **Logique ipsative** : comparer le score avec d'autres scores de la même personne (analyse le profil : on interprète un score en référence à d'autres scores de la même personne).

Attention : selon le contexte, le terme ipsatif peut prendre un sens partiellement différent. Par exemple, en psychométrie, les mesures « à choix forcé » dans lesquelles on demande de choisir entre deux branches d'une alternative s'appelle aussi parfois mesure ipsative. Un exemple de question correspondant à une mesure ipsative : *je préfère : (a) la psychométrie ; (b) les statistiques.*

2. Usage des tests

2.1. Standardisation

La standardisation des tests est essentielle à respecter. C'est elle qui donne la valeur aux tests et permet de réaliser la comparaison : « **toutes choses étant égales par ailleurs** ». Elle permet donc de s'assurer que les différences entre les scores observés habituellement et les scores observés lors de la passation du test ne sont pas la conséquence de variations de la situation (les événements particuliers à la situation d'évaluation devront donc être pris en compte lors de l'interprétation d'un test). La standardisation doit permettre d'assurer :

- que les conditions de passation sont les mêmes que celles du groupe de référence utilisé pour [l'étalonnage](#) ;
- que le matériel et la procédure sont les mêmes ;
- que la notation et le calcul des scores sont identiques pour tous afin de garantir le résultat quel que soit l'examineur.

ATTENTION

→ La standardisation n'est jamais parfaite et sa bonne application dépend en partie de l'expertise de l'observateur. La passation des tests est une expertise qui s'acquière sur le terrain et l'apprentissage de la passation des tests devrait être toujours accompagnée.

- Respecter la standardisation permet de minimiser les biais que pourrait introduire l'observateur.
- La standardisation *ne signifie pas des consignes stéréotypées et mécaniques.*
- La valeur de la standardisation dépend du mode de passation : collectif - informatisé - individuel.
- Des facteurs introduisant des biais de standardisation dans les passations individuelles existent. Ce sont essentiellement :
 - les attentes de la personne ;
 - l'attitude de l'observateur ;
 - les caractéristiques personnelles de l'observateur (sexe, âge, apparence physique, etc.). Ces effets sont faibles mais plus marqués chez l'enfant.

2.2. Diffusion des tests

Les tests sont des instruments standardisés permettant de situer un individu par rapport à d'autres ayant été dans les mêmes conditions de passation. L'élaboration des tests est un processus long et les questions posées (questionnaires de personnalité, intelligence, autres) sont l'objet d'un traitement et d'une élaboration qui permettent de s'assurer que ce l'on mesure correspond à ce que l'on veut mesurer. Cependant, cette hypothèse reste vraie si les personnes ne connaissent pas préalablement le contenu des tests.

Si les tests sont diffusés et accessibles plus ou moins facilement certaines personnes connaîtront le test ou son contenu avant et d'autres non. L'interprétation des résultats devrait alors prendre en compte cet aspect (ce qui serait un moindre mal). La [validité](#) du test est en fait alors largement remise en question. En effet, si on connaît préalablement un test, il est probable que les réponses n'évaluent plus la même chose ou la/les mêmes dimensions psychologiques. Par exemple, si on apprend ou si l'on s'entraîne sur des épreuves typiques des tests d'intelligence (d'intelligence fluide par exemple), le score sera probablement élevé mais la comparaison avec les scores habituellement observés n'aura plus le même sens.

En fait, diffuser un test diminue la validité du test voire l'invalide. En conséquence, le psychologue ne diffuse pas le contenu des tests ni ne les cède à des non psychologues de façon à préserver leur [validité](#).

2.3. Qui peut utiliser les tests ?

Les tests. Un usage réservé aux psychologues ?

Actuellement de nombreux psychologues et enseignants-chercheurs considèrent que l'utilisation des tests est un acte psychologique inséparable du titre de psychologue. Ils en déduisent ou affirment que seuls les psychologues peuvent acheter et utiliser des tests. Formellement cette règle est fautive et dépend des lois et règles en vigueur dans un pays.

En France par exemple, il n'existe pas de réglementation concernant l'utilisation des tests et seul l'usage du titre de psychologue est réglementé. L'utilisation des tests, comme l'achat des tests, n'est donc a priori réservé à aucune profession particulière. Certains réclament (et il existe une réflexion européenne à ce sujet) que l'utilisation des tests (passation, cotation, interprétation) soit réglementée. Cette réglementation pourrait prendre la forme d'une certification et toutes formations donnant le droit à l'usage des tests ou de certains tests devraient voir un enseignement minimum de psychométrie.

Remarque : les éditeurs de test restreignent néanmoins la vente aux personnes possédant les titres et qualifications professionnelles pour l'usage des tests (ce n'est pas uniquement les psychologues). Ce

principe est conforme aux "Standards" adoptés par des organisation telles que [l'American Psychological Association](#) (APA), [l'European Test Publisher Group](#) (ETPG), et [l'international Test Commission](#) (ITC). Selon le test, les professions pouvant l'acheter sont différentes.

Principaux utilisateurs des tests.

Les tests sont utilisés dans différents domaines et dans différents secteurs (école, santé mentale, orientation scolaire et professionnelle, médical, recherche, etc.). Parmi les utilisateurs on peut citer : les psychologues cliniciens, les neuropsychologues, les psychologues scolaires, les psychologues du travail (sélection du personnel et orientation). Ces exemples concernent essentiellement différentes facettes du métier de psychologue mais d'autres professions peuvent utiliser des tests comme les orthophonistes ou encore, avec des épreuves plus ou moins standardisées, les médecins, les psychomotriciens, les ergothérapeutes les enseignants. Dans le domaine de la recherche (en psychologie ou en sciences de l'éducation) le test peut jouer des rôles variés et contrairement aux exemples précédents, l'objectif est rarement de contribuer à une évaluation individuelle. Le plus souvent, en recherche, ces tests contribuent à sélection d'échantillon, à la description de population ou encore peuvent servir de mesures d'intérêt (soit comme variable dépendante ou comme variable contrôlée).

Référentiel européen.

Il existe un « référentiel » général européen qui a été élaborée conjointement par l'EFPA ([European Federation of Psychologists' Associations](#)) et l'EAWOP ([European Association of Work and Organizational Psychology](#)) pour définir des niveaux d'habilitation pour l'usage des tests. Le schéma ci-dessous résume de façon très simplifiée ces différents niveaux qui par ailleurs restent une recommandation :

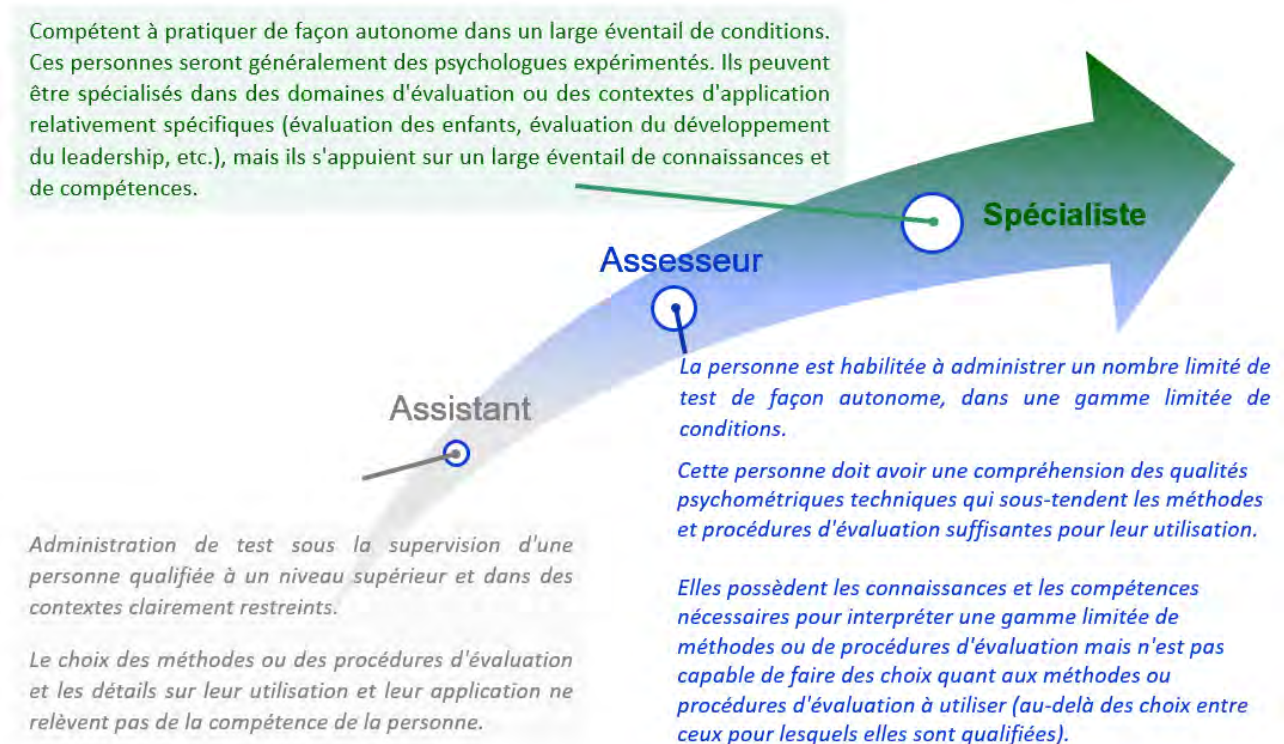


Figure B.2 : Niveau d'habilitation à l'usage des tests (document source : rubrique "Test User standard documents" sur le site de l'EFPA1 : https://www.efpa.eu/sites/default/files/2023-08/115-EFPA_BoA_Qualification_specifications_Assessment_All_Levels_15-03-2013.pdf - site consulté le 04 juillet 2025).

2.4. Manuel des tests

Le manuel des tests est un outil important pour les psychologues. Un bon manuel (donc un manuel qui respecte les règles de la déontologie professionnelle) doit préciser a minima :

- Les objectifs généraux du test
- le référentiel théorique des concepts utilisés ;
- Les grandes étapes et les justifications ayant conduit à sélectionner les questions constitutives de l'épreuve.
- les consignes précises de passation et de cotation (indispensable pour une bonne standardisation) ;
- les données permettant d'apprécier les qualités psychométriques de l'instrument ;
 - la définition de la population de référence (ou des populations de références) et la méthode d'échantillonnage.
 - Les qualités métrologiques : sensibilité, [fidélité\(s\)](#), [validité](#) (travaux de validation).
- la(les) table(s) d'étalonnage ou les règles de transformation des scores bruts en scores étalonnés ;
- la date de publication des normes d'étalonnage ;
- éventuellement des exemples cliniques quand l'épreuve s'y prête.

Attention

Il y a de plus en plus de recherches pour adapter des tests conçus initialement dans d'autres pays (essentiellement nord-américains). Toutefois, en pratique, beaucoup de tests publiés ne comportent pas les données psychométriques permettant de les évaluer et d'apprécier leur validité.

Il est de la responsabilité du psychologue de prendre en compte dans l'évaluation d'une personne toutes les données psychométriques utiles : groupe de référence pour l'étalonnage, fidélité(s), validation, date à laquelle le test a été étalonné, etc. Le psychologue doit toujours pouvoir assurer qu'il utilise des outils dont il connaît les qualités psychométriques et il doit être capable de justifier ses choix (utiliser des instruments non validés pour en tirer une conclusion est a minima une faute morale mais aussi et surtout un non respect des personnes évaluées comme un non respect du code de déontologie des psychologues).

2.5. Règles d'utilisation des tests

Il existe un certain nombre de principes ou règles (devoirs) pour le psychologue dans l'utilisation des tests. Ces règles se retrouvent de façon explicite ou implicite dans les codes de déontologie professionnelle (France, Suisse, Belgique, USA, Canada, etc.).

- (1) Dans l'exercice de sa profession, le psychologue doit tenir compte des principes scientifiques généralement reconnus en psychologie.
- (2) Le psychologue ne diffuse pas le contenu des tests ni ne les cède à des non psychologues ([pour des problèmes de validité](#)).
- (3) Le test n'est pas une fin en soi, c'est un outil standardisé qui complète, éclaire des données recueillies sur la personne. Il doit s'inscrire dans une démarche hypothético-déductive.
- (4) Le psychologue ne doit pas remettre à autrui, sauf à un psychologue, les données brutes et non

interprétées inhérentes à une consultation psychologique.

(5) Le psychologue doit éviter toute possibilité de fausse interprétation ou d'emploi erroné des informations qu'il transmet à autrui.

→ le psychologue doit rédiger des comptes rendus qui dans leurs formes doivent s'adapter aux destinataires (parents, psychiatres, collègues, etc.)

→ Lors d'un bilan, la personne concernée doit toujours avoir une restitution du bilan (orale et écrite). Cette restitution doit être expliquée et discutée et on doit s'assurer que la personne a compris.

(6) Le rapport psychométrique :

→ Ne doit pas être le relevé du résultat des tests passés. C'est un composé cohérent de toutes les données relatives à l'évaluation.

→ Il doit éviter de faire des commentaires sur ce qui est moyen ou « normal » et mettre l'accent sur ce qui concerne spécifiquement la personne.

→ Le rapport doit répondre aux questions posées et doit éviter tout ce qui ne concerne pas ces questions.

→ Il doit être rédigé en fonction des besoins et des connaissances des personnes auxquelles il est destiné (en respectant le code de déontologie).

Attention.

Le test n'est pas, et ne doit pas être, le seul outil du psychologue lors d'un bilan psychologique. Le psychologue dispose de 4 outils : l'entretien (anamnèse, situation actuelle, etc.), l'observation, les sources d'informations externes éventuelles (proche, institution, bilan antérieur, etc.) et les tests.

2.6. Quand ne pas utiliser un test ?

Voici selon Urbina (2014), les 10 raisons qui devraient conduire un psychologue à ne pas utiliser un test :

1. Les objectifs du test sont inconnus ou peu clairs pour le psychologue.
2. Le psychologue n'est pas familier avec le test et pas assez entraîné à l'utilisation de ce test.
3. Le psychologue ne sait pas à qui est destiné le test ou comment seront utilisés les résultats au test.
4. Les informations que pourraient fournir le test sont déjà disponibles ou peuvent être obtenus par d'autres moyens ou des sources plus sûres.
5. La personne devant être testée n'est pas d'accord ou pas prête pour coopérer à une situation de test.
6. Le test ou la situation de test peut engendrer un préjudice à la personne testée.
7. L'environnement et/ou les conditions de passation ne sont pas adaptés à la situation de test.
8. Le format du test n'est pas adapté en raison de l'âge, d'aspect linguistique, culturel, ou de tous les autres facteurs qui rendent invalides les données obtenues.
9. Les normes sont trop anciennes ou inadaptées et inapplicables à la personne testée.
10. Le manuel du test (documentation) concernant le test ne donne pas d'informations suffisantes concernant la fidélité et la validité des scores observables.

2.7. Rapport psychométrique

L'étape finale du processus d'évaluation est la rédaction d'un rapport psychométrique (bilan). Ce rapport n'est pas un simple relevé des résultats des tests qui ont été passés, c'est un composé cohérent de toutes les données relatives à l'évaluation. Les principes de base du rapport psychométrique ont été largement discutés dans la littérature et plusieurs auteurs proposent d'organiser le contenu d'un rapport psychométrique suivant un schéma (Wolber et Came, 2002) en 8 points :

1. Les données personnelles (nom et prénom, date de naissance, date de l'évaluation);
2. Mandat (raison de l'évaluation);
3. Méthodes et instruments d'évaluation (entrevue, tests, étude de dossiers);
4. Présentation de la personne (statut social, relations familiales, scolarité, histoire du développement, situation actuelle, éléments les plus significatifs de la vie, portrait clinique);
5. Observations et conditions de passation (comportements et attitudes lors de l'entrevue, coopération, motivation, motricité, empathie);
6. Résultats, impressions sur le plan clinique et interprétation (inférences basées sur des variables significatives des tests et sur les observations compte tenu de l'objectif de l'évaluation; discussion sur les résultats par thèmes);
7. Éléments de diagnostic (appartenance à une catégorie psychologique ou clinique);
8. Résumé et recommandations (conseils reliés au but de l'évaluation).

De façon générale le rapport doit éviter de faire des commentaires sur ce qui est moyen ou « normal » pour mettre l'accent sur ce qui concerne spécifiquement une personne dans son environnement particulier. Les rapports doivent tenter de répondre à des questions spécifiques posées par la personne évaluée (ou par celle qui l'a adressée). Ce rapport évite donc tout ce qui ne concerne pas ces questions. Le rapport doit être rédigé en fonction des besoins et des connaissances des personnes auxquelles il est destiné (tout en respectant les principes du code de déontologie). La responsabilité des conclusions présentées dans un rapport relève du psychologue.

Afin de vérifier la qualité d'un rapport d'évaluation psychométrique, Tallent (1993) propose une série de questions que l'on devrait se poser suite à la rédaction d'un rapport :

- Est-ce que le rapport respecte les normes relatives à l'éthique et indique que le professionnel prend ses responsabilités face au client ?
- Est-ce que les interprétations qui sont présentées ont été faites d'une façon responsable ?
- Est-ce que les concepts trop abstraits ou trop théoriques ont été écartés ?
- Est-ce que le rapport est organisé de façon efficace et logique ?
- Est-ce que le rapport se contredit ?

Enfin, 87 principes à respecter pour une utilisation compétente des tests ont été posés par Eyde et al. (1993, p. 213-215). Les principaux principes sont repris dans le tableau ci-dessous (extrait de Bernier, J.J., Peitrulewicz, B., 1998). Il existe aussi de nombreux textes (chartes, guidelines, etc.) reprenant ces principes. Un psychologue devrait suivre systématiquement l'évolution des textes de référence édités par la commission internationale des tests : https://www.intestcom.org/files/guideline_test_use.pdf. Remarque : ce texte date de 2013. Il a été traduit dans de très nombreuses langues sauf en Français ! (<https://www.intestcom.org/page/15>)

Extrait de 87 principes pour une utilisation compétente des tests (Bernier, Peitrulewicz, 1997)

11. Empêcher les personnes qui passent les tests de consulter ceux-ci avant la passation.
12. Conserver en lieu sûr les clefs de correction et le matériel des tests.
13. Ne pas modifier la procédure de passation prévue afin de l'adapter à des individus en particulier (c'est-à-dire lire les items du test à une personne, définir des termes spécifiques à l'intérieur d'un item ou encourager un individu à reconsidérer une réponse).
14. Évaluer les tests et détecter le matériel de promotion trompeur (Connaître les tests et leurs limites).
15. Veiller à ce que la passation des tests soit assurée par un personnel qualifié.
16. Choisir pour l'examen un endroit permettant l'optimisation du rendement du sujet (par exemple, un bureau).
18. Prendre conscience que les scores à un test représentent seulement un point dans le temps. Ils sont sujets à changer avec l'expérience.
20. Considérer les erreurs de mesure dans les résultats d'un test.
22. Être conscient de la nécessité d'avoir plusieurs sources de données convergentes.
25. Comprendre les normes et leurs limites.
26. Reconnaître que le contenu du test est limité.
27. Reconnaître les répercussions de la validité d'un test.
28. Garder le contact avec son domaine d'activité et vérifier ses propres interprétations avec des confrères.
29. Appliquer les principes de la théorie des tests et les principes d'interprétation des épreuves.
30. Résister aux pressions du milieu visant à trop écarter la planification, le diagnostic et les processus d'interprétation des tests
32. Considérer l'erreur standard de mesure.
33. Prendre en considération des conditions éveillant des doutes sur la validité de l'information à propos d'une situation particulière.
34. Voir si la raison pour faire passer un test correspond au but dans lequel le test a été créé.
41. Comprendre les scores standards et les rangs centiles.
42. Comprendre la validité de construit.
43. Comprendre la relation entre validité et fidélité.
45. Choisir un nombre suffisant de tests pour échantillonner les comportements, et ce afin d'en arriver à un objectif spécifique (comme l'évaluation neuropsychologique).
48. S'abstenir d'utiliser la version de recherche d'un test qui n'a pas de normes pour un groupe ou une personne qui ne parle pas français afin de prendre des décisions.
54. En s'appuyant sur une information valide, prendre en considération les éléments d'un test qui peuvent défavoriser certains groupes.
55. Éviter les erreurs au cours de l'évaluation et de l'enregistrement.
56. Vérifier les scores avec précision.
57. Faire de fréquentes vérifications durant l'évaluation afin de découvrir les erreurs.
58. Suivre les instructions concernant l'évaluation.
59. Partir de l'idée qu'une norme pour un travail ne s'applique pas à un travail différent (et que les normes pour un groupe ne s'appliquent pas automatiquement à d'autres groupes).
67. Choisir des tests appropriés à la fois à l'objectif de la mesure et aux personnes qui passent le test.
68. Choisir des tests qui sont le plus possible exempts de discrimination sociale par rapport à l'échantillon standardisé et à la population qui passe le test
78. Se référer au test comme base de l'interprétation uniquement lorsqu'on a fait passer et corrigé celui-ci dans le respect des règles et lorsque l'interprétation a été bien validée.
79. Éviter de se référer à un test comme base d'interprétation, même quand il est utilisé par un bon clinicien, sans tenir compte de la validité de l'interprétation, mais s'y référer seulement dans un cycle de formation et de vérification d'une hypothèse pour une entrevue clinique et une étude de cas.
80. Utiliser un test dans un cycle d'élaboration et de vérification d'une hypothèse dans le respect d'une bonne validité de l'interprétation.
87. S'abstenir de rapporter les scores sans faire une interprétation adéquate.

2.8. Codes, standards, directives

International Test Commission : La commission Internationale des Tests (ITC) définit des règles d'usage et de bonne conduite dans l'utilisation des tests (guidelines) : <http://www.intestcom.org/>

Les "**Standards for Educational and Psychological Testing**". Ces standards (2014) sont développés conjointement par l'APA ([American Psychological Association](#)) et l'AREA ([American Educational Research Association](#)). Une des "sections" de l'APA (la section American Psychological Assessment) précise aussi à quelles exigences techniques et professionnelles doivent répondre les tests et leurs utilisateurs. Ces éléments concernent les psychologues nord-américains mais ces recommandations sont aussi utiles pour ceux qui veulent approfondir leur réflexion dans ce domaine.

EFPA ([European Federation of Psychologists' Associations](#)). Cette association européenne s'est donné comme mission (entre autres) de définir des directives ("guidelines") et des principes à l'origine des codes de déontologie nationaux. Parmi les directives, on peut trouver des textes concernant la communication avec les médias, des recommandations pour l'enseignement de l'éthique, etc. Depuis 2007, EFPA a constitué un groupe de travail avec EAWOP (European Association of Work and Organizational Psychologists) dont l'objectif est de définir des standards pour l'usage mais aussi la constructions des tests.

Par ailleurs un texte révisé concernant la description et les critères d'évaluation des tests (validé en 2013 par l'EFPA) est disponible ici : [EFPA 2013 TEST REVIEW MODEL Version](#). Un psychologue devrait parcourir (voire plus) ce texte qui permet de rappeler les principales exigences attendues pour définir "une bonne épreuve" ou un "bon questionnaire".



Pour les curieux, voir les travaux de l'EFPA Board Assesment (document de 2023) -> [ICI](#)

3. Classification des tests

Il n'existe pas de classification unique des tests mentaux. Les tests mentaux peuvent se distinguer par :

- des caractéristiques formelles (tests papier-crayon, tests de performance),
- le mode de passation (individuel ou collectif),
- les caractéristiques de la population à laquelle ils s'adressent (enfants, adultes, baby-tests, etc.),
- les objectifs (on parle alors de classification fonctionnelle).

La classification la plus souvent utilisée est de type fonctionnel. On distingue les **tests cognitifs (ou tests d'efficience)** des **tests de personnalité** (ce qui permet d'effectuer une typologie grossière des tests mais elle ne couvre pas toutes les mesures).

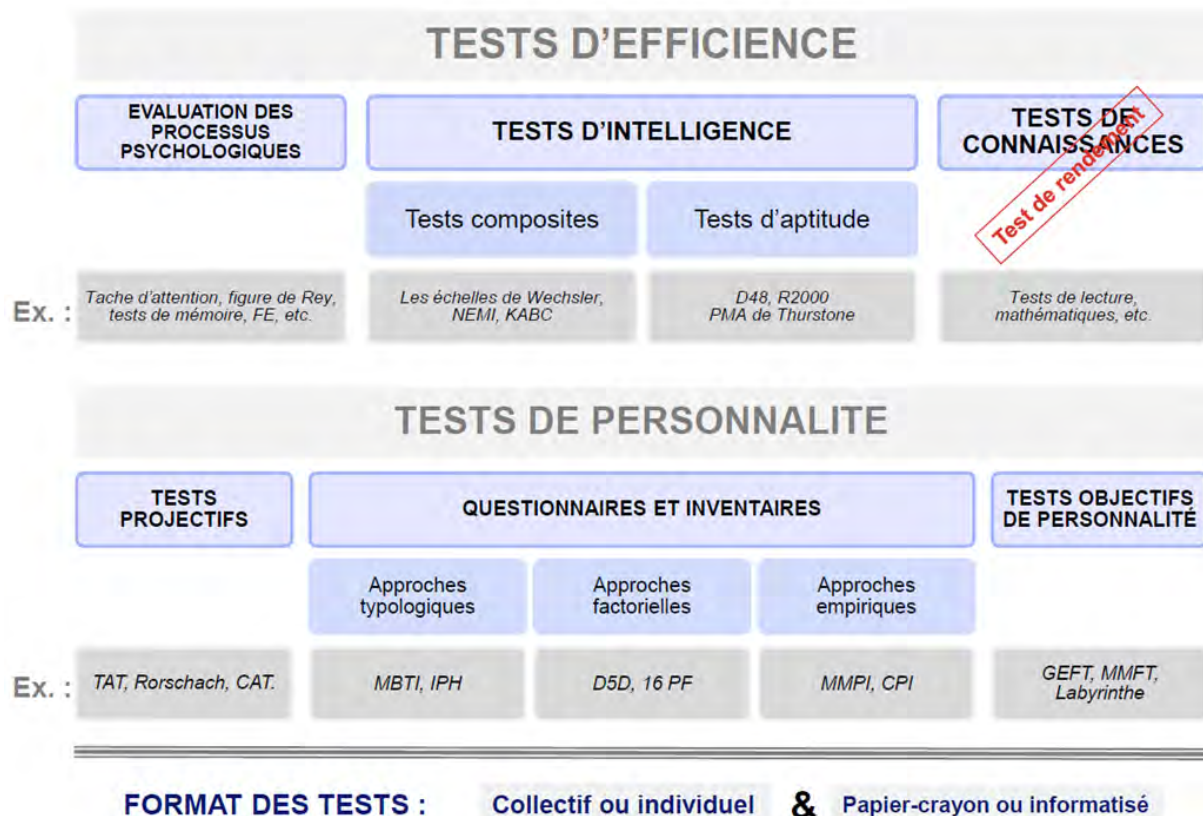


Figure B.3 : Exemple de classification des principaux tests mentaux

Remarques

- Une des différences essentielles entre test d'efficacité et tests de personnalité est que les premiers sont associés à des performances (il y a des bonnes et des mauvaises réponses, le temps d'exécution peut être pris en compte) alors que pour les seconds, il n'y a pas des bonnes et des mauvaises réponses. Par ailleurs, les tests de personnalité, contrairement aux tests d'efficacité sont le plus souvent sans contrainte temporelle.
- En Amérique du Nord (Canada), les tests de connaissances sont parfois classés dans une catégorie plus large (**test de rendement**) qui évalue "le niveau de connaissance ou d'habileté d'une personne dans un domaine particulier" (Hogan, 2017, p. 5). Dans cette catégorie on retrouve cependant des épreuves qui dépassent le champ classique de la psychologie et parfois s'éloignent largement de la définition habituelle des tests.
- Parmi les tests de personnalité (questionnaires et inventaires), on distingue aussi (ce qui n'est pas fait dans le schéma précédent) : (a) les tests d'intérêts ou de valeurs ; (b) les grands tests de personnalité. Les tests d'intérêts, de valeurs ou de motivation concernent surtout l'orientation et le monde du travail. Ils sont nombreux et ne respectent pas toujours les standards de construction des tests.
- La classification des éditeurs de tests ne recoupe pas obligatoirement cette classification. Il en est de même pour les ouvrages ou les sites proposant un recensement des tests (cf. par exemple, le site officiel de l'Institut Buros qui est spécialisé dans la publication d'analyses critiques de très nombreux tests (<https://marketplace.unl.edu/buros/test-reviews/categories>))

3.1. Tests d'efficacité

Les tests d'efficacité (ou cognitif) permettent d'évaluer la capacité à acquérir des connaissances, la

capacité à traiter certaines informations, des niveaux de connaissances, des aspects du fonctionnement cognitif. Habituellement on distingue :

- ❑ **Tests d'aptitudes** : permettent d'évaluer la capacité à acquérir des connaissances ou à traiter des informations dans des domaines particuliers (aptitudes : verbale, spatiale, numérique, etc.).
- ❑ **Les échelles composites d'intelligence** : évaluent un niveau de performance ou un niveau de développement global (les résultats sont exprimés le plus souvent en [quotient intellectuel](#) [QI] ou quotient de développement [QD])
- ❑ **Les tests de connaissances (achievement test)** : permettent d'évaluer les connaissances acquises (tests de connaissances scolaires et professionnelles). Comme nous l'avons mentionné précédemment, cette catégorie est incluse dans une catégorie plus large (dans les classifications nord-américaine) qui sont les "tests de rendement". Ces derniers englobent les grandes évaluations internationales (comme les enquêtes PISA), les tests de connaissances, les tests d'évaluation scolaire, etc. Ils s'écartent donc parfois de notre définition des tests (définition qui est plus restreinte).
- ❑ **Tests spécifiques (évaluation de processus) ou tests neuropsychologiques** : test d'attention-concentration, test de mémoire, etc.

Remarque : parmi les tests d'efficience on distingue aussi les tests de vitesse qui privilégient l'évaluation par le temps d'exécution des problèmes : les items sont le plus souvent simples. A l'inverse les tests de puissance n'ont pas de limite de temps et les items sont complexes.

3.2. Tests de personnalité

Les tests de personnalité permettent d'évaluer l'ensemble des aspects « non cognitifs » de la personnalité : les intérêts, les attitudes, les valeurs personnelles, et les traits de personnalité. Certaines des épreuves de personnalité n'ont parfois qu'à la marge les caractéristiques d'un test (standardisation, référence à une norme, fidélité, validité). On distingue actuellement :

- ❑ **Les tests objectifs** : dans une perspective classique, ce sont des épreuves cognitives perceptives ou motrices à partir desquelles on tire des indications sur le comportement. Elles sont dites objectives car elles ne laissent pas de place à la subjectivité du sujet dans sa réponse ni à celle de l'évaluateur dans la cotation (exemples : le labyrinthe de Porteus, anxiété et test du dessin en miroir).
Cette définition est différente de celle que l'on trouve en Amérique du nord. Elle est souvent reprise dans les traductions d'ouvrage sur la psychométrie. Dans cette tradition psychométrique, on ne retient que la seconde partie de la définition classique : les tests objectifs sont les tests de personnalité à cotation objective (exemples : Le test des Big Five [OCEAN], le NEO PI-R, etc.) . Ce sont donc, pour la majorité, des questionnaires et des échelles, cf. dessous.
- ❑ **Les tests projectifs** : la personne doit interpréter des images ambiguës (épreuves de Rorschach, TAT, etc.). Les travaux de validité concernant l'interprétation sont peu nombreux et l'[accord interjuge](#) est parfois faible.
- ❑ **Les questionnaires/échelles** : Le sujet doit répondre à de nombreuses questions (souvent par oui ou non). Cette méthode présente des inconvénients (tendance à l'acquiescement et désirabilité sociale) mais ces « biais » peuvent être en partie contrôlés. On distingue : les mesures de traits de personnalité ; les échelles d'intérêts ; les mesures d'attitudes ou de valeurs.

Remarques

- le terme de test projectif a été introduit par Lawrence K. Frank en 1939 pour signifier la proximité

entre différentes épreuves.

- Le même Lawrence K. Frank propose en 1948 une classification plus précise des tests de personnalité, classification souvent citée qui ne correspond cependant plus aux tests actuellement en développement. Il distinguait les techniques constitutives (organiser un matériel non structuré comme l'épreuve des tâches d'encre de Rorschach), les techniques constructives (la personne doit, à partir d'un matériel défini, construire des structures plus larges comme le test du monde de Bühler), les méthodes interprétatives (la personne doit interpréter une expérience ayant une signification affective, par exemple le Thematic Apperception Test ou test de Frustration de Rosenzweig) et enfin les techniques réfractives (la personnalité est appréhendée par la distorsion qu'il fait subir à un moyen de communication).

3.3. Le Quotient Intellectuel

Si le concept de test mental a été introduit en 1890 par le psychologue américain James McKeen Cattell, le premier test d'intelligence a été élaboré par [Alfred Binet](#) et [Théodore Simon](#) (Échelle métrique de l'intelligence) pour le ministère de l'Education nationale Français en 1906. L'objectif de ce test était de détecter les enfants en échec scolaire. La réussite aux différentes tâches proposées (items) était typique d'un âge de développement ce qui permettait, à partir des réponses de l'enfant, de calculer un âge mental. C'est [William Stern](#) qui introduit le terme et le calcul du QI. Depuis ces premiers travaux, les épreuves ont largement évoluées (en fonction des conceptions de l'intelligence des psychologues) et le terme de QI a pris des sens multiples. Les psychologues distinguent cependant assez classiquement le QI classique (QI de ratio) et le QI standard.

Pourquoi parler plus particulièrement du QI dans un texte concernant les tests et la mesure en général ? Pour une raison simple : le test de QI est souvent celui auquel on pense en premier lorsque l'on parle de test. C'est aussi le premier test psychométrique moderne (mais pas le premier test, c'est James McKeen Cattell qui avait parlé de tests mentaux). Par ailleurs, le terme de QI donne souvent lieu à des présentations erronées, dans la presse, les séries télévisées ou autres. Une précision sur cette "mesure" particulière est donc nécessaire. L'objet n'étant pas d'ouvrir un débat mais de rappeler ce dont on parle quand on parle de QI.

3.3.1 Le QI classique (QI de ratio)



[Alfred BINET](#), auteur de la première échelle d'intelligence, proposait de mesurer l'intelligence en terme d'âge mental (âge de développement) en utilisant comme indicateur des questions typiques d'un âge donné. La mesure de l'âge mental, par comparaisons à l'âge réel, était ainsi un indicateur d'avance ou de retard de développement et permettait.

C'est William STERN en 1912 qui introduit la notion de QI. Sa proposition était simple, le QI était simplement le rapport de l'âge mental (AM) sur l'âge réel (AR) multiplié par 100. C'est donc un pourcentage de "chemin parcouru" le long du "développement". Le QI classique (QI_c) est donc un cas particulier de ce qu'on appelle les quotients de développement (QD).

$$QI_c = \frac{AM}{AR} \times 100$$

Les Limites du QI_c (peu ou plus utilisé actuellement) :

→ Un AM est obtenu en additionnant des points de différents niveaux d'âge. La structure mentale d'un

âge mental donné est cependant probablement différente selon l'âge réel (par exemple un AM de 6 ans ne correspond probablement pas à la même structure mentale si l'AR est de 5 ans ou de 12 ans).

→ Cette méthode ne s'applique pas aux adultes. Comment mesurer l'intelligence après l'âge d'achèvement du développement ? Elle laisse alors une question en suspend, quand s'arrête le développement ?

→ Au QI classique, si la moyenne des QI est constante d'âge en âge (100) la distribution (et donc l'écart-type) des QI varie d'âge en âge. Un QI constant ne signifie pas que la personne reste classée (positionnée) de la même façon par rapport aux enfants du même âge.

3.3.2 QI standard



[David WECHSLER](#) va proposer d'abandonner le QIc et la notion d'âge mental pour construire un test d'efficacité intellectuelle. Ce nouveau QI sera obtenu à partir d'épreuves variées verbales ou non verbales (inspirées de l'alpha test et du beta test destinés aux jeunes recrues américaines). Le score sur chacune des épreuves (sous-tests) est transformé en score standard ([échelle normalisée](#) en 19 classes). Ces scores sont additionnés et le total est transformé en QI standard (QIs).

La transformation effectuée fait que la distribution reste normale. La moyenne des QIs est fixée à 100 et l'écart-type à 15. Par ailleurs 50% des sujets ont entre 90 et 110 (quartile de la distribution des QI). Cette distribution est la même quel que soit le groupe d'âge. Le QIs peut donc être considéré comme un étalonnage particulier.

Attention : Le QIs devient donc l'indicateur d'un rang et non plus indicateur d'un niveau de développement.

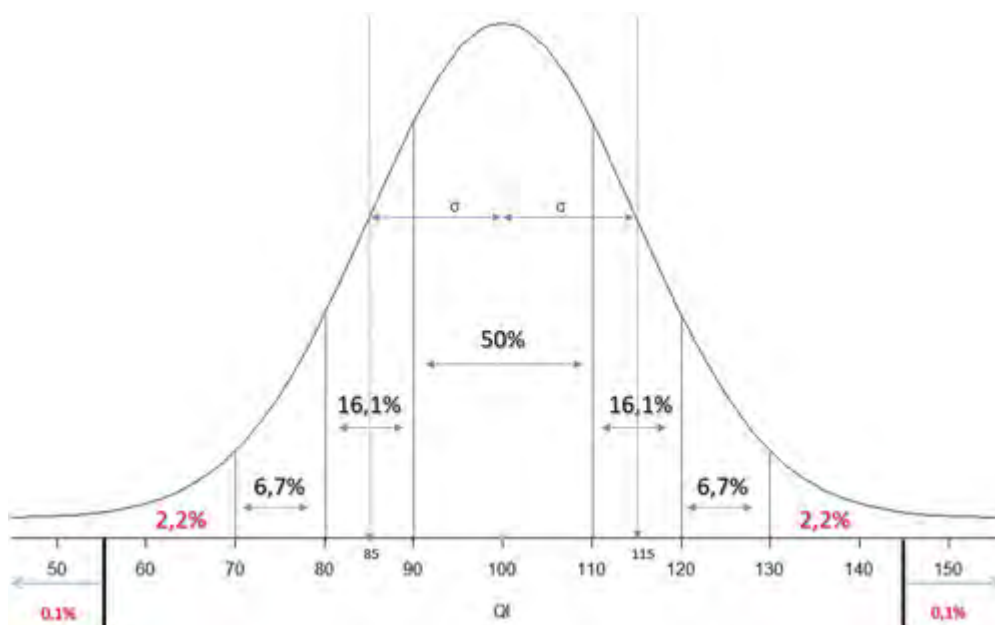


Figure B.4 : : Distribution des QI et fréquences entre différentes bornes

Avantage et inconvénient :

(+) Rend comparable les QIs de tous les âges, puisque le QIs positionne une personne par rapport à ses pairs du même âge. La distribution des QIs est normale et **la moyenne des QIs est toujours de 100 et**

l'écart-type de 15. Un QIs de 110 (un écart-probable au-dessus de la moyenne) signifie à tous les âges que 25% des personnes du même âge ont mieux réussi sur les tâches proposées. 68% des QIs se trouvent entre 85 et 115 et près de 95% entre 70 et 130.

(-) Le QI standard ne renseigne plus sur le retard ou l'avance de développement et n'est plus un quotient (le nom est totalement impropre).

Remarques :

- Attention aux confusions : avec David Wechsler, on a fixé a priori la moyenne du QI standard à 100 et l'écart-type à 15 lors de l'étalonnage des tests. Il existe cependant des tests postérieurs à ceux de Wechsler qui ont, pour le QI, un écart-type différent (16 ou 24). Par exemple le Culture Fair Intelligence Test (CFIT) composés de 5 sortes d'items non verbaux a un écart-type de 24 et non de 15 (on devrait utiliser le terme **QI Cattell** quand on en parle). Pour le Standford-Binet l'écart-type est de 16 (la différence est faible mais ce n'est plus un QI standard). **Donc la signification d'un QI dépend de la nature de l'échelle utilisée ! Il faut donc toujours préciser le type de QI utilisé** (QI classique, QI standard, QI Cattell, etc.).

Il existe des tables d'équivalence entre les QI des échelles de Wechsler (QI standard) et celle du CFIT (QI Cattell). Cette conversion est cependant facile à faire soi même avec une simple calculatrice ! On divise le (score - 100) par l'écart-type de l'échelle initiale et on multiplie par l'écart-type de la nouvelle échelle. La valeur obtenue est ajoutée à 100. Le tableau ci-dessous présente 3 exemples de conversion (avec détail de calcul) :

QI Standard (s=15)	QI Cattell (s=24)	QI stanford-Binet (s=16)
130	$100 + \frac{(130 - 100) * 24}{15} = 148$	$100 + \frac{(130 - 100) * 16}{15} = 132$
$100 + \frac{(76 - 100) * 15}{24} = 85$	76	$100 + \frac{(76 - 100) * 16}{24} = 84$
$100 + \frac{(108 - 100) * 15}{16} = 107.5$	$100 + \frac{(108 - 100) * 24}{16} = 112$	108

- Le QIs renseigne sur la position d'un individu dans un groupe de référence (population pour laquelle on a étalonné le test). Ce n'est pas une [échelle de rapport](#). En effet une personne avec un QIs de 150 n'est pas deux fois plus intelligente qu'une personne avec un QIs de 75. Ce n'est pas non plus une [échelle d'intervalle](#) (même si souvent on la traite ainsi). En effet une différence de 10 points entre un QI de 100 et 110 ou entre 130 et 140 ne correspond probablement pas à une "même différence d'intelligence". Toujours se rappeler que le QI positionne simplement un individu par rapport à un groupe (dans les épreuves présentés, la personne est plus efficace ou moins efficace que x% des personnes placées dans la même situation)..
- Les échelles de Wechsler (WPPSI, WISC et WAIS) sont probablement les échelles les plus utilisées au monde et les plus traduites. Pour ceux qui veulent connaître l'origine des sous-tests constituant ces échelles, on vous recommande l'article de [Boake](#) (1982).
- Le QI positionnant une personne par rapport à un groupe, il est relatif à ce groupe. Les conditions d'apprentissage et de développement évoluant régulièrement, il faut donc étalonner régulièrement (tous les 10 ans environ) les test de QI. En effet, on a observé pendant longtemps une augmentation des performances sur les épreuves utilisées pour mesurer le QI (les enfants actuels réussissant

mieux ces épreuves que les enfants du même âge des générations précédentes). Cet effet appelé effet Flynn (du nom de celui qui l'a mis en évidence) est plus marqué sur les épreuves en relation avec le facteur *g*. Depuis les années 1990-2000, cet effet semble cependant moins important voir pourrait s'inverser ([Dutton & Lynn, 2013](#)) dans certains pays. Il existe de multiples interprétations à cet effet ([Rindermann, Becker, & Coyle, 2017](#), [Bratsberg & Rogeberg 2018](#)).

4. Le code de déontologie

La déontologie est, de façon générale, l'ensemble des règles ou des devoirs régissant la conduite à tenir pour les membres d'une profession. Elle peut être cadrée par la loi ou non. Les plus connus sont les règles concernant les professions médicales (le serment d'Hippocrate) ou celles des journalistes (Charte de Munich).

En psychologie, les premiers codes de déontologie élaborés par les organisations professionnelles des psychologues datent des années 50. En Europe les codes s'inspirent le plus souvent de la charte européenne de déontologie votée le 5 novembre 1994 à Malte et du méta-code de la Fédération Européenne des Associations de Psychologie (EFPA) adopté le 1 juillet 1985 à Athènes. Un psychologue se doit donc de respecter le code de déontologie de son pays mais il se doit aussi, de connaître des règles de conduites comme celles proposées pour l'usage des tests par la commission internationale des tests ((ITC, International Test Commission (cf. [Code, Standards et directives](#))).

Exemples de code de déontologie (et protection du titre de psychologue. Situation au 26 octobre 2025)

France (code en vigueur : date 2021)

En France, la Société Française de Psychologie (SFP) publie son premier code de déontologie en 1961, le second code établi en collaboration avec d'autres organisations de psychologues a été adopté en 1996. En février 2012, les organisations de psychologues, associations et syndicats réunies au sein du GIRÉDÉP ont actualisé le Code de déontologie. Cette version a été à nouveau actualisée et adoptée en 2021 par 21 organisations, réunies dans le CERéDéPsy (Construire ensemble la réglementation de la déontologie des psychologues). Il faut savoir que depuis septembre 2024, les professions libérales réglementées doivent respecter des principes déontologiques susceptibles d'être sanctionnés légalement. Le code est téléchargeable sur de nombreux sites (site des associations de psychologues). Il existe aussi un site spécifique pour le code de déontologie : <http://www.codededeontologiedespsychologues.fr> , site sur où vous pourrez trouver aussi les anciennes version du code de déontologie et les signataires.

Rappel : en France, la psychologie est une profession réglementée (son exercice n'est pas libre, [article 44 de la loi n°85-772 du 25 juillet 1985 complétée par des ordonnances en 2002, 2005 et 2010](#)). Pour faire usage du titre de psychologues, ceux-ci ont obligation d'avoir un numéro ADELI (qui signifie Automatisation DES Listes) donc d'être référencés dans le système d'information national concernant aussi les professionnels relevant du code de la santé publique et du code de l'action sociale. Ce numéro doit être indiqué lors de la réalisation d'un bilan avec les coordonnées du psychologue.

Suisse (code en vigueur 2024)

Le code de déontologie que l'on trouve sur le site de la [fédération suisse des psychologues](#) (FSP) se base comme la plupart des codes européens sur le métacode de l'European Federation of Psychologists Association (EFPA). Il est rédigé en 3 langues : https://www.psychologie.ch/storage/images/6181/Berufsordnung-D_F_I_E_Stand-2024_08_01.pdf.

Rappel : la pratique professionnelle est aussi réglementée en Suisse. La [loi LPsy du 18 mars 2011](#) (entrée en vigueur le 1 mars 2013) protège l'utilisation du titre de psychologue. Seules les personnes possédant un diplôme de fin d'études reconnu d'une université (un Master en psychologie) ont le droit de porter le titre de psychologue (art. 2 et art. 4 LPsy). Ces dispositions s'appliquent aussi aux noms composés (psychologue du sport, psychologue de l'éducation, etc.). Cette loi définit aussi les conditions régissant l'octroi des titres fédéraux de formations postgrade et du droit de pratique au niveau cantonal.

Belgique (code en vigueur 2014, amendé en 2018)

Le code de déontologie des psychologues de Belgique est entré en vigueur le 26 mai 2014 (amendé en 2018). Le code de déontologie est "contraignant" (i.e. on doit le respecter) pour les psychologues et impose un respect absolu des personnes rencontrées dans le cadre de leur activité professionnelle. Une de ces caractéristiques est qu'il permet de se positionner dans le contexte de certaines situations engageant sa responsabilité (exemple : quand peut-on ou doit-on avoir une levée partielle du secret professionnel). Vous pouvez le consulter ici : <https://www.compsy.be/le-code-de-deontologie>

En Belgique, la loi du 8 novembre 1993 protège le titre de psychologue. La loi du 21 décembre 2013 modifiant la loi de 1993 précise par ailleurs que le titre de psychologue offre une garantie non seulement de compétence professionnelle, mais également d'engagement éthique. Ce droit se matérialise à travers deux conseils créés au sein de la Commission des Psychologues. Ces commissions veillent au respect du code de déontologie du psychologue et interviennent en cas d'infraction (pour plus de détail : <https://www.compsy.be/fr>). Pour porter le titre de psychologue, il faut être titulaire d'un master et d'une licence de psychologie et être inscrit sur la liste officielle des psychologues. Plus de détail sur le site de la Fédération Belge des psychologues (<https://www.bfp-fbp.be/fr>)

Luxembourg (code en vigueur, 2001)

La Société Luxembourgeoise de Psychologie (SLP) a adopté en 2001 un code de déontologie destiné à servir de référence aux personnes qui exercent la profession de psychologue. C'est une adaptation de différents codes européens. Il est disponible sur le site de la SLP (<https://www.slp.lu/fr/ethique/>). L'adhésion à la SLP implique l'engagement aux critères éthiques et le respect du code de déontologie de la profession.

Rappel : la psychologie ne fait pas partie des professions réglementées au Luxembourg (un projet de réglementation existe dans le Plan national santé mentale (PNSM), approuvé en juillet 2023, mais il n'y a toujours pas de loi à ce sujet en 2025). La chambre des députés a cependant voté en 2015 une loi portant création de la profession de psychothérapeute (qui devient une profession réglementée). Pour exercer en qualité de psychothérapeute, il faut être titulaire d'un diplôme en psychologie clinique ou de médecin mais aussi (en plus) d'un diplôme de psychothérapeute (pour plus de détails : <https://www.slp.lu/fr>).

Canada (code en vigueur, 2017)

La quatrième édition du code d'éthique des psychologues (<https://www.cpa.ca/aproposdelascp/comites/ethics/codeofethics>) est un cadre général pour les psychologues canadiens (le premier code datait de 1986). On notera que dans cette dernière édition le nom du premier principe du code devient « *Respect de la dignité des personnes et des peuples* » (page 12). Par ailleurs il est clairement affirmé la responsabilité du psychologue qui se doit d'avoir une démarche permettant le développement continu de ses connaissances sur l'éthique et de ses

compétences en matière de prise de décisions éthiques.

Rappel : les psychologues doivent posséder un "permis" (licence en anglais) pour exercer la psychologie au Canada. L'autorisation d'exercer est accordée par les organismes réglementaires de chaque administration canadienne. Les exigences relatives à l'autorisation d'exercer varient d'une administration à une autre. Dans certaines provinces et certains territoires, le doctorat professionnel est exigé pour pouvoir s'enregistrer et dans d'autres c'est la maîtrise. Les psychologues titulaires d'un doctorat peuvent utiliser le titre « Dr ». Vous trouverez plus d'information sur le site de la société canadienne de psychologie (<https://www.cpa.ca>).

C - Construction d'un test et création des items

La construction des tests est un processus long qui ne se termine pas nécessairement au moment de sa publication puisque la [validation](#) de l'épreuve peut donner lieu à de nombreux travaux complémentaires permettant de préciser les propriétés de la mesure. Dans un manuel de test on doit retrouver tous les éléments qui vont permettre au praticien d'évaluer la qualité de l'outil qu'il souhaite utiliser : objectif du test, sa forme finale, ses qualités métrologiques, l'étalonnage et la date de publication (cf. chap C §1.3 - [Manuel des tests](#)). Ces éléments sont, pour la plupart, ceux que l'on retrouve lors de la construction d'un test.

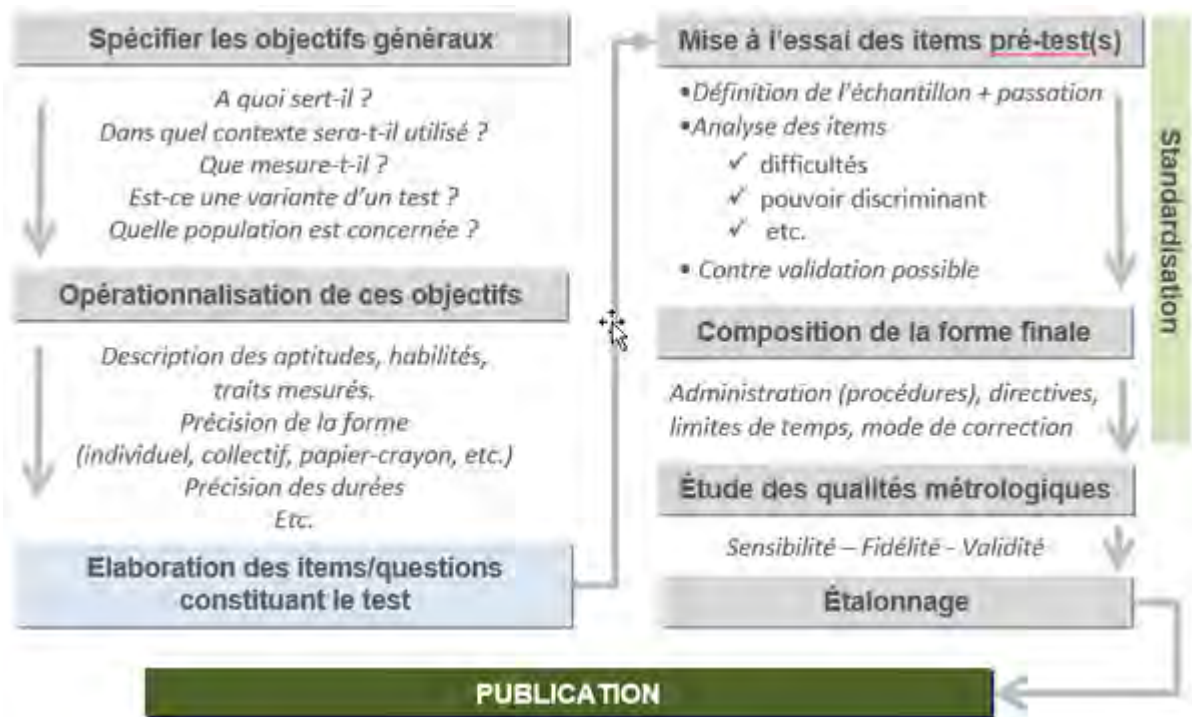


Figure C.1 : : Résumé des principales phases de construction des tests

1. Théorie classique des tests

La théorie classique des tests (TCT) est ce qu'on peut appeler la "théorie du score vrai". L'hypothèse initiale de la TCT (premier postulat, Spearman, 1904b) est qu'il est impossible d'obtenir une mesure complètement exempte d'erreurs. Cet aspect est illustré par l'équation :

$$X = T + \varepsilon$$

où X correspond au score observé à un test ;

T est le vrai score du sujet (inconnu) ;

ε est l'[erreur aléatoire \(non-systématique\)](#) qui s'ajoute au vrai score du sujet et

ε suit une loi gaussienne de moyenne 0 et d'écart-type σ .

Cette équation simple signifie que le score observé à un test n'est pas nécessairement le score vrai. Les scores qui s'éloignent du score observé sont cependant moins probables et plus on s'éloigne de ce score vrai, plus cette probabilité est faible. La variable X est une variable aléatoire qui se distribue normalement et l'écart-type de la variable aléatoire X est lié à l'erreur de mesure ε . Cet écart-type correspond à l'[erreur standard de mesure](#). Cette hypothèse concernant la relation entre score vrai et

score observé permet de calculer des [intervalles de confiance](#) dont l'importance (la taille) dépend de cette erreur standard de mesure. Plus l'erreur (ε) est faible, plus la mesure sera précise (= plus le score observé a de très forte chance d'être proche du score vrai).

Cette conception est à la base de la construction des tests. Il est important de noter :

- qu'un des problèmes les plus importants ou les plus connus de la TCT, concerne le fait que les caractéristiques de [l'échantillon de standardisation](#) et les caractéristiques du test ne peuvent pas être séparées. Les caractéristiques des items et des tests peuvent changer selon l'échantillon étudié. Il est par conséquent difficile de comparer des items dont les caractéristiques sont obtenues en utilisant des groupes différents de sujets.
- que l'erreur standard dans le cadre de la TCT (important dans le cadre de l'évaluation de la [fidélité](#)) est supposée la même pour tous les participants quel que soit leur position sur la dimension examinée. Cette hypothèse est hautement improbable.
- que la TCT est "*test oriented*" et qu'il n'est pas possible de faire des prédictions sur la façon dont une personne ou un groupe réussit un item particulier du test. En d'autre terme, la TCT ne nous permet pas de savoir quelle est la probabilité pour un individu particulier de répondre correctement à un item donné.

2. Elaboration des items d'un test

Un test ou les différentes épreuves d'un test sont 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 d'un test 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 cette réflexion on doit aussi intégrer les effets de contexte (d'une question sur la suivante), le temps de passation maximum souhaité, etc. Ce choix est complexe et il existe des études qui portent sur ces aspects. Il n'est pas possible de tout aborder dans le cadre d'un cours ou d'un manuel et nous présentons uniquement quelques éléments importants soulignant la complexité de la construction et de la sélection des items d'un test.

2.1. 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 rarement utilisé car il pose des problèmes de standardisation de la notation.
- **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 (ces questions sont parfois appelées questions de performances mais ce sont des questions ouvertes avec une réponse juste).
- **Les questions "VRAI - FAUX"** avec une réponse à fournir parmi deux. On peut distinguer deux modes 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 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 (prémises ou amorces). Dans ces épreuves les fausses réponses sont appelées « les distracteurs » ou les "leurres".

(*) *Les premiers tests QCM apparaissent en 1914-1915. On considère que F.J. Kelly en est l'inventeur. Professeur à l'Institut de formation des enseignants du Kansas, il crée le premier QCM de lecture silencieuse chronométré (Kansas Silent Reading Test) pour le recrutement en masse d'immigrants aux USA dans les entreprises après la guerre.*

- **Les questions d'appariement** : ce sont des questions proches du format à choix multiples dans lesquelles on demande de mettre en relation (appairer) des énoncés (phrases, mots, expressions) qui sont souvent présentés sur 2 colonnes. Exemple :

BINET <input type="radio"/>	<input type="radio"/> QI
STERN <input type="radio"/>	<input type="radio"/> TEST MENTAL
CATTELL <input type="radio"/>	<input type="radio"/> AGE MENTAL
SPEARMAN <input type="radio"/>	<input type="radio"/> 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

On distingue habituellement les **échelles paires** dites à choix forcés (comme l'exemple précédent) des **échelles impaires** dans lesquelles le niveau central permet de n'exprimer aucun avis.

Remarque : Cette classification ne recouvre pas tous les formats existants et l'inventivité dans ce domaine reste ouverte ! Il existe d'autres façons de classer les formes à corrections objectives (cf. par exemple Laveault et Grégoire, 2014, p.30).

2.2. Difficulté et validité des questions

Les VRAI-FAUX sont plus complexes à élaborer que ce que l'on pense habituellement. La difficulté et la validité des questions vont dépendre de la formulation de la question. Par exemple, dans les questionnaires oui/non portant sur l'acquisition de connaissances la formulation des questions (négatives, avec ou sans connecteur logique, etc.) permet d'augmenter la difficulté des questions mais le risque est alors d'évaluer non plus seulement des connaissances, mais l'activité de raisonnement (ce qui peut entraîner une perte de validité de la mesure).

Le problème paraît plus simple avec les questionnaires à choix multiples, mais il faut savoir que la difficulté de la réponse dépend alors fortement des distracteurs et pas uniquement de la connaissance de la réponse. Par exemple à la question "quel est le poids d'un électron ?" peu de gens peuvent répondre. La même question évaluée par le choix multiple (A) suivant donnera probablement 100% de réponses justes et ce ne sera pas le cas dans la formulation (B).

(A) Le poids d'un électron est de : 1000g $9,1 \cdot 10^{-28}g$ 500g 1 g

(B) Le poids d'un électron est de : $9,1 \cdot 10^{-24}g$ $9,1 \cdot 10^{-26}g$ $9,1 \cdot 10^{-28}g$ $9,1 \cdot 10^{-30}g$

En fait le choix multiple permet dans la formulation (A) d'évaluer grossièrement la connaissance de l'ordre de grandeur du poids d'un électron. En proposant des valeurs plus proches de la réponse attendue (formulation B), on évalue une connaissance plus précise de cet ordre de grandeur.

Exemples d'items d'un des premiers tests à choix multiples : Alpha Test, Yerkes, 1917

- Les thermomètres sont utiles parce que :
 - Ils régulent la température
 - Ils indiquent la température
 - Ils contiennent du mercure
- Chaussure – pied ; chapeau – ?
 - chaton
 - tête
 - couteau

Exemples d'items évaluant des connaissances

- Binet a été le premier à introduire :
 - La notion de QI
 - La mesure de l'âge mental
 - La notion d'âge réel
 - Les tests mentaux
- Qui serait à l'initiative des "QCM" ? :
 - J. Watson

- B.F. Skinner
- F. J. Kelly
- E. Thorndike
- A. Binet

Réponses : 2 ; 2 ; 2 ; 3

2.3. Cotation des QCM et des VF

Une des caractéristiques des questionnaires de type VRAI-FAUX ou à choix multiples (QCM) est que l'on peut répondre juste par hasard. Par exemple pour une question avec 4 réponses alternatives dont une seule correcte, la probabilité de répondre juste au hasard est de 25%.

Par convention un score brut (non corrigé par les erreurs) est égal au nombre de réponses correctes dans une question (r_i) divisé par le nombre de bonnes réponses possibles à cette même question. Ce chiffre constitue une borne supérieure (si on fait l'hypothèse qu'il n'est fait aucune réponse au hasard, même les fausses). Par contre, le score corrigé totalement pour tenir compte du hasard constitue une valeur inférieure où l'on suppose que toutes les réponses sont faites au hasard. La réalité situe probablement entre ces deux bornes mais on ignore exactement où. C'est pourquoi, il existe plusieurs systèmes de correction possible en fonction de la façon dont on va positionner notre curseur (qui sera plus ou moins sévère).

Cas des choix multiples (lorsque la personne ne connaît pas le nombre de bonnes réponses).

Il existe de nombreuses façons de calculer le score selon que l'on privilégie, lors de la cotation, de valoriser la prudence, l'exactitude ou de pénaliser les stratégies aléatoires. Il n'est pas possible de présenter toutes ces méthodes, mais pour les choix multiples en voici 2 :

Pénaliser les mauvaises réponses proportionnellement au nombre de bonnes réponses possibles: cette cotation, sans divisé par $k_i - d_i$ est très sévères et est souvent utilisé dans les concours (décourage le hasard par une espérance mathématique négative). Dans la dernière version qu Quiz de Scalp on l'utilisait mais on normait le score en divisant par $k_i - d_i$. On atténuait ainsi cette pénalisation associée à des réponses aléatoires.

$$(1) \quad x_i = \begin{cases} -1 & \text{si } r_i - e_i \geq k_i - d_i, \\ 0 & \text{si } r_i + e_i = k_i, \\ \frac{2 \cdot r_i - e_i}{k_i - d_i} & \text{sinon.} \end{cases}$$

avec pour la question i :

- r_i le nombre de réponses correctes données
- e_i le nombre de réponses fausses données (choix erronés)
- d_i le nombre de distracteurs (non réponses attendues)
- k_i le nombre de choix possibles (corrects ou incorrects).

Scores normalisés : Cette méthode, plus simple, est utilisée en général lors d'évaluation (dont le Quiz associé à ce cours). Il correspond à la proportion de bonne(s) réponse(s) parmi les cases à cochers auquel on retire la proportion de mauvaise(s) réponse(s) parmi les distracteurs. Le score se situe dans l'intervalle [-1 ; +1]. Il est de 0 si tout coché ou si aucune case n'est cochée. L'espérance pour ce score est nulle en cas de réponses choisies au hasard. Ceci assure que seule la connaissance réelle génère un score positif. Le score supprime les biais de stratégie (que les réponses soit faites avec peu de certitude ou nécessitant beaucoup de certitude = stratégie audacieuse vs prudente).

$$(2) \quad x_i = \frac{r_i}{(k_i - d_i)} - \frac{e_i}{d_i}$$

avec pour la question i :

- r_i le nombre de réponses correctes données
- e_i le nombre de réponses fausses données

(choix erronés)

d_i le nombre de distracteurs (non réponses attendues)

k_i le nombre de choix possibles (corrects ou incorrects).

Dans le cadre d'un choix forcé on peut distinguer 2 situations

1 choix possible parmi n réponses possibles (donc $d_i = n_i - 1$). La cotation est simple. On attribue la note de 1 quand la réponse est correcte, 0 en cas de non réponse et une pénalité de $-1/d_i$ en cas d'erreur (cela correspond à la formule 2 précédente).

le nombre de bonne réponse est connu et est supérieur à 1. Il existe à nouveau plusieurs méthodes de cotation. On peut utiliser la proposition (1) mais elle est très sévère. On préfère alors pénaliser les fausses réponses en fonction du nombre de distracteurs ce qui conduit à nouveau à la formule 2.

Pour ces systèmes, le score d'une personne à l'ensemble du questionnaire sera la somme (pondérée éventuellement) des scores à chaque question (x_i).

Remarques :

- Pour un choix multiple, on suppose toujours qu'il existe au moins une réponse à donner.
- Ces systèmes sont souvent mal compris mais on doit donc avoir un système de cotation qui tient compte des réponses au hasard. Si on ne prend pas en compte les erreurs dans la cotation, une stratégie consisterait, dans les choix multiples, de tout cocher !
- Pour les VRAI-FAUX, cas particulier du choix forcé connu, ($n=2$) la règle pour compenser le hasard conduit à mettre 1 pour une réponse correcte, -1 pour une réponse incorrecte et 0 pour une non réponse. Si on ne l'utilise pas, il est facile pour celui qui répond d'optimiser son score.
- On peut utiliser d'autres règles pour pénaliser les erreurs et tenir compte du hasard. Celles présentées ci-dessus sont des règles habituelles (adoptées pour certaines dans le [SCALP-Quiz](#) sur la psychométrie).

2.4. Les biais de réponses

De façon générale on parle de biais de réponse lorsque la réponse à un item a tendance à être déterminée par des éléments externes à ce que l'item (la question) est censé mesurer. Dans l'élaboration d'un test, et plus particulièrement les questions constituant l'épreuve, on se doit de prendre en compte ces biais possibles de réponse, biais qui varient selon la nature et le format de la question. On peut mentionner :

- La réponse non sincère (effet de désirabilité sociale par exemple ou tendance à l'acquiescement) dans les questionnaires où l'on oblige à exprimer un avis.
- La tendance à l'indécision qui peut conduire dans les [échelles de Likert](#) impairs à choisir la catégorie centrale.
- La tendance à privilégier la vitesse sur l'exactitude (effet moins fréquent sur les personnes les plus âgées. Il peut même s'inverser).
- La réponse au hasard (normalement pris en compte dans [la cotation de ces items](#)).

Le choix des items et de la consigne accompagnant l'épreuve sont essentiels pour minimiser la plupart de ces biais de réponses. Pour contrôler ces effets, il existe des méthodes permettant de mesurer la

sincérité, l'effet de désirabilité sociale, la tendance à l'indécision, etc. Ces méthodes consistent le plus souvent à regarder la cohérence des réponses à certaines questions, les fréquences des réponses, etc.

Pour certains tests publiés (le plus souvent les questionnaires très sensibles à ces biais) des indicateurs supplémentaires permettent au psychologue de déterminer, lors de l'utilisation du test, si le protocole recueilli peut-être pris en compte ou non. Par exemple, pour contrôler la tendance à l'indécision on peut regarder la fréquence des réponses intermédiaires d'une personne par rapport à l'échantillon de standardisation. On peut aussi calculer des scores de cohérence pour contrôler les réponses au hasard (toujours pour les questionnaires).

3. 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 des psychologues à l'origine de l'épreuve jouent un rôle important. Ensuite, lors de la sélection des items, les définitions implicites et/ou explicites de la dimension que l'on souhaite « mesurer » peuvent contribuer ou non à l'élimination d'un item mais la sélection des items repose essentiellement sur une analyse plus technique (étude des propriétés de chaque question en fonction des objectifs).

Pour rappel, plusieurs étapes sont donc nécessaires pour la construction d'une épreuve. L'élaboration d'un premier ensemble d'items constitue une version provisoire du test qui est administrée à un échantillon de personnes. 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 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, etc.). Pour les épreuves d'évaluation de « performances » cognitives, on prend en compte la difficulté des items et leur discriminabilité. Pour les épreuves de personnalité ou des tests qui ne sont pas des épreuves de performances, c'est l'[homogénéité](#) interne ou la structure interne (en lien avec [la validation](#)) de l'épreuve qui sera aussi pris en compte. On peut aussi appliquer des techniques qui ne feront pas référence à la [théorie classique des test](#) (TCT) mais aux [modèles de réponses à l'item](#).

Il n'est pas possible de présenter toutes ces méthodes et l'objectif est de comprendre et d'illustrer les différents critères pouvant être pris en compte lors de la sélection d'item. Cette présentation qui est une sensibilisation est donc partielle. Nous illustrerons les principes utilisés en nous concentrant sur quelques exemples d'indicateurs classiques pouvant être utilisés avec des échelles de performance (mais pas uniquement). Le principe général de sélection reste similaire lorsque l'on utilise d'autres techniques.

3.1. Indice de puissance (p-index)

Les items doivent le plus possible différencier les personnes qui passeront le test. L'épreuve doit être [sensible](#). Le premier indicateur pour s'assurer de cette sensibilité est l'indice de difficulté ou p-index (power en anglais). En français on parle indice de puissance. Il s'applique aux items pouvant être réussis ou échoués. Le 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 ($p \times 100$ 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).

Utilisation du p-index lors de la sélection des items.

Si l'on ne prend que des items d'indice p élevé (items faciles) l'épreuve ne permettra de différencier que les sujets les plus en difficulté (les autres réussiront). A l'inverse si l'on ne prend que des items difficiles (à l'indice p trop faible), l'épreuve trop difficile ne discriminerait que les très bons (les autres échoueraient à tous les items). 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. On fait l'hypothèse que la majorité des personnes se trouvent dans une zone centrale (on maximise pour 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.

Remarques

- Cet indice dépend directement de l'échantillon. Si les personnes sont plus performantes que la moyenne, les items vont avoir des p_index plus élevés (plus faciles) et l'épreuve sera trop difficile dans la population générale (donc peu sensible pour différencier les personnes). Inversement si l'échantillon est constitué de personnes peu performantes sur cette épreuve.
- Cette dépendance à l'échantillon lors de la sélection des items est forte et nécessite donc d'effectuer un échantillonnage correct. Il existe un modèle d'analyse des items qui permet de dépasser ce problème ([les modèles de réponses à l'item](#)) et qui rend les critères de sélection des items indépendants de l'échantillon (nous reviendront sur cet aspect).

3.2. Indices de discrimination

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.

Il existe plusieurs indices permettant d'évaluer ce pouvoir discriminant. Nous ne présenterons que 2 indices de discriminations : l'indice de Findley (d-index) et le coefficient de corrélation biserial de point (ou point-biserial). Le choix de l'indicateur dépend essentiellement de la nature des questions (items dichotomiques ou non).

3.2.1 Indice de Findley (d-index)

Principe. Les compétences réelles des personnes étant inconnues lors de l'élaboration d'un test on détermine l'indice de discriminabilité d'un item (d-index) de la façon suivante :

dans l'échantillon, on construit deux groupes de personnes : les performants (le sous-groupe des 27% de personnes ayant les scores les plus élevés sur l'ensemble de l'épreuve) et les « peu performantes » (le sous-groupe des 27% de personnes ayant les scores les moins élevés sur l'ensemble de l'épreuve).

Pour chaque item, et dans chacun des sous-groupes on calcule un p-index.

On calcule le d-index qui est la différence entre les p-index de ces deux sous-groupes (p-

index des "plus performants" moins le p-index des "peu performants".

Plus l'indice (le d-index) est proche de 1, plus l'item discrimine les personnes. Plus cet indice est proche de 0, moins il différencie (dans ce cas on ne retient pas l'item). Si l'indice est négatif cela signifie que les "plus performants" dans l'échantillon réussissent moins cet item que les "peu performants". Un item ayant un d-index faible ou négatif nuit donc à l'[homogénéité](#) de l'épreuve et on l'élimine.

Exemple. Soit l'administration d'un test composé de 9 items. On calcule le score total au test (nombre de bonnes réponses). Pour chaque item on calcule la proportion de sujets ayant réussi l'item parmi les 27% des meilleurs score au test (p_1) et la proportion de sujets ayant réussi parmi les 27% les moins efficaces au test (p_2). Remarque : p_1 et p_2 sont des indices de difficultés pour chaque sous groupe (cf. tableau ci-dessous). Le d-index est simplement la différence entre p_1 et p_2 .

Items	I1	I2	I3	I4	I5	I6	I7	I8	I9
p_1	0.85	0.92	0.99	0.55	0.30	0.37	0.86	0.92	0.60
p_2	0.10	0.08	0.02	0.22	0.32	0.35	0.10	0.05	0.12
d-index	0.75	0.84	0.97	0.33	-.02	0.02	0.76	0.87	0.48

I5, I6 et dans une moindre mesure I4 nuisent à l'homogénéité de l'instrument.

Attention : Si potentiellement, le d-index peut varier entre -1 et +1, sa valeur maximale pour un item i est directement dépendant du [p-index](#):

$$\max(d_i) = \begin{cases} p_i/.27 & \text{si } p_i < .27 \\ 1 & \text{si } .27 \leq p_i \leq .73 \\ (1 - p_i)/.27 & \text{si } p_i > .73 \end{cases}$$

avec : p_i le p-index de l'item i

Cette formule signifie qu'un item très difficile (p-index inférieur ou très inférieur à .27) ne peut avoir qu'un d-index peu élevé. En effet si 5% des personnes réussissent un item et que ces 5% sont réellement les meilleurs, p_1 sera égal à 5/27 donc 0.185 et p_2 sera égal à 0 (si on prend la notation du tableau ci-dessus). Le d-index sera alors de 0.185 (ce qui correspond bien à la valeur maximale possible donnée par la formule soit $0.05 * 0.27$). Le d-index paraît donc peu élevé mais on doit tenir compte que c'est le maximum possible au vu de la valeur du p-index global.

De la même façon, si un item est très facile (p-index supérieur ou très supérieur à .73), le d-index ne peut être que faible. On doit donc tenir compte du p-index pour calculer le d-index maximum possible et comparer avec le d-index observé (pour chaque item).

Pour aller plus loin...

Une variante de l'indice de Findley est l'indice B de "discrimination au seuil de maîtrise" (Brennan, 1972) que l'on peut utiliser pour des tests de connaissances (le plus souvent des tests scolaires). Pour calculer cet indice, on se fixe un seuil (niveau de maîtrise du contenu, par exemple réussite à 80% des items) et on constitue 2 groupes, ceux qui réussissent à 80% ou plus et tous les autres. Cet indice se calcule alors comme l'indice de Findley (indice de difficulté pour ceux qui maîtrisent moins indice de difficulté calculé sur ceux qui ne maîtrisent pas). A nouveau cet indice dépend du niveau de difficulté des items et du seuil de maîtrise que l'on se fixe.

3.2.2 Coefficient de corrélation biserial de point

Une autre façon d'envisager la relation entre le score à un item et le score au test consiste à utiliser le coefficient de [corrélation biserial de point](#) (r_{pbis}) :

$$r_{pbis}(i) = \frac{m_1 - m_0}{\sigma} * \sqrt{p(1-p)}$$

avec

m_1 : la moyenne observée à l'épreuve pour ceux qui ont réussi l'item i

m_0 : la moyenne observée à l'épreuve pour ceux qui ont échoué l'item i

σ : l'écart-type des scores

p : la proportion des personnes ayant réussi l'item i

Remarques :

- En général, pour étudier les items on utilise un **coefficient corrigé** en utilisant un score total calculé sans tenir compte de l'item évalué.
- Il s'agit d'une corrélation item-test. Un item discrimine correctement un test si il existe une corrélation positive entre le score à l'item et le score au test. Ce coefficient peut varier (comme le coefficient de corrélation de Pearson) entre -1 et +1. Il faut savoir cependant que ces valeurs maximales ne peuvent être observées que si la proportion des personnes ayant réussi (p) l'item est de .50.
- La formule du coefficient de corrélation biserial de point est simplement une formule qui permet de simplifier les calculs. Mais on peut appliquer la formule de Bravais Pearson (avec des 1 et 0 pour la variable dichotomique) et on obtient le même résultat.
- L'utilisation du coefficient biserial de point est celui qui est le plus souvent utilisé car les items sont souvent dichotomiques. D'autres coefficients de corrélations peuvent cependant être utilisés.

3.3.1 Principes généraux

La sélection ou l'élimination des items constituant une épreuve est une étape importante. Elle participe à la [fidélité](#) et à la [validité](#) de l'épreuve. La méthode dépend de la nature des items mais aussi des objectifs que l'on se fixe.

Le principe général le plus souvent utilisé consiste à sélectionner les items en prenant en compte leur difficulté et leur capacité à discriminer (ces deux indices ne sont cependant pas totalement indépendants). Par exemple, pour un test d'efficience, on cherche à retenir les items les plus différents possibles (i.e. avec des p -index différents pour maximiser la sensibilité) mais qui ont aussi un score d -index le plus élevé possible.

Cette sélection doit prendre en considération aussi d'autres aspects et pas seulement ces indicateurs. Par exemple, dans des tâches cognitives, des items très "faciles" (ayant indice de puissance élevée) seront conservés même s'ils apportent peu d'information car ils peuvent permettre de mettre en confiance. Dans une épreuve de personnalité, on peut avoir des items que l'on garde mais non pris en compte dans le score total (items de remplissage).

Remarques :

Il existe bien d'autres méthodes de sélection des items et l'élimination d'un item peu intervenir à plusieurs moments lors de la construction d'un test. Concernant l'analyse et la sélection des items, vous pouvez vous référer à l'ouvrage de Lavaut et Grégoire (2016, p. 203-

239).

Parmi les autres critères de sélection que nous évoquons :

Il est possible aussi de prendre en compte la corrélation entre le score à l'item et le score total à l'épreuve multiplié par l'écart-type des réponses à cet item. Cet indice reflète à la fois la cohérence de l'item avec le test et la variabilité des réponses. Plus cet indice est élevé, plus l'item est informatif et discriminant. La variance totale du test correspond au carré de la somme de ces indices. Choisir des items avec des indices élevés permet donc de maximiser la variance et la qualité psychométrique du test.

il faut savoir que souvent on tient compte aussi de la contribution de l'item à la dimension mesurée en effectuant une [analyse factorielle](#).

3.3.2 Exemple avec prise en compte du d-index et p-index

Comme nous l'avons mentionné en introduction de ce chapitre, il n'est pas possible de tout présenter et nous avons choisi simplement d'illustrer ce qui est un "bon item" en montrant comment dans la construction d'une échelle de performance on utilise conjointement le [p-index](#) et le [d-index](#).

En principe, les bons items sont ceux dont l'index de discrimination est élevé. Cette valeur est cependant contrainte par la difficulté de l'item et pour interpréter cet indice on doit tenir compte de cette contrainte.

Par exemple, admettons que sur 100 personnes, 10 réussissent l'item (p-index = 0.1). Si ceux qui réussissent sont tous des sujets appartenant au groupe des 27% qui réussissent la tâche (cas de discrimination parfaite), la proportion de ceux qui réussissent sera donc de $.10/.27 = 0.37$ et les 27 appartenant au groupe des plus faibles échoueront. L'indice de discrimination (cf. [calculer le d-index](#)) sera dans ce cas de 0.37 et ne pourra jamais être supérieur. Si le p-index était de 0.05, l'indice de discrimination maximum serait de 0.185 ! Pour un item facile (p-index = 0.90), la valeur serait aussi contrainte et égale à 0.37.

Pour identifier les "bons items" on peut les projeter dans un espace à deux dimensions. Dans cet espace, les coordonnées des items seront en abscisse le niveau de difficulté de l'item et en ordonnée la valeur de l'indice de Findley (d-index). Sur ce graphique tous les couples (p_index, d-index) ne sont pas possibles (zone "grillagée" sur le graphique). On peut par ailleurs identifier des zones de rejets claires (valeurs négatives de d-index). Enfin, pour les valeurs positives a décision dépend de la valeur maximale du d-index pour le p-index de l'item considéré (cf. exemple ci-dessous). Si le d-index maximal est 1, un d-index supérieur à .50 est habituellement considéré comme acceptable. Par contre si la valeur maximale du d-index est inférieure à 1 pour le niveau de difficulté, on ne retient que les d-index d'autant proches de la valeur maximale que cette valeur maximale est proche de 0.

Exemple : dans la figure suivante, on observe que l'item "a" est clairement un bon item. Les items "b" et "c" sont à rejeter (le "c" à une valeur négative et le "b" à une valeur trop faible étant donné son niveau de difficulté (valeurs possibles vont jusqu'à 1). Pour l'item "d", la valeur du d-index est faible, mais le niveau de difficulté de l'item ne permettait pas un indice supérieur, on peut donc considérer l'item comme bon ou acceptable si l'on souhaite avoir des items faciles, qui apportent cependant peu d'information (car la plupart des personnes réussissent).

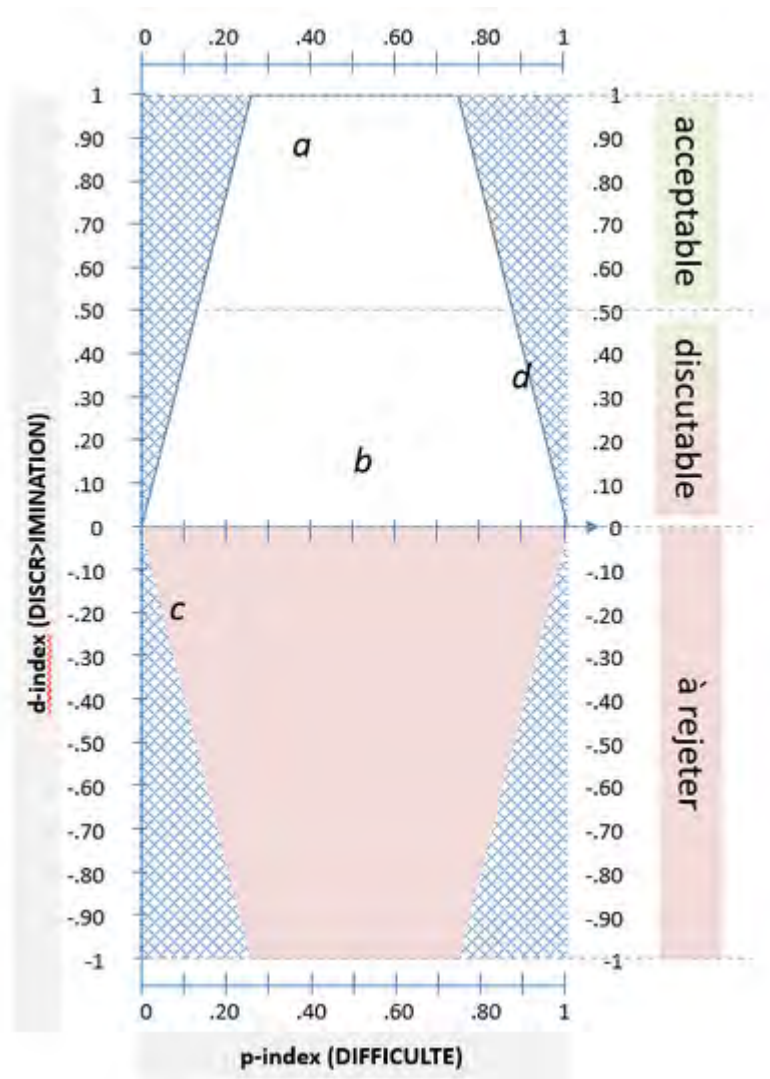


Figure C.2 : Représentation pour une aide à la prise de décision (acceptation ou rejet) concernant les items d'un test. (adapté de Laveault & Grégoire, 2012)

3.4. Le cas des items à choix multiples

L'analyse des items ne consiste pas à étudier uniquement les paramètres de difficulté et de discrimination pour sélectionner les meilleurs items possibles. Elle se doit aussi d'être un ensemble de techniques permettant d'améliorer le contenu et la forme des items. Dans ce cadre, un exemple courant est l'analyse plus détaillée des questions à choix multiples.

Lors de la cotation des items, pour tenir compte des réponses au hasard dans le cas de choix multiples, on suppose que tous les distracteurs ont la même probabilité d'être sélectionnés. Cette hypothèse est rarement vraie. Certains distracteurs peuvent être des attracteurs de mauvaises réponses ou à l'inverse n'être jamais sélectionnés. L'analyse des items doit permettre d'améliorer la formulation ou le choix de ces distracteurs.

Cette analyse des distracteurs pour les questionnaires à choix multiples est très discutée dans le champs de l'édu-métrie (terme préféré à psychométrie pour les tests d'évaluation des connaissances par les enseignants). Pour réaliser cette analyse on ajoute aux indices habituels de difficulté et de discrimination, différents indicateurs concernant les distracteurs ou les items :

- pour chaque distracteur :
 - l'indice de non efficacité (ou non fonctionnement) des distracteurs (NFD pour "Non Functioning Distractor") : il correspond au pourcentage de personnes ayant sélectionné ce distracteur parmi les répondants. En règle générale, cet indice ne doit pas être inférieur à 5% (sinon on considère que ce distracteur ne sert à rien ou n'est pas suffisamment plausible).
 - un indice de discrimination similaire à l'[indice de Findley](#), correspond au pourcentage de personnes sélectionnant ces distracteurs parmi les meilleurs moins les personnes qui sélectionne cet item parmi ceux qui ont les scores les plus faibles (Osterlind, 2002, p.271). Cet indice pour les distracteurs doit être faible voir négatif. Si ces valeurs sont positives, il doit être comparé au même indice calculé pour chaque bonnes réponses de l'item (le plus souvent une seule bonne réponse possible). S'il est proche voir supérieur, c'est que ce distracteur est un "attracteur" de bonnes réponses ou que ce choix est trop proche d'une bonne réponse.
- pour chaque item
 - on calcule parfois un indice d'efficacité des distracteurs (ED ou DE pour "Distractor Efficiency") : il correspond simplement au nombre des distracteurs efficaces (NFD > 5%) sur le nombre de distracteurs de l'item. Cet indicateur est cependant peu utile. L'intérêt de cet indice est qu'il permet de résumer les caractéristiques d'un questionnaire ou, lorsqu'il y a de nombreux items, permet aussi d'identifier rapidement les items qui nécessitent d'être examinés en priorité.

Ces analyses et ces indices s'interprètent facilement lorsque les épreuves sont des épreuves à choix multiples simples sachant que le format le plus habituel est une bonne réponse parmi 4 ou 5 (Tarrant, Ware, & Mohammed, 2009) . Généraliser ces indices à d'autres type de formats (k bonnes réponses parmi n) doit être fait avec prudence : l'interprétation de ces indices dépend du nombre de choix, du nombre des bonnes réponses mais aussi du niveau de difficulté de l'item (cf. un exemple d'analyse de distracteurs : Toksöz & Ertunç, 2017).

EXEMPLE D'ANALYSE DES DISTRACTEURS POUR 3 ITEMS

Un enseignant analyse les réponses données à un questionnaire à choix multiples (1 réponse correcte parmi 4). Il calcule pour chaque distracteur l'indice de non fonctionnement des distracteurs (NFD) et un indice de discrimination. Pour chaque item, il calcule aussi un indice d'efficacité des distracteurs (ED). Le tableau suivant reporte pour 3 des 50 items les résultats observés :

		réponses				ED
		A	B	C	D	
ITEM 1	NFD	0,04	0,45	0,12	0,24	66%
	upper 27%	0,07	0,93	0,04	0,11	
	lower 27%	0,00	0,26	0,37	0,56	
	différence	0,07	0,67	-0,33	-0,44	
ITEM 2	NFD	0,23	0,22	0,35	0,20	100%
	upper 27%	0,30	0,19	0,26	0,67	
	lower 27%	0,11	0,22	0,44	0,00	
	différence	0,19	-0,04	-0,19	0,67	
ITEM 3	NFD	0,04	0,05	0,70	0,21	33%
	upper 27%	0,00	0,15	0,96	0,11	
	lower 27%	0,11	0,00	0,26	0,19	
	différence	-0,11	0,15	0,70	-0,07	

Exemple de grille d'analyse de trois items d'un test avec une bonne réponse parmi 4 (**en grisé, la bonne réponse attendue pour l'item**). Pour la bonne réponse attendue, l'indice NFD correspond à l'indice de difficulté de l'item (équivalent d'un p-index) et l'indice donné par la colonne différence correspond à l'indice de discrimination de l'item).

❖ Analyse

- ITEM 1: Cet item est de difficulté intermédiaire (45% de réussite) et un distracteur (le distracteur A) semble inutile car seul 4% des répondants le sélectionne (ce qui donne un ED de 66% car 2 distracteurs sur 3 semblent efficaces). Le distracteur A pourrait être révisé (mais avec pour conséquence de modifier potentiellement les autres indices dont celui de difficulté de l'item).
- ITEM 2 : Cet item difficile (20% de bonnes réponses) présente des distracteurs dont tous les NFD sont compris entre 23% et 35%. Le distracteur A semble un distracteur attractif pour les meilleurs et le distracteur C pour les moins efficaces. L'indice de discrimination des distracteurs (ligne différence) donne une valeur positive pour le distracteur A. Il est cependant bien inférieur à l'indice de discrimination de l'item (0.67) et pourrait être considéré comme acceptable.
- ITEM 3 : cet item facile (70% de bonnes réponses), présente deux items avec un indice NFD insuffisant (ils ne sont presque jamais sélectionnés). L'indice ED est donc de 33%. Le seul distracteur utile est le D. En modifiant les distracteurs A et B on peut les rendre plus utiles mais cela peut modifier (en conséquence) la difficulté de l'item.

4. MRI-TRI

La construction habituelle des tests repose sur un modèle : la [théorie classique des tests](#). Le problème de ce modèle est la relativité des propriétés métriques qui en découlent puisque les normes sont très dépendantes de l'échantillon de standardisation ou d'étude des items. Par exemple, la difficulté d'un item est définie comme la proportion de personnes qui répondent correctement à l'item dans l'échantillon. Si les personnes testées sont "peu performants", l'item sera considéré comme plus difficile. Par contre, s'ils sont très performants, l'item sera considéré comme plus facile (d'où la nécessité de bien échantillonner même dans la phase de construction et de sélection des items). De plus, la théorie classique des tests repose sur des postulats très forts qui ne sont pas toujours vérifiés

(par exemple : l'erreur standard dans le cadre de la TCT est supposée la même quel que soit la position de la personne sur la dimension examinée).

Lord et Rasch (dans les années 1950 et 1960) ont voulu développer une méthode indépendante de l'échantillon et ont proposé **un premier modèle de réponse à l'item (MRI) appelé aussi théorie des réponses à l'item (TRI)**, qui permet d'estimer le niveau d'une personne en référence à l'ensemble de la population sans se baser sur les qualités intrinsèques des personnes qui ont été évaluées. Cette méthode est probabiliste et suppose (postulat) que la réponse à un item (la probabilité de répondre correctement) est une fonction des caractéristiques de l'individu (traits latents) et des caractéristiques de l'item (niveau de difficulté dans le premier modèle de Lord et Rash). D'un point de vue technique, la relation est formalisée par une fonction appelée [courbe caractéristique de l'item](#) (CCI).

Avantages ...

- Cette méthode présente l'avantage de fournir des informations qui sont indépendantes des caractéristiques des personnes, ce qui permet de créer des banques d'items c'est-à-dire, de vastes ensembles d'items dans lesquels on puise pour construire des tests. A chaque création d'un nouvel ensemble d'items, une étude de leurs propriétés métriques est réalisée.
- Les MRI permettent de développer le testing adaptatif, c'est-à-dire le fait d'administrer des items à une personne en fonction de son niveau supposé qui sera ajusté pendant la passation. Cette méthode augmenterait la sensibilité des épreuves tout en réduisant le nombre des items présentés.

... et inconvénients

- Une analyse des items dans le cadre des MRI requiert une expertise statistique beaucoup plus poussée que la théorie classique des tests.
- Les MRI nécessitent de recueillir de nombreuses réponses pour chaque item (et à minima variées) pour avoir vraiment une bonne indication des caractéristiques d'un item. L'indépendance des caractéristiques métriques des personnes à qui on a fait passer les items supposent qu'il existe une variabilité minimum entre ces personnes sur la dimension évaluée et un nombre suffisant d'observations.

Aller plus loin...

Pourquoi parle-t-on lors de l'étude des caractéristiques des items d'une estimation indépendante des caractéristiques des personnes ?

Le principe consiste à modéliser la probabilité qu'une personne réponde correctement à une question (item) en fonction de sa compétence latente (un score caché qu'on ne connaît pas). Lors de l'étude des paramètres des items (comme la difficulté), le niveau de compétence des personnes (θ |thêta) est inconnu et θ comme les paramètres des items doivent être estimés ensemble. Il existe plusieurs techniques (estimations conjointes, estimations marginales, méthode du maximum de vraisemblance conditionnelle, etc.). La structure spécifique des modèles de réponse à l'item permet d'estimer les caractéristiques des items indépendamment de celles des personnes car ils présentent une propriété mathématique appelée séparabilité des paramètres, qui permet de distinguer l'effet de la compétence des individus de celui de la difficulté des items dans la probabilité de réponse. Ainsi, les paramètres des items (tels que la difficulté) peuvent être estimés de manière fiable, quelle que soit la distribution des compétences dans l'échantillon de personnes, pourvu que celui-ci soit suffisamment varié.

(*) Séparabilité des paramètres : c'est une propriété de certains modèles statistiques qui permet d'estimer un ensemble de paramètres sans avoir à connaître ou estimer précisément un autre ensemble de paramètres liés. Cette indépendance d'estimation repose sur une formulation mathématique dans laquelle les effets des différentes sources de variation apparaissent de manière distincte dans le modèle. Prenons comme exemple un

modèle linéaire classique en statistiques : $Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$ (avec Y_{ij} la mesure observée ; μ une constante ; α_i l'effet d'un facteur ; β_j l'effet d'un second facteur ; ϵ_{ij} l'erreur aléatoire). Ici la contribution de chaque paramètre est additive et peut être séparée mathématiquement. Cela permet d'estimer l'effet des facteurs indépendamment tant que le modèle est bien spécifié et que les données couvrent suffisamment de données.

4.1. Les postulats

Les modèles de réponse à l'item sont un cadre de construction des tests qui est différent de celui utilisé dans la théorie classique des tests (TCT). Comme dans la TCT, ils reposent néanmoins sur un certain nombre de postulats.

1. **Postulat d'unidimensionnalité(*)**. Ce postulat suppose que tous les items qui composent un test sont dépendants d'une seule dimension sous-jacente (d'un seul trait latent). Il est rare cependant que cela soit le cas et en pratique la définition d'unidimensionnalité est considérée comme respectée si une dimension est clairement dominante pour tous les items du test.
2. **Postulat d'indépendance locale** des items d'un test (les relations entre les items s'expliquent uniquement par le trait latent mesuré). Cette propriété n'est pas identique à celle d'unidimensionnalité. Un test peut être unidimensionnel mais la probabilité de bonne réponse à un item peut dépendre des réponses données aux items antérieurs (il n'y a plus alors d'indépendance locale).
3. **Monotonie**. La réponse d'une personne peut-être modélisée par une fonction monotone croissante décrivant la relation entre la probabilité de réussite à un item et la position de cette personne sur la dimension mesurée (trait latent).

(*) *Le postulat d'unidimensionnalité est difficile à juger car les réponses dépendent de nombreux facteurs individuels (motivation, stratégies, anxiété, etc) mais aussi des caractéristiques de la situation (comme la séquence d'items présentés). Pour tester l'unidimensionnalité d'un test on peut utiliser des techniques comme l'analyse factorielle. Il existe aussi des MRI qui sont multidimensionnels, la dimensionnalité d'un ensemble d'items correspond alors au nombre de traits latents suffisant à l'explication des relations entre les items.*

4.2. Courbe caractéristique d'un item (CCI)

Le principe de base des [Modèles de Réponses à l'Item](#) est que la performance d'un sujet à un item peut être expliquée par un facteur appelé trait latent. Il peut s'agir d'un trait de personnalité, d'une aptitude cognitive, d'une compétence générale, etc. Cette variable (ce trait latent) est habituellement notée θ). La relation entre les performances à l'item (probabilité de réussite ou probabilité de fournir une réponse donnée) et le trait latent est décrite au moyen d'une fonction appelée courbe caractéristique de l'item (CCI).

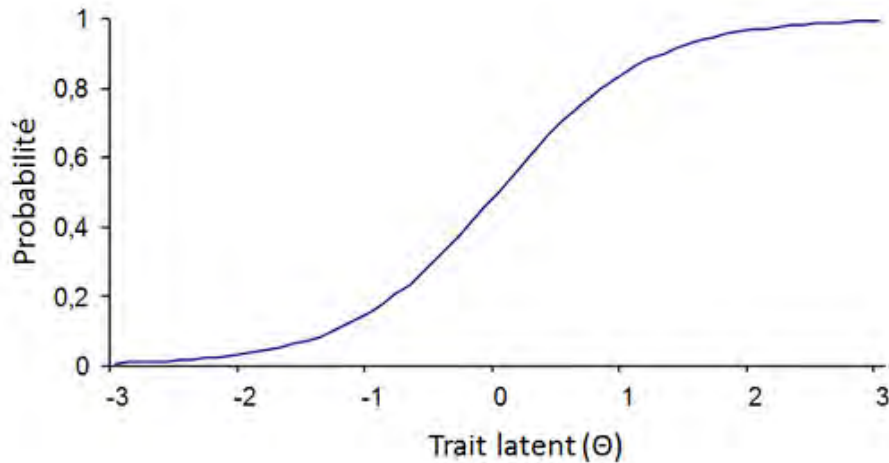


Figure C.3 : Représentation d'une courbe caractéristique d'un item (CCI)

Cette courbe prend habituellement la forme d'un S plus ou moins allongé (sigmoïde). Le trait latent étant supposé distribué normalement au sein de la population, les graduations de l'abscisse correspondent aux valeurs de θ (position vraie du sujet dans la population) exprimé en score z . Le niveau moyen correspond donc à la valeur zéro et la distance d'un écart-type par rapport à cette moyenne est représentée par +1 ou -1. La forme de la courbe traduit une relation entre le trait latent et la probabilité de réussite d'un item.

Remarque : en général l'échelle sur le trait latent prend des valeurs (Θ) entre -4 et +4 mais parfois , comme c'est le cas dans tous nos exemples, l'échelle des valeurs varie entre -3 et +3.

4.3. Paramètres des CCI

D'un point de vue mathématique, la fonction (courbe) caractéristique d'un item ([CCI](#)) est exprimé par une équation particulière (cf. ci-dessous). Cette courbe dépend de plusieurs paramètres dont le nombre varie selon le [modèle](#) utilisé. Les paramètres possibles qui décrivent les propriétés des items sont habituellement au nombre de trois : [le paramètre de difficulté](#) (toujours présent), le [paramètre de discrimination](#) et le [paramètre de pseudo-chance](#).

Pour ceux qui veulent aller plus loin (voir aussi Dickes, Flieller, Tournois et Kopp, 1994).

L'équation de la courbe caractéristique d'un item est :

$$P_j(x = 1|\theta) = \gamma_j + \frac{1 - \gamma_j}{1 + e^{-D\alpha_j(\theta - \delta_j)}}$$

avec :

θ -----	la valeur du trait latent
$p_j(x = 1 \theta)$ -----	la probabilité de répondre correctement à l'item j sachant θ
e -----	la base des logarithmes naturels ($e = 2.71828$)
δ_j -----	le paramètre (delta) de difficulté de l'item
α_j -----	le paramètre (alpha) de discrimination (égal à 1 dans le modèle à un paramètre)
γ_j -----	le paramètre (upsilon) de pseudo chance (égal à 0 dans les modèles à 1 et 2 paramètres)
D -----	Constante (souvent la valeur utilisée est 1.7)

4.3.1 Paramètre de difficulté

Le paramètre de difficulté des items, que l'on appelle le paramètre bêta (β), est présent dans tous les modèles. Par convention, la valeur qui représente la difficulté d'un item est égale à la valeur de θ pour laquelle la probabilité de donner une réponse correcte est de .50. Ainsi, dans le graphique ci-dessous, l'item (qui correspond à la CCI tracée) à une difficulté égale à -1

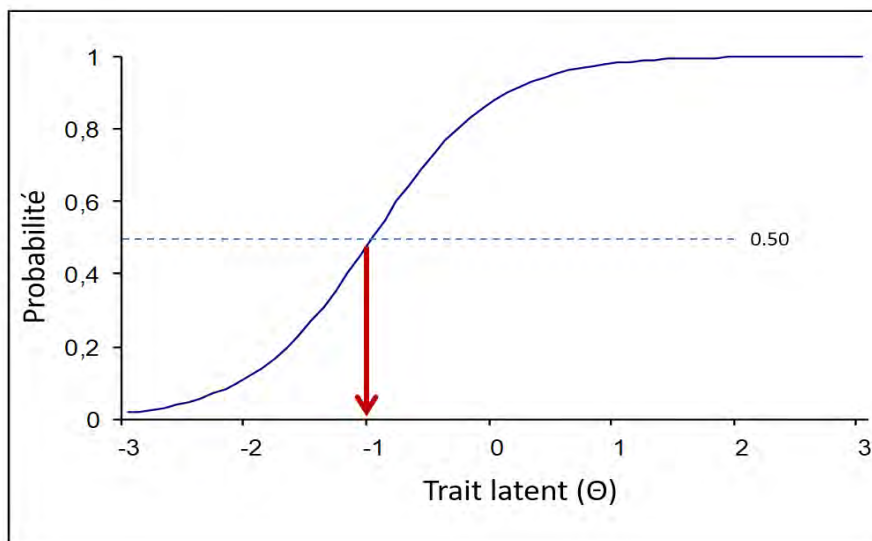


Figure C.4 : CCI d'un item de niveau de difficulté à peine supérieure à -1

Remarques :

- Dans ce cadre, si l'on a affaire à des items de performance, les valeurs négatives traduisent des items faciles à très faciles et à l'inverse, des scores positifs des items difficiles à très difficiles. Une valeur de difficulté de 0 correspondra à un item de difficulté moyenne.
- La fonction décrivant la relation entre trait latent et probabilité de réussite à l'item étant une fonction monotone la probabilité de réussir varie de façon continue selon la valeur de θ . Normalement la fonction est croissante (condition nécessaire mais non suffisante pour que l'item soit acceptable).
- On trouve toujours, quel que soit le modèle utilisé, une CCI qui s'ajuste aux données (comme en régression, on peut toujours trouver une droite qui résume un nuage de points). Cependant, lors du calibrage des items (estimation des paramètres), on doit vérifier l'ajustement des données aux exigences du modèle (le bon ajustement des données doit être vérifié pour chaque item).

4.3.2 Paramètre de discrimination

C'est Birnbaum qui introduit (dans les modèles à deux paramètres), le deuxième paramètre correspondant au niveau de discrimination, que l'on appelle paramètre α . La discrimination de l'item (c'est à dire sa propension à bien discriminer les individus les uns des autres) est représentée par la pente de la CCI (tangente au point d'inflexion de la CCI). Celle-ci peut être plus ou moins inclinée. Plus la pente est "abrupte" (tend à être parallèle à l'axe Y), plus l'item est discriminant et inversement.

Par exemple, ci-dessous, les items ont la même difficulté et ne sont pas aussi discriminants. Les pentes des courbes représentées sont clairement différentes. L'item A est plus discriminant.

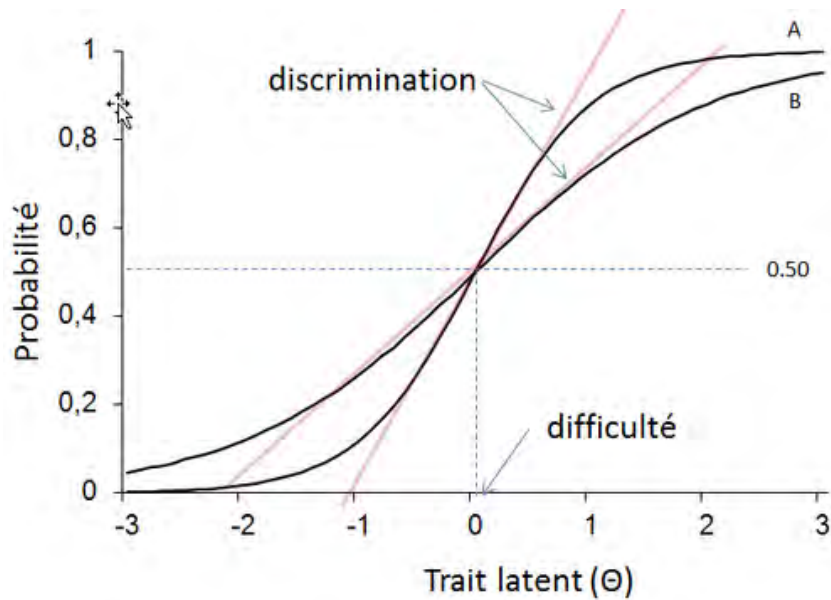


Figure C.5 : CCI de deux items ayant le même niveau de difficulté mais des paramètres de discrimination différents

Comprendre : un item correspondant à un niveau I sur le trait latent (Θ) est discriminant si la probabilité de réussite est minimum lorsque la position sur le trait latent est inférieure à I et maximum lorsque la valeur sur le trait latent est supérieure à I . En fait, l'intervalle de valeur sur Θ pour que la probabilité de réussite passe de son minima à son maxima doit être minimum si l'on veut que l'item soit discriminant (cf. figure E.6).

Remarque : En règle général la valeur de α (paramètre de discrimination) est positive et varie entre 0 et 2 (2 étant une valeur élevée). Plus cette valeur est élevée, plus l'item est discriminant. Si la valeur de la pente était négative cela signifierait que la probabilité de réussir l'item diminue lorsque la compétence augmente (absurde, la fonction serait monotone décroissante).

4.3.3 Paramètre de pseudo-chance

Le troisième paramètre des CCI est celui dit de la pseudo-chance (appelé parfois « paramètre c »). Il correspond à l'asymptote "basse" de la courbe. En fait, dans la réponse à un item on peut concevoir que des facteurs aléatoires puissent influencer la réponse (ou la performance). Dans ce cadre, pour la valeur minimale de θ (Θ) (-3 ou -4 le plus souvent), la probabilité de répondre comme attendu correspondra à une probabilité non nulle (cf schéma ci-dessous).

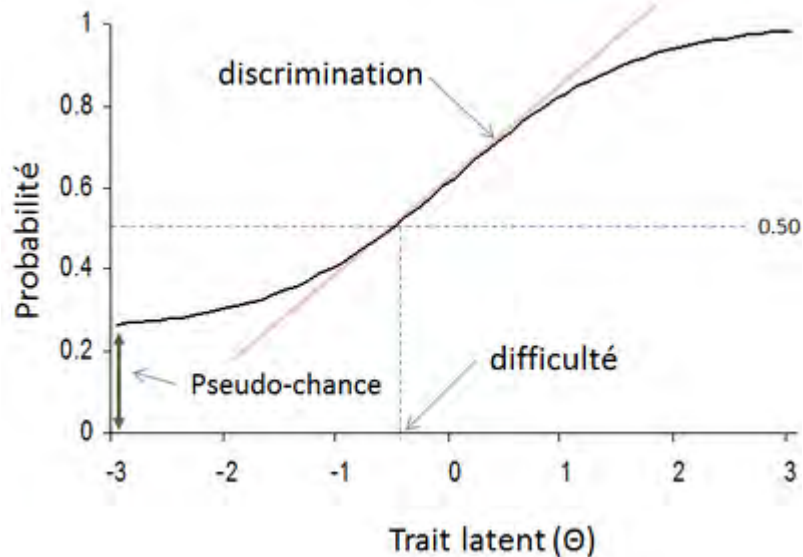


Figure C.6 : CCI d'un item avec un paramètre de "pseudo-chance" différent de 0.

4.4. Les différents modèles

Nous venons de voir que les modèles de réponse à l'item concernent la relation existant entre la probabilité de répondre correctement à un item et les caractéristiques de l'individu et de l'item. La complexité de ces modèles dépend de la fonction de répartition choisie (relation) mais aussi du nombre de paramètres que l'on souhaite prendre en compte. Plusieurs modèles de réponses sont habituellement distingués :

- **Le modèle de Rash** : c'est un modèle simple à un paramètre (difficulté). Tous les items sont supposés avoir le même pouvoir discriminant. Dans ce modèle, la valeur de D ([constante de l'équation de la CCI](#)) est fixée le plus souvent fixé à 1.7 (la courbe est alors proche d'une ogive normale, ce qui correspond à l'intégrale de la courbe normale). Le paramètre de discrimination (α_j) et de pseudo-chance (γ_j) sont fixés respectivement à 1 et 0.
- **Le modèle de Birbaum** : c'est un modèle à deux paramètres (difficulté et discriminabilité). Le paramètre de discrimination (α_j) varie aussi en fonction des items mais le paramètre de pseudo-chance (γ_j) reste fixé à 0.
- **Le modèle à 3 paramètres** : ce modèle ajoute simplement au modèle précédent le paramètre de pseudo-chance. Le paramètre de pseudo-chance, comme les deux autres, varie donc en fonction des items.

Initialement construit pour des données dichotomiques sous-tendues par une seule dimension, on distingue les modèles selon le nombre possible de réponses. Les items V-F ou à choix multiples (avec n réponses possibles) pour lesquels il n'y a qu'une réponse possible (correct-incorrect) sont des modèles pour données dichotomiques. Si dans les choix multiples il y a plusieurs bonnes réponses possibles, ou lorsque l'on utilise par exemple des échelles de Likert, il faut utiliser d'autres catégories de modèles qui sont des modèles polytomiques (pour plus de détails, cf. l'article d'introduction de [van der Linden, 2010](#)).

4.5. Des items aux individus

Les courbes caractéristiques des items permettent, à partir des réponses d'un individu, de situer la réponse donnée par un individu sur le trait latent. Pour illustrer et présenter, les principes généraux de

cette démarche, nous nous placerons dans la cadre des items dichotomiques. Ce chapitre introductif, très simplifié présente successivement : la courbe de vraisemblance, la notion d'information et l'erreur standard de mesure dans les MRI.

Pour aller plus loin, vous pouvez consulter l'article de Géraldine Rouxel (1990) comme un document introductif à l'utilisation des MRI avec des items polytomiques.

4.5.1 Courbe de vraisemblance

Pour un test de n items dichotomiques (avec 1=réussite et 0=échec), on peut avoir 2^n profil de réponses. Par exemple pour 3 items (A_1, A_2, A_3), les profils de réponses sont au nombre de $2^3=8$: [0,0,0], [0,0,1], [0,1,0], [0,1,1], [1,0,0], [1,0,1], [1,1,0], [1,1,1]. Puisque l'on connaît les [CCI](#) des items, on peut calculer la probabilité $p(x)$ de chacun de ces profils pour chaque valeur de θ (thêta) en utilisant d'une part la propriété d'[indépendance locale](#) et d'autre part le théorème qui dit que la probabilité que plusieurs événements indépendants se produisent est égal au produit de leur probabilité.

$$p(E_1 \dots E_j \dots E_n) = \prod_j p(E_j)$$

Dans notre exemple avec 3 items, la probabilité de l'événement [0,1,0], pour une valeur de $\theta = 1$ par exemple sur le trait latent, sera donc le produit de la probabilité de réussite à l'item A2 et la probabilité d'échec à l'item A1 et A3, calculer à partir des CCI de chaque item pour la valeur $\theta = 1$. Sachant que la probabilité d'échec est égale à un moins la probabilité de réussite [$p(A_i/\theta_j)$] donnée par la CCI, on a dans notre exemple :

$$p([0,1,0]/\theta_{j=1}) = [1-p([A_1]/\theta_{j=1})] * [p([A_2]/\theta_{j=1})] * [1-p([A_3]/\theta_{j=1})]$$

On peut calculer cette probabilité pour chaque valeur de thêta (de -3 à +3) et obtenir ainsi, pour un profil de réponse donné, une courbe de vraisemblance (cf. schéma ci-dessous). Cette courbe passe par un maximum qui correspond à la valeur θ qui sera attribuée à la personne qui a ce profil de réponse. Dans l'exemple ci-dessous, ce score, pour le profil [A_1, A_2, A_3]=[0,1,0] est de 1.14.

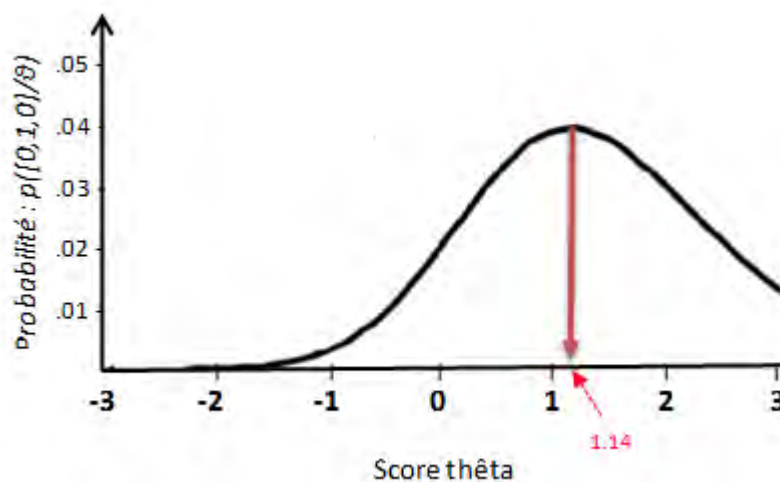


Figure C.7 : Courbe de vraisemblance pour la configuration de réponse à 3 items [0,1,0]

Remarque importante : ce qui permet de déterminer le niveau n'est plus la somme des scores (points obtenus = performance) mais le profil des scores. Le même total de points, qui dans une perspective classique correspond à une seule performance, peut correspondre à deux niveaux de thêta différents donc à des positionnements différents sur le trait latent.

4.5.2 Courbe d'information

Dans un test cognitif, un item difficile apporte peu d'information (voir aucune information) sur la position d'une personne qui présenterait des difficultés (position basse sur le trait latent mesuré) et inversement un item facile apporte peu d'information concernant une personne qui aurait, sur ce même trait latent une valeur élevée. Selon le paramètre de difficulté, un item permettra de différencier plus ou moins les élèves en fonction de leur position sur le trait latent. Cet exemple concernant le paramètre de difficulté est une façon d'illustrer le fait que chaque item apporte une information différente selon la valeur que prend θ (trait latent).

On a vu que la courbe caractéristique d'un item associe à chaque valeur de θ (trait latent) la probabilité de réussir cet item. On peut, à partir des paramètres de cette courbe, calculer une courbe qui renseigne sur le "pouvoir" d'information de l'item en fonction des valeurs de θ . Le graphique suivant illustre cette relation entre θ et le niveau d'information apporté par deux items.

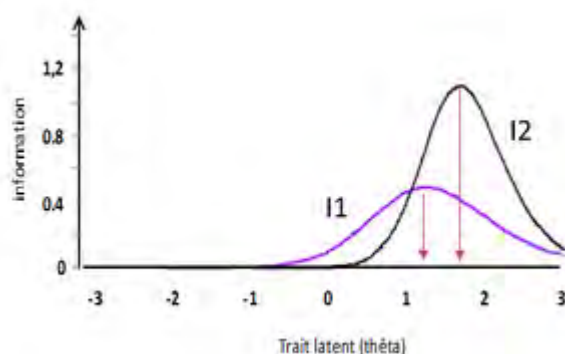


Figure C.8 : Courbe d'information de deux items d'un test (I1 et I2)

Le point haut des courbes renseigne sur le niveau du trait latent pour lequel l'item apportera le plus d'information (sera le plus précis). Dans l'exemple ci-dessus, le pouvoir informatif maximum est observé pour l'item I1 avec une valeur de θ égale à 1.20. Pour I2, le maximum correspond à une valeur de θ égale à 1.75.

Remarques :

- Dans les deux exemples précédents faire passer ces deux items n'apportent aucune information significative pour des personnes dont la position sur le trait latent serait négative (I1 apporterait en fait très peu d'information et I2, aucune information).
- A partir des courbes d'information des items d'un test, on peut établir la courbe d'information du test qui s'obtient simplement en additionnant les courbes d'information des items (pour chaque valeur de θ). Cette courbe d'information peut avoir plusieurs points hauts ce qui signifie que le test présente un pouvoir informatif plus important pour différentes valeurs du trait latent.

(*) L'équation qui définit la fonction d'information d'un item est :

$$I_j(\theta) = D^2 \alpha_j^2 \frac{1 - p_j(\theta)}{p(\theta)} \left[\frac{p_j(\theta) - \gamma_j}{1 - \gamma_j} \right]^2$$

avec :

$I_j(\theta)$ ----- Information associée à l'item j au point θ
 θ ----- la valeur du trait latent

$p_j(\Theta)$ -----	la probabilité de répondre à l'item j lorsque la valeur du trait est Θ
α_j -----	le paramètre de discrimination (égal à 1 pour un modèle à un paramètre)
γ_j -----	le paramètre de pseudo chance (égal à 0 pour les modèles à 1 et 2 paramètres)
D -----	Constante (souvent la valeur utilisée est 1.7)

4.6. Intérêts et limites

La présentation des MRI faite ici est une présentation simplifiée. Le développement de ces modèles a permis de généraliser ces MRI aux questions polytomiques et aux échelles multidimensionnelles avec des modèles paramétriques et non paramétriques. Les avantages de ces MRI sont multiples même si leur utilisation peut paraître complexe. Ils permettent de répondre à des questions pour lesquelles la TCT n'apportait pas de réponse. Par exemple :

- Ces modèles sont utilisés dans les grandes enquêtes internationales comme PISA (*Program for International Student Assessment* de l'OCDE) et permettent de comparer des niveaux de compétence même si une partie des items sont différents d'un pays à l'autre (cf. à ce sujet l'article de [Pierre Vrignaud](#) [2006] qui reprend les avantages et les limites de ce type de modèles pour ces études).
- Ces modèles devraient être à la base d'un renouveau des tests adaptatifs. Dans ces tests les items sont choisis en fonction du niveau de compétence calculé pendant la passation. Après chaque réponse on choisit comme item suivant celui qui a priori apportera le plus d'information. Dans ces tests, le niveau de compétence θ est donc estimé à partir des premières réponses, ce qui permet de choisir un item adapté à ce niveau a priori (à partir des courbes d'information des items). Le test est différent pour chaque individu et s'arrête quand on a obtenu un niveau de précision suffisant fixé préalablement. Le développement de banque d'item et l'informatisation devraient permettre à ces tests adaptatifs de se développer et de raccourcir notablement, sans perte de précision, des questionnaires parfois beaucoup trop longs (dans la TCT on augmente le nombre d'item pour augmenter la fidélité de la mesure).
- Il est facile d'estimer [l'erreur standard](#) de mesure (ESM) en fonction de la performance observée (et donc de θ) et des items passés alors que la TCT, comme nous le verrons, conduit à calculer un ESM identique quel que soit la valeur du trait latent.

Les MRI remplaceront probablement la théorie classique des tests. Ils sont de plus en plus utilisés. Cependant, ces méthodes ne sont pas sans critique ou sans problème :

- Comme la théorie classique des tests, la modélisation de la réponse à l'item repose sur des postulats et surtout une relation entre des probabilités de réussite à un item qui sont exprimées en fonction du trait latent. Malheureusement, il n'est pas possible d'étudier expérimentalement cette relation en contrôlant θ (mesure sur le trait latent que l'on ne connaît pas a priori). Le point zéro de l'échelle de difficulté est déterminé à partir de l'échantillon des personnes ayant passé le test.
- La qualité de l'estimation des paramètres dépend aussi des caractéristiques de "l'échantillon" qui se doit d'être hétérogène et surtout de taille suffisamment importante (selon le modèle utilisé et le nombre d'items, l'effectif minimum peut dépasser très rapidement 100 voir 500 personnes).
- La propriété d'invariance est la caractéristique principale des MRI. Elle postule que les

paramètres des items sont indépendants de l'échantillon et en parallèle que l'estimation relative à l'individu est indépendante de l'échantillon d'items utilisés. Cependant cette invariance est relative et n'est assurée que si certaines conditions sont satisfaites comme l'ajustement du modèle aux données pour la population comme pour chacun des sous-groupes la constituant et pour lesquels ils pourraient exister des différences (exemple, catégorie socioprofessionnelles, sexe, etc.). La prise en compte de ce qu'on appelle le fonctionnement différentiel des items est possible mais complexe dans l'analyse des données.

→ Les postulats concernant les TRI (MRI) sont forts et particulièrement celui concernant l'indépendance locale (la variabilité des résultats doit dépendre exclusivement du trait mesuré).

D - Les qualités métrologiques

Étudier les qualités métrologiques d'un test revient à étudier "les qualités de la mesure". En psychométrie, classiquement 4 propriétés correspondant à 4 questions fondamentales sur la mesure. Ce sont ces propriétés qui sont particulièrement étudiées ou analysées:



- **Sensibilité** : le test différencie-t-il suffisamment les sujets ?
- **Unidimensionnalité - Homogénéité** : les items constituent-ils entre eux une mesure homogène - unidimensionnelle ?
- **Fidélité(s)** des tests : le test mesure-t-il quelque chose ? quelle est l'importance de l'erreur de mesure ?
- **Validité** des tests : le test mesure-t-il ce que je voulais mesurer ? ou le test fournit-il une information suffisante qui correspond à ce dont on a besoin celui qui l'utilise ? »

1. Sensibilité

Il existe plusieurs conceptions de la sensibilité d'un test selon que l'on se place dans le cadre de la mesure d'une dimension (intelligence, aptitude, attitude, motivation, intérêts, trait de personnalité, etc) ou dans le cadre d'un dépistage d'une caractéristique spécifique que l'on peut éventuellement mesurer de façon dichotomique en terme de présence/absence (présence ou absence d'un trouble par exemple).

Dans le premier cas, le test doit différencier au mieux l'ensemble des individus, dans le second cas, que l'on peut appeler abusivement "test diagnostic", il doit détecter au mieux les personnes ayant la caractéristique recherchée. On ne parle plus alors uniquement de sensibilité mais aussi de [spécificité](#). Ces deux caractéristiques impactent la [validité](#) du test. Ce deuxième aspect de la sensibilité n'est pas l'objet du cours mais nous le présentons pour information (cf. aussi le chapitre "[détermination d'un score seuil](#)").

1.1. Sensibilité et mesure d'une dimension

1.1.1 Définition

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. 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 pouvoir différenciateur plus important.

Si la distribution n'est pas une distribution normale, la sélection des questions était probablement

incorrecte et le choix des questions doit être revue et/ou les questions remaniées. Lorsque l'on 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.

Remarque : La sensibilité d'une épreuve dépend du nombre d'items mais aussi des caractéristiques des items. Par exemple, si dans une version provisoire d'un test on retient 10 items sur les 20 initiaux, on peut avoir les deux cas de figures suivants :

- (a) *les 10 items sont de difficulté croissante ce qui permet de classer les sujets en 11 classes (notes allant de 0 réussite à 10 réussites).*
- (b) *Si les 10 items se regroupent en k sous-groupes ayant le même niveau de difficulté, les sujets sont répartis uniquement en $k+1$ classes distinctes même si un sujet particulier peut avoir entre 0 et 10.*

Conclusion : dans le premier exemple le test sera plus sensible ou plus « discriminant », bien que le nombre d'items retenus soit identique. Augmenter le nombre des items sans contrôler leur difficulté n'augmente donc pas nécessairement la sensibilité d'une épreuve.

Pour aller plus loin.

Souvent, naïvement, on pense qu'augmenter le nombre d'items d'un test augmente son pouvoir discriminant, sa sensibilité. En fait ce problème est complexe. Lorsque l'on ajoute des items à un test on doit s'assurer que ces items corrélerent entre eux ce qui permettra d'augmenter la variance de l'épreuve puisque la variance totale est la somme des variances observées à chacun des items plus la somme des covariances entre ces items pris 2 à 2 (pour ceux qui en doutent, il est assez facile de démontrer ce théorème). Si les items ne sont pas homogènes (ne covarient pas), ils contribuent donc peu à l'augmentation de variance du score total dans un test. Pour qu'un test soit discriminatif, une solution consiste donc à augmenter le nombre d'items mais ceux-ci doivent être homogènes (corrélés entre eux) mais on doit aussi préserver la capacité du test à discriminer sur l'ensemble de l'échelle (donc avoir des items avec des niveaux différents de difficulté donc des corrélations qui ne peuvent pas être maximales).

Par ailleurs, il faut rappeler qu'un score total sera aussi "plus facilement" interprétable si les items qui le composent sont les plus homogènes possibles (sinon le même score peut avoir des significations différentes). Ce problème renvoie à l'unidimensionnalité des épreuves et soulignent que tous les choix, lors de la construction d'une épreuve, sont interdépendants. En effet, il faut savoir aussi, qu'augmenter la variance d'un test ou maximiser la sensibilité est une condition nécessaire (non suffisante) pour pouvoir s'assurer d'une bonne fidélité et validité des tests.

1.1.2 Pour le psychologue praticien

Pour un praticien, la sensibilité d'un test peut être parfois appréciée à travers l'analyse critique des tables d'étalonnage. Par exemple, ci-dessous se trouve reproduit une table de conversion des scores bruts en scores standards (table d'étalonnage) d'une version ancienne (donc non utilisable actuellement) d'une épreuve d'évaluation de l'intelligence. Pour chaque épreuve (CUB, SIM, MCH, etc.) on trouve dans les colonnes les notes brutes (scores possibles des personnes) qui sont associés à une note standard (en première colonne) pouvant varier de 1 à 19 (ici, le score standard a pour moyenne 10 et pour écart-type 3). Cette table de conversion concerne des enfants de 6;0 ans à 6;3 ans et on observe (colonnes encadrées en rouge) une faible sensibilité de certaines épreuves qui peut impacter l'analyse qualitative. En effet, si on échoue par exemple à l'épreuve de similitude (SIM, note de 0) la note standard est de 6. Elle serait significativement supérieure à une note standard de 1 obtenue avec un échec complet aussi à l'épreuve mémoire des chiffres (MCH). Par ailleurs on notera que l'épreuve

information (INF) est peu sensible car les scores standards possibles varient de 1 à 19 mais se limitent à : 1, 3, 5, 7, 9, 11, 13, 15, 17, 19. Pour cette classe d'âge, cette épreuve montre une sensibilité réduite des notes standards et les résultats devront être interprétés avec prudence quand on compare aux autres épreuves. .

Ages 6:0-6:3																
Notes standard	Notes										Notes standard	Notes				
	CUB	SIM	MCH	IDC	COD	VOC	SLC	MAT	COM	SYM		CIM	BAR	INF	ARI	RVB
1	0	-	0-2	0-1	0-0	0-6	-	0-3	-	0-3	1	0-3	0-20	0-2	0-1	-
2	1	-	3	2	10-12	7	0	4	0	4-5	2	4	21-23	-	2	-
3	2-3	-	4	3	5	8	1	5	1	6	3	-	24-25	3	3	-
4	4-5	-	5	4	9	2	-	-	-	7	4	5	26-27	-	4	0
5	8-7	-	6	5	10	3	6	2	6	8	5	6	28-29	4	5	1
6	8-6	0	7	6	4	-	3	9-10	6	7	2	32	-	6	2	
7	10-11	1-2	-	7-8	5	7	-	11-12	7	8	35	5	7	-	-	
8	12	3	8	9	6	8	4	13-15	8	9-10	38	-	-	3		
9	4	8	10	10	9	5	16-17	9	11	42	6	8	4			
10	1	10-11	1	10-11	6	18-19	10	12-13	17	-	9	5				
11	1	12-13	11	14-15	11	7	10	6								
12	2	5	-	11	7	-	11	7								
13	2	0	6	12	8	-	12	8								
14	3	6	-	13-14	9	-	13	9								
15	3	9	9	15	10	-	14	10								
16	3	1	10	16	-	-	15	11								
17	4	-	11	17	11	-	16	11								
18	4	-	11	18	12	-	17	12								
19	45-68	18-44	18-32	20-28	63-65	20-28	12-33	18-34	13-24							

Figure D.1 : Extrait d'une table d'étalonnage du WISC IV (Wechsler, 2005). Une partie de la table est masquée pour droits d'auteurs (cette version n'est plus valide actuellement).

(CUB = Cube, SIM = Similitude, MCH = Mémoire des chiffres, IDC = Identification de concepts, COD= Code, VOC = Vocabulaire, SLC = Séquence Lettre-Chiffre, MAT = Matrice, SYM = Symbole, CIM = Complètement d'image, BAR = Barrage, INF = Information, ARI = Arithmétique, RVB = Raisonnement verbale)

1.2. Sensibilité et spécificité

1.2.1 Définitions

Lorsque la mesure a pour objectif de dépister une caractéristique particulière (présence/absence d'un trouble), l'instrument doit avoir le meilleur pouvoir séparateur possible (= une bonne sensibilité) mais doit aussi avoir une forte spécificité.

- **La sensibilité** est, dans ce cadre, la capacité de l'instrument à identifier correctement les personnes présentant la caractéristique que l'on souhaite étudier. On parle de capacité de détection. Elle est mesurée (cf. formule ci-dessous) par la proportion de personnes présentant la caractéristique étudiée qui est identifiée par le test.
- **La spécificité** est la capacité de l'instrument à identifier correctement les personnes ne portant pas cette caractéristique. On parle de capacité de discrimination. Le coefficient de spécificité correspond à probabilité d'identifier correctement une personne ne présentant pas la caractéristique étudiée.

Un instrument peut être sensible et non spécifique ou inversement, spécifique et peu sensible. Un bon test est celui qui aura la meilleure sensibilité et spécificité possible. Avec l'indicateur de spécificité, dans ce cadre de test de dépistage on s'approche du concept de validité du test (notion abordée plus avant)

1.2.2 Calcul : Se et Sp

L'indice de sensibilité (Se) que l'on utilise est le pourcentage de personnes testées positives parmi les personnes étant réellement positives (positives sur un critère externe). L'indice de spécificité (Sp) est le pourcentage de personne testée comme négative (ne possédant pas de trouble) parmi les personnes n'ayant effectivement aucun trouble (on parle de Vrais négatifs).

La méthode de calcul peut être surprenante car elle s'appuie sur un critère externe qui a priori nous permet déjà de savoir si la personne présente le trouble ou non. Qu'apporte alors le test ? En fait plusieurs situations sont possibles : le critère externe peut être coûteux, invasif, long et/ou difficile à mettre en oeuvre. Avoir une épreuve plus simple pour décider permet d'avoir un critère efficace et un préalable avant de faire des examens plus approfondis. Un autre cas concerne des épreuves qui veulent prédire une situation qui arrive ultérieurement. Le critère sera l'apparition ou non de la maladie (l'étude de la sensibilité et de la spécificité peut cependant prendre du temps !) et la possibilité éventuel de prévenir, atténuer ou de retarder l'apparition des troubles.

Pour illustrer le calcul des indices de sensibilité (Se) et de spécificité (Sp), on présente les résultats d'une étude concernant un test permettant de tester s'il existe ou non un trouble psychopathologique (étude fictive). Cette étude porte sur 113 personnes dont 85 ne sont pas porteurs de la pathologie étudiée (donc 28 l'ont !). Après avoir fait passer le test, 26 personnes sont considérées comme positives (test positif) par le test (score supérieur à un seuil fixé) et parmi ces 26 personnes, 25 personnes sont des vrais positifs. Quarante-sept personnes sont considérées comme négatives (test négatif) dont 3 à tort (faux négatifs). Ces données sont résumées dans le tableau suivant où l'on distingue les Vrais Positifs (VP=25), les Faux Positifs (FP=1), les Faux Négatifs (FN=3) et les Vrais Négatifs (VN=84).

Valeur diagnostic		
	Positive	Négative
Test positif	VP (Vrais Positifs)	FP (FauxPositifs)
	25	1
Test négatif	FN (Faux Négatifs)	VN (Vrais négatifs)
	3	84

Calcul de la sensibilité (nombre de vrais positifs divisé pas le nombre des personnes effectivement "positives") :

$$Se = VP/(VP+FN)$$

$$Se = 25/(25+3) = 89\%$$

Calcul de la spécificité (nombre de vrais négatifs divisé par le nombre des personnes effectivement "négatives") :

$$Sp = VN/(FP+VN)$$

$$Sp = 84 / (84+1) = 99\%$$

1.2.3 Importance de la prévalence

L'utilisation des indicateurs de spécificité et sensibilité est une source d'erreur (biais cognitif). Le plus simple pour illustrer ce biais est de vous y confronter.

Soit un test dont la sensibilité serait de 99% et sa spécificité serait aussi élevée(98%).

Question : On vous fait passer le test. Vous n'étiez pas inquiet car la prévalence dans la population est (1 pour 100). Vous êtes malheureusement positif au test ! Quelle est la probabilité d'avoir le

trouble (ou la maladie) ?

Réponse probable : au regard des valeurs de spécificité et de sensibilité la majorité des personnes répondent (à tort) que la probabilité d'avoir le trouble est élevée.

En fait ce n'est pas si simple. Vérifions ce qu'il en est avec un simple calcul.

On peut constater que l'épreuve détecte correctement les vrais positifs dans 99% des cas (sensibilité) et les vrais négatifs dans 98% des cas (spécificité).

Pour répondre à la question je peux utiliser le théorème de Bayes :

$$p(A|B) = \frac{p(B|A) \times p(A)}{p(B)}$$

avec : A = présence du trouble ; B = test positif

nous savons que

$$P(B|A) = 0.99 \text{ (c'est la sensibilité du test)}$$

$$P(A) = 0.01 \text{ (prévalence)}$$

par ailleurs

$$p(-A) = 0.99 \text{ (proportion de personnes sans trouble soit } 1 - p(A))$$

$$P(B|-A) = 0.02 \text{ (correspond à 1 moins la spécificité soit } 1 - Sp)$$

donc

$$p(B) = p(B|A) \times p(A) + P(B|-A) \times p(-A) = 0.99 \times 0.01 + 0.02 \times 0.99 = 0.297$$

(ce qui correspond à la probabilité d'être positif si on présente le trouble auquel s'ajoute la probabilité d'être positif si on ne présente pas le trouble).

Ce qui permet de calculer $p(A|B)$

$$P(A|B) = 0.99 \times 0.01 / 0.297$$

$$P(A|B) = 0.33$$

Bonne Réponse : la probabilité d'avoir le trouble est de 33%.

Être positif ou négatif ne permet pas d'estimer, connaissant uniquement la spécificité et la sensibilité, la probabilité que le trouble soit présent ou non. Spontanément vous auriez probablement répondu que la probabilité d'avoir le trouble si le test est positif est bien supérieure à 0.33 (33%). En fait, si la prévalence est faible, malgré une bonne spécificité et sensibilité, en cas de test positif, la probabilité d'avoir le trouble reste faible. Si la prévalence était de 2%, la probabilité d'avoir le trouble passerait à .50 (50%). A l'inverse si la prévalence est de 1‰, le calcul donne la valeur de 4.17 (soit 4,17% de "chance" seulement d'avoir réellement le trouble si le test est positif !).

La prise en compte de la **prévalence** de la maladie est donc essentielle pour interpréter la spécificité et la sensibilité des tests.

1.2.4 Les valeurs prédictives (VPP et VPN)

L'interprétation des indices de sensibilité et spécificité doit donc toujours tenir compte de la prévalence dans la population de la caractéristique étudiée.

Deux indicateurs, fréquemment cités, devraient donc toujours accompagner la sensibilité et la

spécificité. Il s'agit de deux indices : la valeur prédictive positive (VPP) et la valeur prédictive négative (VPN). Ils indiquent la probabilité de classer correctement une personne quand le test est positif (VPP) ou négatif (VPN) et prennent en compte la prévalence du trouble ou de la maladie dans la population.

Pour calculer ces valeurs prédictives, on utilise le théorème de Bayes. Le calcul de VPP correspond à $p(A|B)$ (A = avoir le trouble ou la maladie et B = le test est positif). Le calcul de VPN correspond à $p(\neg A|\neg B)$ ($\neg A$ = ne pas avoir le trouble ou la maladie et $\neg B$ = le test est négatif). Ces valeurs changent en fonction de la prévalence qui peut être connue ou estimée.

Un **calculateur** de ces valeurs est aussi disponible <[calculateur VPP VPN](#)>

Une petite application web est disponible <[ICI](#)> pour visualiser l'effet de la prévalence sur les valeurs prédictives positives et négatives.

Pour aller plus loin.

Il existe d'autres indicateurs comme :

- **Le rapport de vraisemblance positif** ($RVP = \text{sensibilité} / (1 - \text{spécificité})$) mesure la vraisemblance d'avoir un test positif si on est effectivement positif. Il varie de 0 à plus l'infini. Plus il est élevé, plus le « gain diagnostic » est important. En général la valeur de RVP doit être supérieure à 10. Une valeur entre 5 et 10 est encore correct et discutable entre 2 et 5.
- **Le rapport de vraisemblance négatif** ($RVN = (1 - \text{sensibilité}) / \text{spécificité}$) mesure la vraisemblance d'avoir un test négatif si on est effectivement négatif. Plus il est proche de 0, plus il permet d'exclure le diagnostic. En général la valeur de RVN doit être inférieure à 0.10. Une valeur entre 0.10 et 0.20 est encore correct et discutable entre 0.20 et 0.50.

Une façon de résumer la valeur d'un test consiste à calculer le rapport RVP/RVN . Le test est jugé utile si ce rapport est au moins supérieur à 50.

2. Homogénéité et dimensionnalité

Le test psychologique ou test mental est souvent censé être une mesure d'une variable latente qualitativement semblable pour tous les individus, ceux-ci se différenciant sur cette variable. Une des hypothèses à la base de la construction des tests est donc que l'ensemble des items ou des questions qui le sous-tendent mesure une seule chose (une épreuve peut-être multidimensionnelle mais elle est alors composée de plusieurs sous-ensemble d'items respectant chacun cette condition d'unidimensionnalité).

Cette propriété unidimensionnalité à la base des tests a initialement été utilisée dans un sens similaire à celui d'homogénéité des tests mais ce dernier a plusieurs contexte d'utilisation et plusieurs sens. Ces deux termes ne sont donc pas identiques mais il serait trop long, ici, d'introduire une discussion bien introduite par Hattie en 1985. Nous présenterons donc succinctement une facette du concept d'homogénéité et des éléments concernant l'unidimensionnalité.

2.1. Homogénéité

Le terme homogénéité a été historiquement utilisé dans deux acceptations : soit comme synonyme d'unidimensionnalité, soit comme un indicateur de relations cohérentes entre items. Dans ce cadre, l'évaluation de l'homogénéité n'est pas forcément la même dans le cas d'une échelle ordinale que dans celui de l'échelle d'intervalle.

Échelle ordinale

L'homogénéité par implication : des items de difficultés différentes sont dits homogènes si l'on ne peut pas réussir les plus difficiles sans réussir les plus faciles (toutes les personnes résolvant une question de difficulté p parviennent à résoudre une question de difficulté moindre). Un test conforme à cette propriété (échelle de [Guttman](#)), n'est bien sûr pas forcément homogène (mais le contraire est vrai : une échelle homogène respecte cette propriété).

L'homogénéité par implication est nécessaire car si elle n'existait pas cela signifierait qu'un score total (somme des scores aux items) pourrait être obtenu avec des patterns de réponses très différents (ce qui nuirait à la signification que l'on pourrait lui attribuer). Par exemple, si cette propriété n'est pas vérifiée sur une échelle en 10 items (items dichotomiques notés 1 ou 0), un même score de 3 peut-être obtenu par certains en réussissant les 3 items les plus difficiles et en échouant tous les autres alors que ce même scores pourraient être réussis avec 2 items faciles et un item de difficulté moyenne.

La technique statistique privilégiée pour mettre en évidence l'implication entre les questions est l'analyse hiérarchique proposée initialement par Guttman (1944) en psychologie sociale pour la mesure des attitudes. Si plusieurs questions (où le sujet doit répondre par oui ou par non, d'accord ou pas d'accord) sont supposées a priori exprimer une même attitude avec des degrés différents d'intensité, (par exemple, attitude à l'égard d'étrangers ou d'une religion), on doit pouvoir ordonner les réponses et on doit constater qu'une personne qui donne une réponse positive à une question d'intensité x doit avoir donné une réponse positive à toutes les questions d'intensité inférieure. S'il n'en était pas ainsi, l'échelle ne peut pas être considérée comme homogène ou unidimensionnelle.

L'homogénéité par équivalence : quand plusieurs questions sont de même difficulté, toutes les personnes résolvant une de ces questions doit réussir les autres questions de même difficulté.

Remarque : pour une épreuve donnée, on n'a pas obligatoirement qu'un seul item par niveau de difficulté. Plusieurs items peuvent donc être équivalents et la hiérarchie (homogénéité par implication) existe alors uniquement entre groupes d'items. L'homogénéité des items peut être mise en évidence par des techniques comme l'analyse hiérarchique (coefficient d'homogénéité de Loewinger, échelle hiérarchique de Guttman ou de Lazarfield).

Échelle d'intervalle

Dans ce type d'échelle, le test est considéré comme homogène si les différents items demandent des capacités équivalentes pour être réussis. On peut mettre en évidence ce type d'homogénéité par le calcul des [corrélations](#) entre items ou encore par l'utilisation de technique comme [l'analyse factorielle](#).

2.2. Unidimensionnalité

La notion d'unidimensionnalité est une question complexe en psychométrie. En principe on parle d'unidimensionnalité lorsque chaque item ne dépend que d'une seule dimension (une seule variable latente). On considère cependant, qu'à côté de cette définition stricte de l'unidimensionnalité, on se doit de considérer une définition plus "lâche" (unidimensionnalité essentielle ou dominante). En effet, les processus mis en œuvre lors de la réalisation d'une tâche ou les facteurs déterminant une réponse (pour un questionnaire) peuvent non seulement varier en fonction du contexte et des personnes mais il est aussi très probable que la réponse à une question n'implique pas qu'un seul trait latent. On parle donc d'unidimensionnalité essentielle lorsque qu'une variable latente domine pour expliquer les réponses aux items (Stout, 1987).

Au delà de cet aspect, discuté dans la littérature, qui fait de la dimensionnalité une notion parfois ambiguë ("fuzzy" en anglais), les méthodes qui existent pour déterminer le nombre de dimensions (e.g. la dimensionnalité d'une épreuve) sont nombreuses et fonction des modèles de mesure (Hayti, 1985, Tate, 2003). En règle général les plus fiables s'appuient sur des techniques comme l'analyse factorielle exploratoire ou l'analyse en composantes principales des résidus d'un modèle mais il n'existe pas d'approche unifiée ou faisant l'unanimité. Haiti en 1985 recensait l'utilisation de 30 indices d'unidimensionnalité (regroupés en 5 grande catégories). Certains indices (parfois très utilisés) apparaissent comme très insatisfaisants et aucun semblait totalement approprié. Il faut peut-être considérer que cette recherche d'indices est illusoire et la question à se poser est non pas de rechercher l'indice "absolu" mais un ensemble de critères qui permettent de s'assurer qu'il existe une variable latente dominante.

Actuellement, on peut cependant recommander (Levy & Roy, 2014) soit l'utilisation des analyses factorielles non linéaires ou dans une approche dite d'unidimensionnalité essentielle des procédures comme DIMTEST (Nandakumar & Stout, 1993) ou, pour les items polytomiques, PolyDIMTEST (cf. pour ceux qui veulent aller plus loin, Kieftenbeld, & Nandakumar, 2015).

Remarques :

- Il ne faut pas oublier que la dimensionnalité est une propriété conjointe de l'ensemble d'items et d'un échantillon particulier de sujets. Selon l'échantillon, cette unidimensionnalité pourrait être présente ou non (par exemple, les personnes doivent être différentes sur le trait latent pour que l'on puisse vérifier l'unidimensionnalité, ou encore selon les groupes de personnes (âge par exemple) les processus impliqués sont différents et pour certains on peu observer le respect de l'unidimensionnalité de la mesure et pas pour d'autres).
- Certains indices de [fidélité](#) comme le coefficient alpha, (ou plus généralement des indices [de consistance interne](#)), sont parfois (trop souvent et à tort) utilisés pour assurer de l'unidimensionnalité d'une épreuve. En aucune façon (sauf dans des cas particuliers) ils ne permettent d'assurer l'unidimensionnalité d'une épreuve.

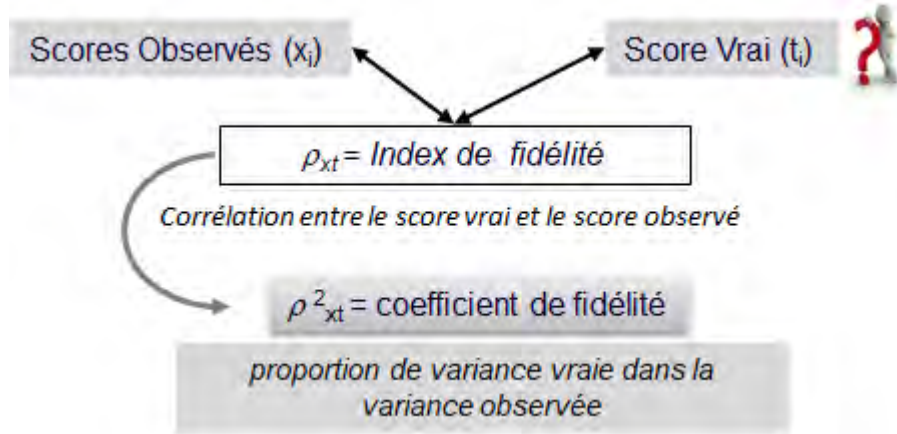
3. Fidélité

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).

3.1. Définitions

Pour comprendre le concept de fidélité et ses différentes méthodes d'études, nous nous inscrivons dans le cadre de la [théorie classique des tests](#). Dans ce cadre, une mesure observée X , peut être décomposée en deux sources T et ε : $X = T + \varepsilon$ (rappel : T est la quantité représentant le score vrai et ε l'erreur de mesure). On appelle, **index de fidélité** la corrélation existante entre les scores observés (X) et les scores vrais (T). Le **coefficient de fidélité** (ρ^2_{TX}) est égal au carré de l'index de fidélité.

Le coefficient de fidélité est donc (cf. pré-requis : [variance expliquée](#)) le rapport entre la variance de T dans la population (qui évaluent l'amplitude des différences réelles entre les individus) et la variance de $T + \varepsilon$ (qui correspond à la variance de X observée donc l'amplitude des différences observées). En d'autres termes, c'est la proportion de variance des scores observés imputable à la variance des scores vrais (des différences réelles entre individus).



Le calcul du coefficient de fidélité peut paraître impossible car, si on ne connaît pas les scores X , on ne connaît pas les scores vrais (par définition). Dans le cadre de la théorie classique des tests, le coefficient de fidélité (ρ^2_{TX}) peut cependant être estimé en calculant par exemple la corrélation entre deux séries de mesures prises sur les mêmes individus ([sous conditions](#)). Cette estimation est souvent notée r_{xx} (ce qui parfois prête à confusion car il s'agit bien d'une simple corrélation qui est une estimation du coefficient de fidélité qui est lui le carré d'une corrélation).

Cette propriété a conduit à développer différentes méthodes de mesure de la fidélité comme la [méthode du test re-test](#), la [méthode des tests parallèles](#), la [méthode du partage](#), les méthodes s'appuyant sur la [consistance interne](#), etc. (Revelle, & Condon, 2018). Il existe une quatrième méthode distincte des précédentes. Cette méthode dite [méthode inter-juges \(ou accord inter-juges ou encore accord inter-cotateurs\)](#) est utilisée quand l'erreur de mesure à apprécier a pour origine "la difficulté de cotation" (contextes particuliers comme certaines épreuves de personnalité).

A savoir

- La fidélité consiste à estimer la part [des facteurs aléatoires](#) dans la mesure. Un coefficient de fidélité n'est pas suffisant pour interpréter ou même garantir l'existence d'une dimension ou de la [validité](#) d'un test. Un test peut-être fidèle mais non valide (on mesure quelque chose mais on ne mesure pas ce que l'on voulait mesurer !).
- La fidélité est le rapport entre la variance vraie (due à un ou plusieurs facteurs de différenciation des sujets) et la variance observée dans le test (on la note souvent r_{xx}).
 - le coefficient de fidélité varie entre 0 et 1.
 - si le coefficient de fidélité est de .80, cela signifie que 80% de la variance observée est de la variance vraie et 20% de la variance d'erreur.
 - plus la fidélité est grande plus, plus l'erreur de mesure est faible. Une bonne fidélité assure donc que le test mesure quelque chose.
- Plusieurs méthodes permettent d'évaluer la fidélité et ces méthodes évaluent l'importance de l'erreur de mesure mais ces méthodes n'évaluent pas nécessairement la même [source d'erreur de mesure](#).
- Il ne faut pas confondre indice (ou index) de fidélité (corrélation entre score vrai et score observé) et coefficient de fidélité (qui est le carré de l'indice de fidélité). C'est toujours le coefficient de fidélité qui est reporté dans les manuels.

ATTENTION / IMPORTANT : dans de nombreux ouvrages, sur internet ou lors d'interventions orales, on

donne comme définition de la fidélité, la constance ou la reproductibilité des scores d'un test (ce qui fait que les IA conversationnelles se tromperont aussi si vous demandez une définition !). Constance ou reproductibilité, ne sont pas des définitions de la fidélité mais des manifestations d'un score fidèle (*i.e.* avec une erreur aléatoire de mesure minimale). Malheureusement actuellement c'est parfois la seule que l'on donne alors que la définition à retenir, nous le rappelons, est que **la fidélité d'un test est le degré auquel les scores observés sont exempts d'erreurs de mesure aléatoire ou encore** "Reliability refers to the precision of measurement (or degree of error) in an instrument." ([AERA, APA, & NCME, 2014](#)).

Pour aller plus loin

Il est toujours surprenant de calculer un coefficient estimant la relation qui existe entre un score observé et un score vrai que l'on ne connaît pas. Il est indiqué ci-dessus que ce coefficient est estimé à partir de deux formes parallèles d'un test. Dans la théorie classique des tests (TCT), deux formes d'un test sont dites parallèles si leurs scores vrais et l'erreur type de mesure sont égales. A partir de ce postulat, on peut démontrer que la corrélation entre deux tests parallèles est une estimation du coefficient de fidélité (donc une estimation du carré de l'index de fidélité). Pour ceux que la démonstration intéresse, ils peuvent se reporter à la page 107 de l'ouvrage de Laveault et Grégoire (2014).

3.2. Erreur systématique - Erreur aléatoire

De façon générale, l'erreur de mesure correspond à l'écart existant entre la valeur réelle que l'on veut mesurer (inconnue) et la valeur mesurée. On doit cependant distinguer deux types d'erreurs.

Erreur systématique.

Le premier type d'erreur est ce qu'on appelle l'erreur systématique. Cette erreur est une "déviation" constante, négative ou positive introduit par l'instrument. De façon plus générale on parle d'erreur systématique lorsque, par rapport à une valeur de référence x , l'instrument donnera toujours comme valeur observée $x+b$ (déviation positive ou négative). Par exemple, pour un instrument comme votre balance, si elle affiche systématiquement "+ 2 kilogrammes" par rapport au poids réel, l'erreur systématique est de +2 kg. Dans les tests mentaux une des causes possibles de cette erreur systématique est le biais d'échantillonnage lors de [l'étalonnage](#) de l'épreuve.

Erreur aléatoire.

Dans la construction des tests et l'analyse de la fidélité, quand on parle d'erreur de mesure, on fait référence à ce qu'on appelle l'erreur aléatoire. Cette erreur est le résultat d'un ensemble de facteurs (inconnus) qui font que parfois la mesure sera légèrement supérieure à la valeur réelle et parfois légèrement inférieure. Un instrument de mesure est toujours construit pour minimiser cette erreur aléatoire (la mesure observée doit être toujours proche de la mesure de référence ou plus exactement la dispersion autour de cette valeur de référence, lors d'observations multiples, est faible). **Cette erreur aléatoire est celle qui est associée à la notion de fidélité** et celle à laquelle on fait le plus souvent référence lorsque l'on parle d'erreur de mesure dans la construction des tests mentaux.

Remarques :

- Lors d'une opération de mesure, ces deux erreurs s'additionnent mais dans le cadre de la [théorie classique des tests](#), ($x = T + e$) l'erreur systématique est confondue avec T (le score vrai). L'erreur systématique affecte donc la [validité de la mesure](#) alors que l'importance de l'erreur aléatoire est en relation avec la [fidélité](#) d'une épreuve.

- Si l'on répète une mesure et qu'on calcule la moyenne de ces mesures, l'effet de l'erreur systématique reste identique sur la moyenne mais, à l'inverse, l'effet de l'erreur aléatoire sur la moyenne diminue (en effet parfois l'erreur de mesure augmente la valeur et parfois la diminue et la moyenne de cette erreur tend vers 0).


3.3. Sources de l'erreur aléatoire de mesure

Un test sera fidèle si l'on minimise l'erreur de mesure aléatoire par rapport à la variance totale, c'est à dire si l'on s'assure que les différences interindividuelles ne sont pas attribuables ou sont très peu attribuables à une erreur aléatoire de mesure. Il est donc important de cerner les sources possibles de l'erreur de mesure. Pour résumer les quatre principales sources d'erreurs sont :

- l'erreur engendrée par l'instrument (le test) lui-même ;
- l'erreur engendrée par les variations de conditions de passation du test (cette erreur devrait être minimale) ;
- l'erreur engendrée par des facteurs associés aux répondants au test ;
- l'erreur engendrée par les cotateurs.

La première catégorie d'erreur comprend des facteurs contrôlables et fait l'objet de l'attention des psychologues qui construisent les tests. La deuxième catégorie d'erreur et la quatrième catégorie d'erreur sont aussi contrôlables et concernent les procédures de passation du test. Elle justifie que ces procédures soient parfaitement standardisées et que le psychologue professionnel suive parfaitement les instructions, le minutage et les consignes (de passation comme de cotation).

La troisième catégorie d'erreur est bien plus difficile à contrôler et la liste des sources d'erreurs associées aux répondants est très longue. Les plus fréquentes sont, la motivation, l'anxiété, l'habitude de passer des tests et les variables d'ordre physiologique (fatigue, concentration). Toutes ces variables doivent donc être prises en compte lors de l'analyse non pas quantitative des résultats mais qualitative. Les résultats d'un test demandent donc une interprétation à intégrer dans une démarche clinique plus générale. C'est ce que prônait déjà Binet (pour des raisons partiellement différentes) pour le premier test d'intelligence (Binet, Simon, 1908).



A. Binet

Mon test, «*n'est pas une machine qui donne notre poids imprimé sur un ticket comme une bascule de gare*». Les résultats ont besoin d'être analysés, situés dans un contexte, interprétés. L'échelle métrique «est un instrument qu'on ne doit pas mettre entre les mains d'un imbécile».

Quelle est l'intelligence d'un enfant ayant un âge mental de 6 ans 4/5 et d'âge chronologique 5 ans 1 mois ? **« Qu'est-ce-que cela veut dire ? Cela veut dire que dans les conditions où nous venons de l'examiner, cet enfant s'est comporté comme un enfant de 6 à 7 ans. Placé en face de difficultés que représente la série d'épreuves auxquelles nous l'avons soumis, il paraît avoir les mêmes moyens de les résoudre qu'un enfant instruit par une expérience de 6 à 7 années, bien qu'il n'ait que 5 ans d'âge. Cela ne veut pas dire autre chose. Nous apprécions un degré de développement, nous en préjugeons qu'hypothétiquement de ses causes - ou plus exactement nous en préjugeons d'après les autres informations que nous possédons à ce sujet. Un niveau d'intelligence est un résultat qui se doit d'être commenté... »**

source, Binet, Simon (1908)

3.4. Méthodes pour évaluer la fidélité

Le coefficient de fidélité est un coefficient que l'on peut estimer en utilisant différentes techniques (ou méthodes) qui n'évaluent cependant pas exactement de la même façon les sources de l'erreur de mesure. Nous ne présentons pas ici toute les méthodes d'étude de la fidélité mais celles qui sont les plus utilisées même si certaines sont très critiquables (cf. pour aller plus loin Revelle & Condon (2018)).

3.4.1 Postulats

Les méthodes d'étude de la fidélité s'inscrivant dans le cadre de la TCT reposent sur l'acceptation de postulats fondamentaux. En plus de ceux déjà mentionnés (l'équation $X = T + e$, et le postulat concernant les formes parallèles d'un test (2 tests sont parallèles si leurs scores vrais et leur erreurs de mesure sont identiques) trois postulats additionnels sont formulés :

- **Postulat 1** : $\mu_e = 0$. La moyenne des erreurs commises aux différents items d'un test est nulle. Autrement dit, les erreurs se compensent mutuellement, ce qui signifie qu'il n'existe pas de biais systématique (l'espérance mathématique de X est égal à T).
- **Postulat 2** : $\rho_{Te} = 0$. La corrélation entre les scores vrais et les scores d'erreur vaut zéro. Il n'existe donc pas un mécanisme qui conduirait à accroître ou à réduire l'ampleur des erreurs en fonction de la valeur du score vrai.
- **Postulat 3** : $\rho_{e1e2} = 0$. La corrélation entre les erreurs associées aux différents items est nulle. Cela signifie que les erreurs de mesure sont indépendantes les unes des autres. On n'observe donc pas des erreurs d'autant plus grandes à certains items qu'elles le sont à d'autres.

3.4.2 Test-retest

La méthode du test-retest 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. On calcule ensuite la corrélation entre les

performances observées lors de la première puis de la seconde passation. Cette corrélation est une estimation du coefficient de fidélité. Ce coefficient de fidélité associé à cette méthode est parfois appelé « 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.

Cette méthode est coûteuse dans sa mise en oeuvre et parfois impossible (pour des raisons d'effet de test-retest, d'apprentissage ou autres).

3.4.3 Tests 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 (on se retrouve alors dans le cas du test-retest !).

Remarques

L'hypothèse de base de cette méthode est qu'un test doit mesurer une dimension relativement indépendante des situations. Donc, si l'on construit une forme A d'un test, on doit pouvoir construire par la même méthode de construction un test mesurant la même dimension avec d'autres items (forme B). S'il n'y a pas de corrélation entre ces deux formes différentes, c'est qu'on ne peut pas faire confiance à ce que mesure le test. La possibilité de construire une forme parallèle est une garantie que l'on maîtrise ce que l'on construit.

Cette méthode est coûteuse et exige beaucoup de temps et deux formes parallèles ne sont jamais équivalentes à 100 %. L'erreur de mesure peut être surestimée.

Pour aller plus loin

On peut combiner la méthode de test-retest (qui teste aussi la stabilité de la mesure) et la méthode des tests parallèles (parfois appelée méthode d'équivalence) pour définir une nouvelle méthode "stabilité-équivalence". Le principe consiste à évaluer deux formes différentes d'un test (méthode des tests parallèles) à deux moments différents (méthode test-retest). Les deux premières méthodes n'évaluent pas tout à fait les mêmes sources de l'erreur de mesure. En les combinant, la valeur trouvée est donc souvent plus faible. Cette méthode est très rarement utilisée.

3.4.4 La méthode du partage

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).

Calcul

La méthode de calcul du coefficient de fidélité par la méthode du partage est simple :

1. On calcule le score pour chaque groupe d'items (par exemple : pairs et impairs)
2. On calcule la corrélation r_{12} entre ces scores. Cette corrélation est une estimation de la fidélité r_{xx}
3. Pour tenir compte que l'on a réduit la longueur du test par deux on doit appliquer la [formule de Spearman-Brown](#) :

$$r_{xx} = 2 * r_{12} / (1 + r_{12})$$

Une autre méthode de calcul est ce qu'on appelle la **formule de Rulon** (plus rarement utilisée). Elle donne le même résultat que la corrélation entre les deux groupes d'items corrigée avec la formule de Spearman-Brown. Cette formule consiste à estimer l'erreur en rapportant la variance des différences (s_d^2 qui est directement une évaluation de l'erreur de mesure) à la variance du test (s_x^2). Le coefficient de fidélité devient alors :

$$r_{xx} = 1 - (s_d^2 / s_x^2)$$

Inconvénients de cette méthode : le coefficient obtenu va être différent selon la méthode de bissection utilisée et le nombre de bissection possible explose très rapidement avec le nombre des items (126 bisections possibles pour 10 items puis 92378 pour 20 items). La formule générale concernant le nombre de bisections :

$$k = \frac{C_n^2}{2} = \frac{n!}{2 \times \left(\frac{n}{2}\right)!^2}$$

Remarque : comme pour les deux méthodes précédente, la méthode du partage s'appuie sur le postulat de la TCT concernant les formes parallèles des tests. Avant de calculer le coefficient de fidélité et après avoir séparé (quelle que soit la procédure) les deux groupes d'items, on devrait s'assurer que les moyennes et les variances sur les deux parties du test sont similaires. Dans le cas contraire, l'estimation de la fidélité risque d'être incorrecte. Les simulations sur des jeux de données montrent que la valeur du coefficient peut varier de façon significative selon la partition utilisée. La partition en deux moitiés peut donc engendrer une erreur d'estimation de la fidélité (le hasard peut mal faire les choses !).

Il existe des méthodes (s'appuyant sur la covariances entre items) qui s'affranchissent de la méthode de partage pour mesurer la fidélité (pour simplifier, si un certain nombre de conditions sont respectées, ils fonctionnent comme si on effectuait tous les partages possibles et que l'on faisait la moyenne des coefficients obtenus). Ces méthodes cherchent à évaluer plus directement la consistance interne (cf. le paragraphe suivant).

3.4.5 Consistance interne

(a) Alpha de Cronbach

Les méthodes d'estimation de la fidélité s'appuyant sur la consistance interne sont différentes de [la méthode du partage](#) qui en reste une mesure très indirecte. En effet, la partition en deux moitiés peut engendrer une erreur d'estimation de la fidélité. Pour résoudre ce problème, différentes méthodes, selon que les items sont dichotomiques ou non, sont souvent utilisées :

⇒ **Le KR20 (Kuder Richardson)** : KR20. Ce coefficient ne s'applique qu'aux items dichotomiques.

$$KR20 = \left[\frac{n}{n-1} \right] \cdot \left[1 - \frac{\sum p_i q_i}{s^2} \right]$$

n = nombre de questions du test
 s^2 = variance observée au test (sur le score global)
 p_i = proportion de réussites à l'item i
 q_i = proportion d'échecs à l'item i

Pour l'anecdote : le terme de KR-20 vient du nom des auteurs de l'article (Kuder-Richardson) pour les lettres et le 20 fait référence au numéro de la formule dans l'article originale paru en 1937 présentant cet indice.

⇒ **Le KR21** : Le KR21 suppose que tous les items ont le même niveau de difficulté !. Ce coefficient ne s'applique qu'aux items dichotomiques.

$$KR21 = \frac{n}{n-1} \cdot \left[1 - \frac{m \cdot (n-m)}{n \cdot s^2} \right]$$

n = nombre de questions du test
 s^2 = variance observée au test (sur le score global)
 m = moyenne observée au test (sur le score global)

⇒ **L'alpha de Cronbach** : similaire au KR20 mais concerne des items non dichotomiques.

$$\alpha = \frac{n}{n-1} \cdot \frac{s_t^2 - \sum s_i^2}{s_t^2}$$

n = nombre de questions du test
 s_t^2 = variance observée au test (sur le score global)
 s_i^2 = variance observée à l'item i

Remarques :

- Si et seulement si les items sont tau-équivalents*, le coefficient alpha représente la moyenne de tous les coefficients de bissection (méthode du partage) possible. On devrait éviter de l'utiliser sauf dans les cas où son utilisation est parfaitement justifiée. (Cho 2016, Mcneish & Mcneish 2017, Flora 2020).

(*) dans la TCT, deux tests ou deux items sont tau-équivalents (τ -équivalent) si et seulement si les scores vrais diffèrent par une constante ($Test1 = V + \varepsilon_1$; $Test2 = V + K + \varepsilon_2$)

- Il importe de mentionner que ces coefficients sont plutôt conservateurs. Par exemple, si le coefficient alpha est probablement le coefficient le plus connu et le plus utilisé, il repose sur l'hypothèse que chaque item est "parallèle" aux autres, dans le cas contraire (le plus fréquent) il sous-estime la consistance interne.
- Le coefficient alpha n'est pas une mesure de l'homogénéité du test ni de l'unidimensionnalité du

test. Il indique que le test mesure quelque chose (lorsqu'il est élevé) mais pas quoi (ce peut-être plusieurs choses !). En fait, plus le nombre d'items est important, plus le coefficient alpha va avoir tendance à augmenter (si les items corrélerent un minimum 2 à 2). Donc il est facile d'augmenter la valeur de ce coefficient en augmentant le nombre des items même si ceux-ci mesurent des aspects différents. Un coefficient alpha élevé, contrairement à ce qu'on affirme souvent, ne garantit par l'unidimensionnalité ou l'homogénéité (Laveault, 2012).

- On présente parfois l'estimation de la fidélité avec le coefficient alpha comme une méthode des covariances. On pourrait être surpris de cette expression car dans la formule (cf. ci-dessus) il n'y a pas de covariances. En fait, celles-ci sont bien présentes, puisque la variance observée à un test (s_t^2 dans la formule) est égale à la somme des variances des items plus deux fois la somme des covariances (entre les items pris 2 à 2). Dans la formule, on divise (numérateur) la variance du test moins la somme des variances observés aux items constituant le test par (dénominateur) la variance du test. Le numérateur est donc égal à deux fois la somme des covariances des items entre eux (2 à 2). En fait, plus la covariance entre les items va être élevée au regard de la variance du test, plus la consistance interne est élevée.

Pour aller plus loin...

Selon le modèle de mesure (unidimensionnel ou non, tau-équivalent, congeneric, parallèle) la formule de calcul du coefficient de fidélité doit être différente. Dans de nombreuses publications on utilise le coefficient alpha de Cronbach à tort (échelle non unidimensionnelle ou modèle non tau-équivalent). Si vous voulez vraiment aller plus loin, que cette simple présentation, voir à ce sujet l'article de [Cho \(2016\)](#) sur l'usage et le mésusage de ces coefficients. Cho (page 667) propose une nouvelle dénomination de ces coefficients en fonction du modèle de mesure utilisé (nous n'avons pas repris ces dénominations pour simplifier car il demande de bien connaître tous les coefficients habituellement utilisés).

Guttman (1945) a proposé 6 mesures différentes pour estimer la limite inférieure de la fidélité (coefficients lambdas : λ_1 à λ_6). Le coefficient λ_3 est similaire à l'alpha de Cronbach. Tous ces indices sont aussi basés sur des hypothèses plus ou moins restrictives (comme λ_3 qui suppose que les items soient tau-équivalents).

(b) Omega (ω)

Historiquement, l'alpha de Cronbach (α) a été la mesure de référence pour la fidélité (consistance interne). Les conditions d'application pour cet indicateur sont cependant rarement respectées dans la pratique (*modèle non tau-équivalent, échelles multidimensionnelles*). D'autres solutions ont été développées. Le coefficient oméga de McDonald (ω) est une alternative plus robuste (Watkins, 2017) et surtout permet des estimations plus précises concernant les échelles multidimensionnelles.

Principe

Le calcul est plus complexe et aucune formule ne sera donnée dans cette introduction. La mise en oeuvre nécessite d'avoir des hypothèses sur les relations entre scores observés sur les items et les facteurs sous-jacents "expliquant" ces scores. Le calcul de ces coefficients (il y en a plusieurs) suppose l'utilisation de l'analyse factorielle qui seront présentés en fin de ce manuel. Si l'on prend l'exemple d'une échelle multidimensionnelle, et d'un simple modèle hiérarchique (le plus fréquent), la variance totale du test se décompose ainsi (Revelle, 2016) :

- ✓ un facteur général (variable latente) source d'une variance commune à toutes les variables mesurées ;

- ✓ des facteurs de groupe (variables latentes communes à certaines variables ou items mais pas à toutes). Ces facteurs sont sources d'une variance commune aux items ou variables des sous-échelles) ;
- ✓ des facteurs spécifiques avec une variance unique à chaque variable mesurée ;
- ✓ une erreur aléatoire

Les deux dernières sources de variation ne sont pas différenciées dans les modèles et sont sources de ce qui constitue la variance résiduelle. On utilise aussi, quand on les regroupe, d'un facteur d'unicité ("uniqueness" en anglais). Plusieurs variantes du coefficient ω peuvent être calculées.

Une famille de coefficients Omega

Omega Total (ω_T) : ω_T est similaire à α (sans les défauts) et peut être interprété de la même manière. Des valeurs ω élevées indiquent un composite multidimensionnel global fidèle mais cet indicateur ne permet pas de différencier la précision des scores totaux et des scores des sous-échelles.

Omega scale (ω_s) : dans une échelle multidimensionnelle on peut calculer ω pour chacune des sous-échelles à l'aide de la même logique de calcul que ω_T . Il s'agit alors de la proportion de variance totale d'une sous-échelle attribuable à la fois au facteur général et au facteur de groupe associée à cette échelle). Cet indicateur s'interprète de la même façon que le coefficient α mais ne permet toujours pas de distinguer la précision du facteur général et du facteur de groupe.

Oméga hiérarchique (ω_H) : Lorsque les échelles sont multidimensionnelles (c'est-à-dire qu'elles mesurent plus d'un facteur), ω_H fournit une estimation de la fidélité du facteur général. Il s'agit donc de la proportion de la variance attribuable à un facteur général. Une valeur supérieure à .50 est acceptable. Une valeur élevée signifie que le facteur général est la source de variance dominante. A l'inverse, une valeur faible indique que les facteurs de groupe et/ou le facteur d'unicité (facteur systématique+facteur aléatoire) sont la source dominante de variance.

Omega hiérarchique pour les sous-échelles (ω_{H_s}). Appliquée aux échelle ω_{H_s} représente la proportion de la variance du score de l'échelle qui est expliquée par le facteur de groupe associé à cet échelle. Il doit être comparé à ω_s . Par exemple, une valeur faible indique, si ω_s est élevé, que la majeure partie de la variance de cette échelle est due au facteur général. Le score à cette échelle n'est donc pas un indicateur sans ambiguïté de la variable que l'on voulait mesurer.

Pour conclure

L'analyse de ces coefficients peut parfois déborder le cadre de la fidélité, et pour le praticien psychologue lui permettre d'avoir une information sur la validité de sa pratique. En effet, lorsque l'on utilise une échelle multidimensionnelle et qu'au delà du score général, ce sont les sous-dimensions qui intéressent, il faut comparer ω_s et ω_{H_s} car selon les valeurs observées, au delà de l'information sur la fidélité, on peut s'interroger sur la pertinence de cette sous échelle si ω_{H_s} est très inférieur à ω_s .

On tient enfin à souligner que le développement des programmes statistiques comme le logiciel R (R Core Team, 2025) facilite l'utilisation de ce coefficient et permettent actuellement une utilisation plus

régulière en lieu et place de l'alpha de Cronbach. Malheureusement cette pratique n'est pas encore généralisée en 2025 dans les manuels des tests psychologiques.

Pour aller plus loin

Il existe d'autres coefficients que le coefficient ω . Par exemple, le **coefficient H de Hancock et Mueller (2001)**. C'est un indicateur utilisé principalement en analyse factorielle confirmatoire (AFC) ou en modélisation par équations structurelles (SEM). Alors qu'un coefficient hiérarchique oméga représente la corrélation entre un facteur et un score composite, H indique dans quelle mesure une variable latente particulière est représentée par ses indicateurs. Il est considéré comme une mesure de la fidélité ou de la reproductibilité du concept. Lorsque sa valeur est faible, la variable latente n'est pas bien définie par ses indicateurs et il aura tendance à être instable d'une étude à l'autre.

La multiplicité des indicateurs peut paraître source de confusion. Elle permet cependant de relativiser les valeurs trouvées et de s'interroger sur les mesures. Tous les manuels de test devraient proposer plusieurs indicateurs de fidélité à leur lecteur.

3.4.6 Accord inter-juges

Cette technique, différente des précédentes, est utilisée dans les cas où il peut y avoir ambiguïté dans l'évaluation (cotation) des résultats au test, évaluation qui peut être entachée de subjectivité (par exemple : certaines mesures utilisant des techniques projectives). La principale source de l'erreur de mesure étant le coteur, la mesure de la fidélité consiste à évaluer s'il existe un degré d'accord suffisamment élevé entre les jugements de plusieurs observateurs.

Plusieurs indicateurs statistiques permettent d'évaluer l'accord inter-juge, mais contrairement aux autres indices de fidélité, il n'y a pas de consensus véritable sur l'interprétation de ces coefficients qui font parfois l'objet de critiques importantes (pour le kappa de Cohen, cf. par exemple le billet de Stéphane Vautier "[Le kappa de Cohen : une solution à un faux problème](#)").

Pour ceux que cela intéresse, on peut aussi utiliser des coefficients de corrélations intraclasse (Shrout & Fleiss, 1979). Cette méthode se base sur les résultats d'une analyse de variance prenant en compte comme sources de variation, le facteur sujet (facteur aléatoire), le facteur juge et l'interaction entre ces deux facteurs. Ce coefficient s'interprète comme les autres coefficients de fidélité.

Deux exemples de coefficients :

- **Le κ (kappa) de Cohen.** Ce coefficient proposé en 1960 par Cohen est destiné à mesurer l'accord inter-juge pour une variable qualitative (échelle nominale ou ordinale). Ce coefficient est compris entre -1 et +1 et le plus souvent on considère que l'accord est moyen entre 0.40 et .60. Il est satisfaisant à partir de .60 et excellent pour plus de .80.

Cette grille de lecture ne fait cependant pas consensus car le nombre de catégories de l'échelle utilisée influe sur la valeur du coefficient. Par ailleurs, ce coefficient ne fonctionne que si il y a deux juges uniquement. La formule du Kappa de Cohen peut-être adaptée si l'échelle est ordinale et non nominale (Kappa de Cohen dit pondéré) de façon à donner plus d'importance à l'erreur introduit par des jugements distants (= très différents) qu'à des jugements proches.

- **Le κ (kappa) de Fliess.** Ce coefficient introduit dans les années 80 par Joseph L. Fliess est utilisé

lorsqu'il y a plus de deux observateurs ou cotateurs. Il est utilisable cependant uniquement pour les échelles nominales ou binaires (mais pas pour les échelles ordinales). Ce coefficient est compris entre -1 et +1. Son interprétation est similaire à celle du Kappa de Cohen mais cette interprétation est remise en question car les valeurs dépendent aussi du nombre des catégories.

Utiliser ces coefficients. Pour ceux que cela intéresse il existe sous R un paquetage (ou package) qui propose une interface graphique pour calculer ces coefficients ([KappaGUI](#)). Il existe aussi sur internet des plate-formes de calcul faciles à trouver.

3.5. Interprétation du coefficient

Il n'existe pas de règle stricte pour l'interprétation des valeurs du coefficient de fidélité. Toutefois, un certain consensus (sauf pour les coefficients issus de la méthode inter-juge) existe quant à l'ordre de grandeur des coefficients et à leur signification. Le tableau suivant résume les seuils d'interprétation les plus souvent admis pour ce type de coefficient.

Fidélité	Signification
.95 à 1	Test ayant très bonne précision. Les mesures contiennent un minimum d'erreur aléatoire.
.85 à .95	Test ayant une bonne précision. Les mesures contiennent peu d'erreur aléatoire.
.70 à .85	Test dont la précision est acceptable.
.50 à .70	Test très peu précis. Selon la nature de l'épreuve peu contenir de l'information utile.
0 à .50	Test inutile. Ne pas l'utiliser.

Attention

- Ces seuils ne sont pas des règles absolues, mais des repères interprétatifs largement utilisés. Ils peuvent varier selon le contexte (test clinique, recherche fondamentale, nature du test, etc).
- Pour les modèles multidimensionnels, un coefficient Omega de .50 à .60 peut être acceptable si les dimensions sont nombreuses ou faiblement corrélées.
- En règle général, en dessous de .65, le coefficient de fidélité devient trop faible pour que l'instrument soit suffisamment précis (il est trop entaché d'erreur de mesure. Le résultat n'est pas fiable !).
- Le coefficient de fidélité est le rapport entre la variance vraie (due à un ou plusieurs facteurs de différenciation des sujets) et la variance totale du test. Si la fidélité est de .80, cela signifie que 80% de la variance du test est de la variance vraie et 20% de la variance d'erreur.
- Lorsqu'on compare deux coefficients de fidélité, il est important de noter que de petites différences de fidélité peuvent correspondre à de grandes variations du rapport signal/bruit. Par exemple, une augmentation de 0.10 du coefficient de fidélité entraîne une hausse du rapport signal/bruit d'environ 1.77 lorsque la fidélité initiale est de 0.70, mais d'environ 10 lorsque la fidélité initiale est de 0.80 (*Cronbach, 1965*). Ces calculs sont faciles à vérifier en utilisant la

formule :

$$\frac{\text{signal}}{\text{bruit}} = \frac{\rho_{xx}}{1 - \rho_{xx}}$$

Remarques

- Si la fidélité est un concept simple, les méthodes d'estimation de la fidélité sont nombreuses et nous en avons présentées qu'un nombre limité (par exemples nous n'avons pas abordé la théorie de la généralisabilité ou les indices comme ceux de Guttman (lambda 1 à 6), ou encore glb (greatest lower bound), ni les indices concernant les échelles multidimensionnelles).
- Dans les manuels de tests, il ne devrait plus être acceptable de rapporter un seul coefficient mais au moins deux (ou plus) en indiquant la raison pour laquelle chacun est approprié pour l'inférence qui est faite. L'utilisation systématique du coefficient alpha de Cronbach ou du KR-20, qui pouvait s'expliquer dans les années 60-70 par la facilité de calcul de ces coefficients, devient difficilement acceptable.

Pour aller plus loin

Les coefficients de fidélité sont largement discutés dans des revues comme Psychometrika ou Applied Psychological Measurement. Les articles de Zinbarg, Revelle et Yovel (2005), Cho (2016) ou le dernier chapitre de Revelle et Condon (2018) peuvent être de bonnes introductions pour approfondir cette notion et les discussions autour de ces indicateurs.

3.6. Propriétés

La fidélité est une notion essentielle à prendre en compte lors de la construction ou l'utilisation d'un test. Un psychologue se doit de connaître la fidélité des instruments qu'il utilise (ou en avoir un ordre de grandeur). Les coefficients de fidélité dépendent cependant en partie [de la sensibilité des épreuves](#), de la [longueur des épreuves](#) (nombre d'items), de [la dispersion des scores](#). Ce coefficient de fidélité impacte aussi [l'intervalle de confiance](#) (un chapitre spécial est consacré aux intervalles de confiance).

3.6.1 Sensibilité et fidélité

L'estimation de la fidélité repose sur la variance des scores observés, et la statistique la plus fréquemment utilisée à cette fin est la corrélation. Par conséquent, quelle que soit la méthode employée, il est essentiel de comprendre que la fidélité dépend directement de la sensibilité de l'épreuve.

En effet, lorsqu'une épreuve est peu sensible, la corrélation avec une autre épreuve tend à être sous-estimée. Par exemple la variation d'un point sur une échelle de 20 points par exemple aura moins d'importance sur le positionnement du score dans la distribution qu'un déplacement d'un point sur une échelle en dix points. Une autre façon de comprendre l'effet de la sensibilité de l'épreuve est que si l'épreuve est peu sensible, la variance totale est réduite, la part de l'erreur de mesure sera plus importante.

Remarque : C'est précisément pour cette raison que, dans la méthode du split-half, on applique la correction de Spearman–Brown : chaque moitié du test, composée de $N/2N/2N/2$ items, possède une sensibilité réduite par rapport au test complet de NNN items, et la formule de Spearman–Brown permet de compenser cet effet.

3.6.2 Longueur des épreuves

Lorsque l'on modifie la longueur d'un test en ajoutant des items comparables (qui mesurent la même chose) à ceux qui existent, on augmente la fidélité de ce test (ce phénomène est la conséquence du fait que la moyenne des erreurs aléatoires tend vers 0 quand le nombre d'items augmente). La fidélité r_{kk} attendue d'un test **k fois plus long** que le test original dont on connaît sa fidélité r_{xx} peut être calculée en utilisant la formule de Spearman-Brown :

$$r_{kk} = \frac{kr_{xx}}{1 + (k-1)r_{xx}}$$

Cette formule est la formule générale. Cette formule permet aussi d'estimer la fidélité d'un test plus court (particulièrement lorsque l'on estime la fidélité à partir de la moitié d'un test, [méthode du partage](#)). Dans ce cas, on corrige en utilisant comme valeur de k , la valeur 2. C'est cette forme qui est la plus connue (après simplification).

$$r_{22} = \frac{2r_{xx}}{1 + r_{xx}}$$

Conséquence

La formule de **Spearman-Brown** permet aussi de calculer l'allongement nécessaire pour obtenir un degré de fidélité donné. Il suffit simplement d'isoler k dans la première équation et de remplacer r_{kk} par la valeur souhaitée :

$$k = \frac{r_{kk}(1 - r_{xx})}{r_{xx}(1 - r_{kk})}$$

Exemple de calcul

Un test à pour fidélité .70. Cette fidélité est peu importante et on souhaite l'augmenter en augmentant le nombre d'items (actuellement ce nombre d'items est de 25). On souhaite que cette fidélité soit au moins de .80. Quel est le nombre d'item que l'on doit ajouter à cette épreuve ?

Étape 1 : on utilise la formule : $k = \frac{r_{kk}(1 - r_{xx})}{r_{xx}(1 - r_{kk})}$

$$k = \frac{.80 * (1 - .70)}{.70 * (1 - .80)}$$

$$k = 1.71$$

on doit donc augmenter le test de façon à ce qu'il soit 1.71 fois plus long.

Étape 2 : sachant qu'il y avait 25 items, le nombre d'items minimum pour augmenter la fidélité sera de :

$$n = 25 * 1.71$$

$$n = 42.74$$

Il faudra donc que l'épreuve soit constituée de 43 items (on arrondi à l'entier supérieur)

⇒ **18 de plus que l'épreuve initiale.**

3.6.3 Fidélité et dispersion

Lorsque la fidélité d'un test est estimée sur une population x ou la variance à ce test est s_x^2 on peut

estimer la fidélité sur une autre population y à partir de la fidélité observée sur la première population et de la variance au test (s_y^2) sur cette seconde population. La formule utilisée est la suivante :

$$r_{yy} = 1 - \frac{s_x^2(1 - r_{xx})}{s_y^2}$$

On peut en déduire que :

- si la dispersion (variance) dans la seconde population est plus faible, alors la fidélité diminuera pour cette population (cela peut être assimilé à une baisse de sensibilité sur ce test : dans cette population les différences interindividuelles observées sont plus faibles).
- si la dispersion (variance) dans la seconde population est plus grande, alors la fidélité augmentera pour cette population (cela peut être assimilé, pour cette seconde population à une augmentation de sensibilité sur ce test : dans cette population les différences interindividuelles observées sont plus importantes)

Cela traduit, le fait que lorsque l'on travaille sur une population homogène, les différences réelles sont moins importantes (variance vraie) et l'importance de la variance d'erreur (qui ne change pas) devient proportionnellement plus importante et la fidélité diminue pour cette population. A l'inverse, si la population est hétérogène, les différences réelles sont plus importantes et comme la variance d'erreur ne change pas, elle devient proportionnellement moins importante et la fidélité augmente pour cette population.

Conséquence :

- Lors des études des propriétés d'un test, on a intérêt à avoir un échantillon le plus représentatif possible pour que l'estimation de la fidélité ne soit pas sur-estimée ou sous-estimée (si l'échantillon est trop hétérogène ou trop homogène).

4. Validité et validation

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) mais aussi de l'utilisateur du test (psychologue) qui doit s'assurer que l'usage qu'il fait du test correspond à celui indiqué par les concepteurs du test ([AERA/APA/NCME, 2014](#)).

Ce concept important a subi une évolution progressive au cours du XX^{ème} siècle dans la littérature scientifique.

- Historiquement, trois formes de validité ont été distinguées dans la littérature (Cronbach & Meehl, 1955) : la [validité de contenu](#), la [validité empirique](#) (ou critérielle) et la [validité de construit](#).

Toutefois, la validité étant considérée comme une question de degré plutôt que comme une propriété absolue, l'accent s'est progressivement déplacé vers la notion de **processus de validation**. En effet, la validation d'une épreuve doit être envisagée comme un processus continu, au cours duquel les preuves empiriques et théoriques s'accumulent, et qui ne s'achèvent pas avec la publication initiale du test.

- A partir des années 1990, cette conception s'élargit encore avec l'évolution du concept de validité vers celui de *processus de validation intégré* (Messick, 1995). La validité ne se réduit plus à la question « le test mesure-t-il bien ce qu'il est censé mesurer ? », mais englobe également les *conséquences sociales* de son utilisation : « le test produit-il les effets attendus lors de sa mise en œuvre ? ». Cette approche met en avant la responsabilité partagée entre les concepteurs et les utilisateurs du test. La validité dépend autant des caractéristiques intrinsèques du test que des conditions de son administration et de son interprétation.
- A la fin du XX^{ème} siècle, début du XXI^{ème} siècle émerge la notion de [validité incrémentale](#), qui met l'accent sur la contribution spécifique d'un test par rapport aux instruments d'évaluation existants.
- Enfin, on admet que la validation ne doit pas concerner que le score ou les scores observés mais doit concerner aussi l'interprétation de ces scores. En conséquence, lorsque l'on utilise ou interprète des scores d'un test de façon différente (nouvelle), on doit apporter des preuves de validité de cette utilisation ou interprétation.

Cette évolution traduit un passage d'une logique de prédiction ou de description à une logique d'explication. En effet, l'utilité, la pertinence et l'applicabilité d'une mesure ne peuvent pas être évaluées sans faire référence à une interprétation théorique des résultats. Les différentes méthodes de validation ne doivent donc pas être envisagées comme des approches concurrentes, mais comme des dimensions complémentaires d'un même objectif : celui de la compréhension et de l'explication du construit mesuré. Chaque méthode de validation constitue ainsi une preuve supplémentaire à apporter dans le cadre du processus global de validation d'un test. S'assurer de la validité d'un test ne donne pas lieu, comme pour la fidélité, à un ou plusieurs indices sur lesquels il existe un consensus. La validation d'un test est une démarche progressive qui commence dès la construction du test (validation de contenu).

Pour résumé, quel que soit l'époque, s'assurer de la validité d'un test ne donne pas lieu, comme pour la fidélité, à un ou plusieurs indices sur lesquels il existe un consensus. La validation d'un test est une démarche progressive qui commence dès la construction du test (validation de contenu).

4.1. Les preuves de la validité

Les "[standards for Educational and psychological testing](#)" (AERA/APA/NCME, 2014) rappellent que le concept de validité est un concept unitaire et que la validation est un processus qui apporte progressivement des preuves de la validité de la mesure. Une "bonne" ou "solide" validité suppose d'accumuler des preuves variées. On distingue actuellement :

- Les [preuves basées sur le contenu des tests](#) ("evidence based on test content")
- Les [preuves basées sur les processus à la base des réponses](#) ("evidence based on response processes")
- Les [preuves reposant sur l'analyse de la structure interne](#) de l'épreuve ("evidence based on internal structure")

- Les [preuves basées sur les relations avec d'autres variables](#) ("evidence based on relations to other variables")
- Les [preuves basées sur les conséquences](#) ("evidence based on consequences of testing")

Cette présentation est celle recommandée par les "standards" ou "guidelines", et celle que nous adopterons ici car elle souligne que la validité s'appuie sur l'accumulation progressive de preuves (évidences) théoriques et empiriques.

4.1.1 Le contenu du test

Il s'agit ici, d'analyser l'adéquation du contenu du test avec le construit qu'il veut mesurer. La notion de contenu est large et renvoie à la fois à la formulation des questions, au type de tâche proposée, au format des items, à la représentativité des items, mais fait aussi référence aux procédures d'administration comme de cotations. Cet aspect de la validation d'un test a donc plusieurs facettes. Elle implique le jugement d'experts et peut aussi concerner la question de l'interprétation des scores observés dans différents sous-groupes de façon à s'assurer que la sélection des items ne conduise pas à donner (par sélection d'items non pertinents ou autres) des biais spécifiques à un ou plusieurs sous-groupes.

Cette preuve de la validité d'une épreuve est rapprochée de la [validité de contenu ou validité représentative](#), (ancienne terminologie) mais elle est bien plus large. Elle est différente de ce qu'on appelait (qu'on appelle toujours) la validité apparente (face validity) qui repose sur une analyse de surface des items, analyse souvent effectuée par des non experts.

Remarque : dans les manuels de tests on se doit de préciser comment ont été construit et sélectionnés les items (questions) et justifier les choix et la méthode utilisée.

4.1.2 Les processus de réponses

L'analyse (théorique et empirique) de la façon dont les personnes répondent aux questions d'un test peut contribuer au processus de validation et peut permettre de s'assurer de l'adéquation qu'il doit exister entre la nature de la performance et le construit évalué. Cette forme de preuve de validité passe par l'analyse des réponses mais peut aussi concerner l'analyse des stratégies utilisées, l'enregistrement des modalités de réponses, l'enregistrement d'indicateurs physiologiques, l'enregistrement des mouvements oculaires, etc. En fait, tout élément permettant d'analyser les réponses et pouvant participer aux preuves de validité sont possibles. Il est évident que les moyens utilisés peuvent être très différents selon le test (test cognitif, tests de personnalité, test de connaissances, etc.).

Deux exemples peuvent illustrer ce qu'on entend par preuves basées sur les processus de réponses :

- Dans un test voulant évaluer le raisonnement mathématique, on doit s'assurer que les réponses données mettent bien en jeu ce raisonnement et n'implique pas un algorithme classique permettant de donner la bonne réponse (sans véritable raisonnement mathématique). Une analyse des stratégies utilisées (approche cognitive) doit assurer la nature de ce qu'on évalue (le raisonnement et non des connaissances ou des algorithmes de résolution automatiques préalablement acquis).
- Dans un questionnaire, on doit s'assurer qu'un effet de conformité sociale n'affecte pas les réponses ou la façon de répondre.

L'analyse des processus de réponses peut aussi, lorsque cela est pertinent, porter sur les procédures de cotation de façon à s'assurer qu'un utilisateur de l'épreuve (psychologue par exemple) ne soit pas influencé, lors de la cotation, par des aspects non pertinent des réponses ou d'autres aspects de l'épreuve.

4.1.3 La structure interne

Il s'agit ici de s'assurer que les relations entre les items ou les sous-composants du test sont conformes au(x) construit(s) qu'il est censé mesurer. Le type d'analyse à mettre en œuvre dépend de la nature des items et des scores observés. D'une certaine façon et dans certains cas, cet aspect peut être rapproché de la notion plus classique de [validité de construit](#). Les méthodes pour évaluer la structure interne s'appuient le plus souvent sur les [corrélations](#) inter items, inter-échelles d'un test, l'analyse factorielle et confirmatoire, etc. Certaines de ces méthodes sont décrites dans le chapitre [d'introduction à l'analyse factorielle](#).

4.1.4 Relations avec d'autres variables

Cet aspect de la démonstration de la validité d'un test concerne l'étude des relations entre le ou les scores observés à l'épreuve et les "scores" observés sur d'autres variables (critères).

Ces variables peuvent être d'autres tests dont la mesure peut converger (critère mesurant le même construit) ou diverger (critère mesurant un autre construit) avec l'épreuve en cours de validation. On peut aussi utiliser des critères autres que des tests. Par exemple, si ce que l'on veut mesurer prédit des différences entre groupes (sociaux, pathologiques, etc.) on doit observer des différences en fonction de ces groupes sur le test. Ce type de validité est à rapprocher de la [validité empirique](#) et de la [validité de construit](#).

4.1.5 Utilité et conséquences

Dans l'appréciation de la validité ou les preuves de validité d'un test, on devrait prendre en compte l'utilité (à rapprocher parfois de la validité incrémentale) et des conséquences de l'usage des tests. **L'utilité d'un test** fait référence aux bénéfices attendus lors du processus de décision ou lors d'une évaluation. L'utilité peut donc être évaluée par le coût/bénéfice de son utilisation. **Les conséquences de l'usage des tests** renvoient à l'utilisation (positive ou négative) comme aux conséquences sociales de l'usage du test. Cela renvoie donc à des aspects éthiques plus qu'à la validité en tant que telle.

Remarque : cet aspect de la validité est complexe et on renvoie le lecteur qui souhaite approfondir au livre [d'Urbina \(2014\)](#) ou au "standards" ([AERA, APA, & NCME, 2014](#)).

4.2. Terminologie plus ancienne

Dans certains manuels de psychométrie, le concept de validité est encore présenté en reprenant la distinction proposée initialement par Cronbach (1984) puis Messick (1989). **Cette terminologie ne doit pas conduire à penser qu'il existe plusieurs formes de validité**. Ce sont des aspects ou des méthodes qui participent au processus de validation ou aux preuves de la validité d'un test.

4.2.1 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.).

Remarques

- La validation de contenu joue un rôle important dans le développement des différents tests utilisés en psychologie et en éducation. Par exemple, pour valider un test de connaissances en mathématique correspondant à un scolaire, on va comparer ces items avec tous les points du programme de mathématique de ce niveau scolaire et s'assurer que les différents items couvrent bien tout le programme.
- Il est bien entendu nécessaire, dans les étapes préliminaires de construction des tests de s'assurer que les items sont pertinents et représentatifs des concepts ou des définitions sous-tendant la mesure. On ne pourra pas cependant se contenter de ce type de validation même si cette forme de validité est toujours prise en compte lors de la construction des tests.
- On considère le test comme un échantillon représentatif d'une population d'items (de questions) bien définis. Cela requiert de définir l'univers de ces questions et soulève des problèmes d'échantillonnage des questions. Par exemple, dans un test d'opérations arithmétiques, la validité ne serait pas suffisante si l'on n'y mettait que des problèmes d'addition en négligeant les autres opérations (à moins qu'on ne décide qu'il s'agisse d'un test d'addition !).
- Bernaud (2007) propose plusieurs règles essentielles de la validation du contenu (que l'on résume pour certaines ici) :
 - Définir correctement le construit mesuré.
 - Lors de la création d'items et plus généralement de l'instrument s'appuyer sur des experts du domaine.
 - Consulter aussi des experts pour valider le contenu et quantifier ces jugements à l'aide d'échelles (exemple échelle de Likert)
 - Évaluer la représentation des items relativement aux différentes facettes du construit que l'on veut mesurer.

En résumé : 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.

4.2.2 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.

Le terme de validité empirique, habituellement utilisé en France, recouvre ce que les anglo-saxons appelle validité de critère, validité critériée ou encore validité pragmatique. Pour établir la validité empirique d'un test x par rapport à un critère y , on se sert d'un échantillon représentatif de la population à laquelle on destine le test et on détermine le degré de covariation entre le test et le critère en utilisant par exemple le coefficient de corrélation de Bravais-Pearson que l'on appelle alors parfois le coefficient de validité.

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 utilisés).
- **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).

Pour chacune de ces deux types de validations, on distingue aussi la validité convergente et la validité divergente.

- **La validité convergente** vise à estimer la validité d'un test par sa ressemblance avec d'autres mesures considérées comme similaires.
- **La validation divergente**, par contraste, confirme la validité d'un test par la divergence (corrélation nulle par exemple) des résultats qu'on obtient entre le test et d'autres tests ou critères dont on fait l'hypothèse qu'ils mesurent autre chose. Cette méthode complémentaire à la validité convergente permet de s'assurer que la variance vraie associée au test (les différences réelles observées) est pour l'essentiel associée au construit que l'on souhaite mesurer et non à un autre construit. Par exemple, lors de la construction d'une épreuve voulant évaluer l'aptitude verbale, la présentation comme la nature des items peut laisser penser que le test est en lien avec d'autres aptitudes (visuo-spatiale, raisonnement, etc.). La validité divergente permettra de s'assurer que le test est peu ou pas corrélé avec ces autres construits.

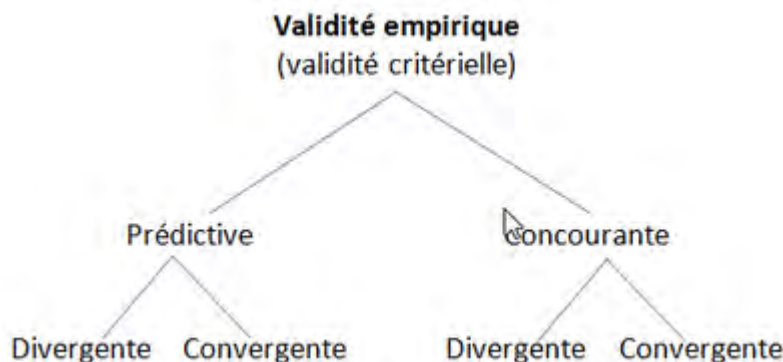


Figure D.2 : Les différentes formes de validité empiriques

Remarque

La validité du test dépend de la [fidélité](#) aussi bien du test que du critère. Il est évident que si le classement des sujets dans l'une et/ou l'autre des variables est arbitraire, le coefficient de validité sera nul ou très faible. Un des problèmes réside dans le fait qu'il est souvent difficile d'obtenir des critères fidèles. L'utilisation de la [correction pour atténuation](#) lors du calcul de la corrélation permet de tenir compte (pour interpréter la relation entre test et critère) de la fidélité du test et du critère.

4.2.3 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.

La validation empirique ne permettait pas forcément l'interprétation du mécanisme psychologique sous-jacent (par exemple, on peut utiliser un test parce qu'il prédit bien la réussite en lecture, mais sans forcément savoir pourquoi). Par contre, la validation de construit vise également à une interprétation théorique. On définit trois étapes dans cette démarche (similaire à ce qu'on appelait anciennement de validation hypothético-déductive : (i) construction d'hypothèses théoriques ; (ii) déduction d'hypothèses testables ; (iii) planification d'une étude expérimentale pour tester les hypothèses. On va s'assurer ainsi que les relations entre les différentes facettes du ou des construits mesurés sont conformes au modèle théorique (on utilisera des méthodes statistiques diverses comme l'analyse factorielle confirmatoire ou plus généralement des méthodes d'équations structurales).

Remarques

- La validité de construit est d'autant plus importante en psychologie qu'elle permet d'aller au delà des limites inhérentes à la validité de contenu et à la validité empirique. En effet, la psychologie s'intéresse à des réalités pour lesquelles il est particulièrement difficile de définir des critères satisfaisants ([validité empirique](#)) et de s'assurer que la totalité du domaine qui intéresse a été pris en compte dans la mesure ([validité représentative ou de contenu](#)).
- Cronbach et Meehl (1955) qui ont formalisé cette notion de validation de construit insistent sur la nature complexe de ce processus de validation. En effet, cette validation nécessite de nombreuses recherches par différents chercheurs travaillant sur différents aspects théoriques du construit mesuré.
- La validation de construit n'est pas une méthode unique et explicite pour établir la validité d'un test, mais bien un ensemble de méthodes qui visent toutes le même but : établir jusqu'à quel point le test fournit une mesure adéquate du construit théorique qu'on prétend qu'il mesure.
- **ATTENTION.** La validité de construit d'un test n'existe que si l'on a une définition explicite du construit que l'on veut mesurer. Le postulat à la base de la notion de validité de construit se fonde sur l'hypothèse que ces construits "existent" (dans la TCT, les tests sont des mesures réflexives) et que l'on peut mettre en relation les variations interindividuelles existant sur ce construit et celles que l'on observe sur les tests. Ce postulat peut être remis en question et la la validité du construit aussi !

4.2.4 Validité incrémentale

La validité incrémentale (ou incrémentielle) est rarement décrite. Cette notion introduite de façon plus spécifique par [Sechrest](#) en 1963, concerne essentiellement les batteries de tests. Dans ce cas, on considère qu'un test doit augmenter de manière significative la puissance de prédiction de l'ensemble des tests présents. Un test est donc valide s'il permet de mieux prédire un critère que ce que ferait la batterie de tests sans ce test lui-même.

De façon plus générale, la validité incrémentale consiste à se poser la question de savoir si un test apporte plus pour prédire un critère que les autres informations déjà disponibles (tests ou autres techniques). La validité incrémentielle est estimée le plus souvent par des techniques de régression multiple hiérarchique. Pour une discussion plus détaillée ([Hunsley, Meyer, 2003](#)).

Remarque.

Ce critère est intéressant mais s'éloigne cependant de la notion de validité telle qu'elle était classiquement définie. Par ailleurs, un psychologue doit avoir à sa disposition plusieurs instruments différents permettant d'évaluer un même construit. En effet, il est parfois utile d'évaluer à distance, la même personne (adulte ou enfant) suite à une prise en charge ou une évolution de la situation personnelle. Disposer de plusieurs instruments valides et différents permet de s'affranchir en partie des effets de test-retest.

5. Validité vs fidélité

"There are no perfectly reliable or perfectly valid instruments; reliability and validity are matters of degree."

NUNNALLY (1978)

La fidélité concerne la précision avec laquelle un test mesure certaines caractéristiques, elle est donc en relation avec l'erreur de mesure et elle est formellement définie comme le rapport de la variance vraie à la variance totale du test (théorie classique des tests). On définit par contre la validité comme la qualité de ce qui est mesuré ; c'est à dire la ressemblance existant entre ce que l'on veut mesurer et ce que mesure le test. De façon formelle, la validité est donc la portion de variance vraie qui est pertinente aux buts de l'utilisation du test (toujours dans le cadre de la [TCT](#)).

Pour résumer cette distinction entre fidélité et validité reprenons la décomposition des scores aux tests :

- $X = T + E$ (avec T, score observé, T score vrai, et E l'erreur)
- T peut être décomposé en deux facteurs : T_p qui est l'effet de la dimension pertinente (celle que l'on veut mesurer) + T_{np} qui est l'effet des dimensions non pertinentes mais qui correspondent à de la variance vraie sur la mesure (non aléatoire comme E). Un test peut en effet mesurer plusieurs choses.

Dans ce cadre, T (détermine la variance vraie) est donc décomposée en deux parties, T_p et T_{np} . Si l'on reprend nos définitions de la fidélité et de la validité, la fidélité est le rapport de la variance de T (donc T_p+T_{np}) sur la variance de X (variance totale) tandis que la validité le rapport de la variance de T_p sur la variance de T (donc T_p+T_{np}).

On peut déduire de ces formules algébriques que si un test est suffisamment fidèle (*i.e.* a une part de variance vraie significative), il n'est pas obligatoirement valide et que pour qu'un test soit valide il est nécessaire que celui-ci soit plaisamment fidèle, c'est à dire que la variance totale ne soit pas que de l'erreur de mesure.

- Il existe donc une relation d'implication entre ces deux notions [**validité** \Rightarrow **fidélité**] : *la fidélité est une condition nécessaire mais non suffisante pour la validité d'un test.*
 - (1) Un test non fidèle est nécessairement non valide.
 - (2) Un test valide se doit d'être, a minima, fidèle.
 - (3) Un test fidèle n'est pas nécessairement valide.
- L'absence de fidélité traduit une erreur aléatoire autour d'un point moyen (qui peut être la cible ou non) l'absence de validité traduit une erreur constante qui éloigne le résultat de la cible visée. Fidélité et validité sont deux notions distinctes liées par une relation d'implication (cf. ci-après).

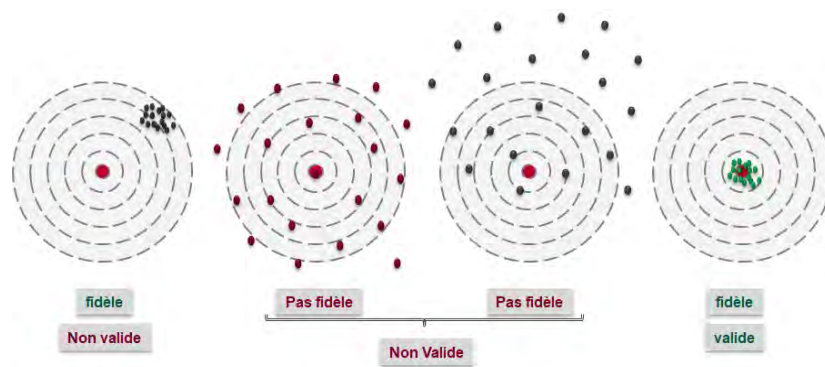


Figure D.3 : Représentation* imagée des notions de fidélité et validité (adapté de Chapanis, 1951)

() représentation souvent utilisée mais la source est rarement citée
(pour cause : cette recherche concerne les systèmes d'information navals)*

6. Contre validation

La **contre-validation** constitue une étape importante (parfois négligée) du processus de construction d'un test. Elle permet de s'assurer que les résultats obtenus ne dépendent pas des caractéristiques spécifiques d'un échantillon. Concrètement les propriétés psychométriques observées (fidélité, structure factorielle, indices de difficulté ou de discrimination des items) peuvent refléter les particularités de l'échantillon initial. Pour s'assurer de la possibilité de généraliser les résultats, la version finalisée du test devrait faire l'objet de vérifications. Dans l'idéal, on devrait administrer le test et reconduire les analyses avec un second échantillon indépendant, souvent appelé **échantillon de contre-validation**. Cette étape permet d'examiner la stabilité des propriétés psychométriques observées et de confirmer la qualité de la mesure..

En pratique, la contre-validation est rarement menée avant la publication d'un test, principalement pour des raisons de coût, de temps et de disponibilité des échantillons. Ce sont souvent des études ultérieures, parfois réalisées par d'autres chercheurs que les concepteurs ou adaptateurs du test, qui assurent la validité de l'outil.

E - L'étalonnage d'un test

1. Présentation

L'objectif de l'étalonnage c'est d'offrir à l'utilisateur du test un repère en permettant de comparer le (les) score(s) brut(s) de la personne évaluée aux scores observés sur une population dont les individus ressemblent à la personne qui vient de passer l'épreuve. L'étalonnage, peut-être assimilé à « *un barème utilisé pour le classement d'une valeur individuelle par rapport à l'ensemble des valeurs caractéristiques d'une population* » (Piéron, 1951).

1.1. Définition

La **définition** habituellement retenu est la suivante « *L'étalonnage consiste à transformer les scores bruts obtenus à un test en scores standardisés, de manière à permettre la comparaison entre individus et avec un groupe de référence.* » (Anastasi & Urbina, 1997, p. 112)

L'action d'étalonner correspond donc à la réalisation d'un ensemble d'opérations établissant la relation entre le score brut et une valeur ayant une signification. Cette valeur permettra d'identifier facilement la position de ce score par rapport aux scores observés dans l'échantillon de standardisation. Suite à cette opération, toutes les notes brutes possibles (et non pas seulement les notes brutes observées dans l'échantillon) appartiendront à l'une ou l'autre des catégories de référence de l'étalonnage.

Pourquoi construire un étalonnage ?

- L'étalonnage donne un sens à la mesure car les catégories ou classes de l'étalonnage permettent dans certains cas de condenser l'information mais surtout de rendre les données plus clairement interprétables.
- Les classes ou catégories de l'étalonnage permettent de rendre comparable des mesures de caractères divers et hétérogènes (par exemple, des temps en secondes et des points par item réussi). Elles permettent de comparer les catégories auxquelles appartient une personne dans des épreuves différentes.
- L'étalonnage permet, dans certains cas et en acceptant un postulat sur la forme de la distribution, de transformer l'échelle ordinale que constitue la note brute en échelle qui sera ensuite traitée comme une échelle d'intervalle dans les recherches.

Observations :

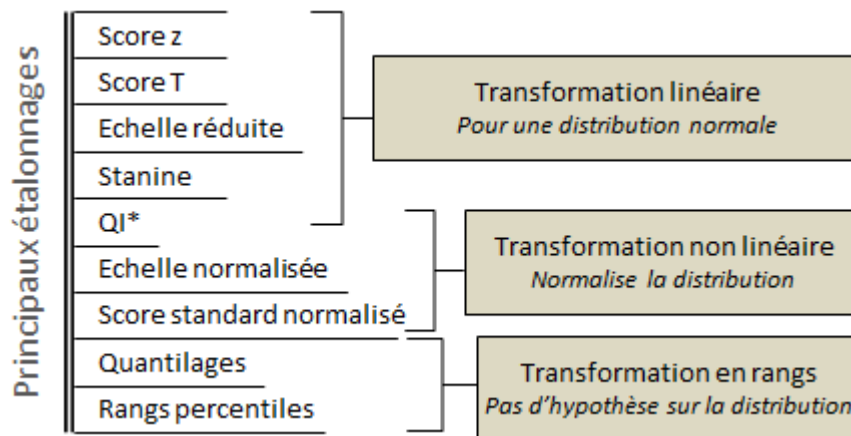
- Le terme étalonner dans la langue française à une définition partiellement différente de celle des psychologues. Étalonner peut avoir deux sens courant : (1) vérifier la conformité d'une mesure ou d'une valeur avec l'étalon conventionnel (2) Graduer, régler un instrument conformément à l'étalon. Le terme d'étalonnage en psychologie fait plutôt référence à cette deuxième définition (réglage d'un instrument) et la mesure obtenue (quel que soit la métrique utilisée) traduit une position dans la population parente.
- L'étalonnage devrait toujours être effectué auprès d'un échantillon représentatif de la population de référence. L'établissement de la représentativité est une tâche difficile à réaliser lors de la construction d'un test et doit donc être effectué avec un soin approprié. Il est important d'identifier les variables qui covarient systématiquement avec ce que l'on souhaite évaluer. Par exemple, pour

des tests de raisonnement, il peut s'agir du niveau d'étude des parents, du lieu de vie, du type d'école, du contexte socio-économique, etc.

1.2. Types d'étalonnage

Les étalonnages ne reposent pas tous sur la même métrique. En psychologie, on distingue généralement trois grandes catégories d'étalonnage couramment utilisées. Elles sont classées ici selon la nature de la transformation appliquée aux scores bruts (cf. schéma ci-dessous) :

- transformations linéaires : s'appliquent sur des scores dont la distribution est normale (après transformation initiale on non des scores bruts) ;
- transformations non linéaires : ils ont objectif principal de normaliser la distribution des scores (les scores étalonnés se distribueront normalement). On peut éventuellement ensuite faire une transformation linéaire ;
- transformation en rangs : ils transforment les scores bruts en rang (sans hypothèse sur la distribution).



Le QI constitue un [type particulier d'étalonnage](#) en deux temps. Dans un premier temps, les scores bruts obtenus aux épreuves sont standardisés (échelles normalisées en 19 classes). Dans un second temps, la somme de ces scores standards est transformée en QI au moyen d'une transformation linéaire.

1.3. Tables d'étalonnage

Nous avons vu que l'étalonnage est la procédure qui permet de faire correspondre des scores brutes et des scores "étalonnés", scores alors interprétables. L'étalonnage désigne aussi le résultat de cette technique présentée le plus souvent dans les manuels de tests sous la forme de tableaux (**tables d'étalonnage**). Les psychologues utilisent ensuite ces tables d'étalonnage (qui se trouvent dans les manuels des tests). Elles sont plus ou moins détaillées en fonction des caractéristiques de l'[échantillon de standardisation](#) (tables d'étalonnage par âge, sexe, profession, etc.).

Une table d'étalonnage, met donc en correspondance des notes brutes et des scores sur une échelle de valeur ayant une signification. Selon la nature de l'étalonnage d'autres informations peuvent être fournies en même temps aux psychologues. Par exemple, dans la table d'étalonnage ci-dessous il est à la simple correspondance "score étalonné - notes brutes" le pourcentage de personnes ayant une performance inférieure ou égale à celle observée. Ceci facilite l'interprétation des notes étalonnées.

Exemple : dans la table présentée ci-dessous une note brute de 48 correspond à une note étalonnée de 2. La table nous indique que 15,8% des personnes de l'échantillon d'étalonnage à un score inférieur ou

égal à cette note de 2. Ainsi, il est possible de déduire que 84,2% réussissent mieux cette épreuve. Ce score 48 correspond à une personne ayant rencontré des difficultés dans l'épreuve qui lui a été administrée lorsque l'on compare sa performance à celle de la population de référence ou plus exactement à l'échantillon de standardisation.

Exemple d'une table d'étalonnage (fictive)

Notes étalonnées	Fréquence (%)	Fréquence cumulée (%)	Notes brutes
0	3.6 %	3.6%	20-42
1	4.6%	8.10%	43-46
2	7.7%	15.8%	47-52
3	11.6%	27.4%	53-67
4	14.6%	42.0%	58-61
5	16,0%	58.0%	62-65
6	14,6%	72.6%	66-69
7	11.6%	84.2%	70-72
8	7.7%	91.9%	73-74
9	4.6%	96.4%	75
10	3.6%	100%	76-80

2. Construction d'un étalonnage

Il existe différents types d'étalonnage, dont les règles de construction varient selon la méthode retenue. De manière générale, la construction d'un étalonnage suppose le choix d'un type spécifique ([quantilage](#), [échelle normalisée](#), [échelle réduite](#), etc.) et, le cas échéant, la détermination du nombre de catégories de référence à utiliser.

Le choix du nombre de catégories d'étalonnage dépend principalement des objectifs de la mesure et du niveau de discrimination souhaité entre les personnes. Il est également influencé par la nature des données, la taille de l'échantillon de standardisation, le nombre de notes brutes observées et la sensibilité de l'épreuve. Certaines règles simples sont souvent utilisées :

- le nombre d'observations indépendantes (personnes) doit être au moins 10 fois plus important que le nombre de catégories de l'étalonnage. Pour un étalonnage en 10 catégories, il faut au moins 100 sujets.
- le nombre des notes brutes observées doit être 3 à 4 fois supérieur à celui des catégories de l'étalonnage (pour un étalonnage en 10 classes, il faut recueillir 30 à 40 notes brutes différentes).
- plus le nombre des catégories est élevé, plus le test doit être fidèle : il serait peu utile de chercher à discriminer finement des personnes dont les performances présentent une forte instabilité (à ce test).

2.1. Quantiles

Ce type d'étalonnage, permet d'exprimer le score d'une personne en terme de rang. Le principe des quantiles est simple. Il s'agit de découper la distribution des notes brutes en k intervalles de telle façon qu'il y ait autant de personnes de l'échantillon dans chacune des classes de l'étalonnage.

Méthode de construction : exemple pour un quartilage = quantilage en 4 classes.

Dans cet exemple la colonne "cat" correspond aux scores bruts observés dans ce test. La colonne "eff" correspond au nombre de personnes (effectif) ayant le score brut observé. La colonne "Cum" est simplement l'effectif cumulé et la colonne "%" est le pourcentage cumulé (facultatif). Le principe est de calculer les valeurs Q_i ($Q_i = i \cdot n/k$ avec n le nombre d'individu constituant l'échantillon et k le nombre de catégories). Il faut rechercher ensuite le score brut (colonne cat) dont le l'effectif cumulé est le plus proche de cette valeur (cf tableau ci-dessous) afin de construire la table d'étalonnage (ici en 4 catégories) avec pour chaque catégorie l'étendue des scores bruts qui appartiennent à cette catégorie.

Cat.	Eff.	Cum.	%
2	1	1	0,90%
3	2	3	2,70%
4	0	3	2,70%
5	3	6	5,41%
6	2	8	7,21%
7	1	9	8,11%
8	3	12	10,81%
9	8	20	18,02%
10	4	24	21,62%
11	6	30	27,03%
12	13	43	38,74%
13	7	50	45,05%
14	8	58	52,20%
15	12	70	63,06%
16	14	84	75,67%
17	10	94	84,68%
18	17	111	100,00%

1. Pour chaque note possible on détermine le nombre des personnes ayant cette note (effectif)
2. On calcule ensuite les effectifs cumulés et les pourcentages cumulés (facultatif).
3. On calcule les valeurs Q qui permettent de fixer les quantiles ($Q_i = i \cdot n/k$):
 - $Q_1 : 1 \cdot 111/4 = 27.75$ Si entier ..., +1
 - $Q_2 : 2 \cdot 111/4 = 55.50$
 - $Q_3 : 3 \cdot 111/4 = 83.25$
4. Surligner la classe la plus proche de chaque Q_i .
5. Construire l'étalonnage

	% Th	notes	%Obs.
1	25%	0-11	27,3%
2	25%	12-14	24,99%
3	25%	15-16	23,47%
4	25%	17-18	24,33%

Remarques

- Ce type d'étalonnage présente l'avantage d'être aisé à établir et de ne requérir aucun postulat sur la forme de la distribution, hormis l'ordre des résultats.
- Il existe différentes formes de quantiles telles que le quartilage (25% des individus dans chaque classe), le décilage (10% des personnes dans chaque classe).
- Il existe un inconvénient majeur au quartilage : il ne présente pas la même finesse discriminative dans toutes les catégories si la distribution des scores bruts est normale ou quasi normale. Dans ce cas les personnes sont classés finement dans la partie moyenne (les notes brutes correspondant aux centiles 40, 50, 60 sont peu nombreuses) alors que le classement des sujets est grossier aux extrémités (le nombre de notes brutes par quantile peut devenir important). En fait, ce type d'étalonnage est plus adapté aux distributions non normales, notamment aux distributions rectangulaires, platicurtiques ou présentant une asymétrie marquée.

2.2. Rangs percentiles

Le rang centile ou rang percentile indique simplement la proportion de personnes ou de scores qui sont égaux ou inférieurs à un score brut. Le mode de calcul* est simple puisque le rang percentile correspond au pourcentage de personnes qui ont un score brut inférieur auquel s'ajoute (valeur de correction) la moitié du pourcentage des personnes qui ont exactement cette valeur. Le mode de calcul formel est le suivant (en général, la valeur est arrondie au nombre entier suivant sauf si la valeur est supérieur à 99) :

$$r_{100(x)} = 100 * (D + 0.5 * E) / n$$

avec : x la valeur pour laquelle on souhaite calculer le rang percentile

D le nombre des scores inférieurs au score x observé

E le nombre des scores identiques (*ex-æquo*) au score x observé

et n le nombre total des scores dans la distribution

Pour un score x , l'utilisation de pourcentage, conduit à utiliser une formule différente :

$$r_{100(x)} = p_c - 0.5p_x$$

avec p_c le pourcentage cumulé correspondant au score x

p_x le pourcentage de score x

Avantage et inconvénients du rang percentile. Le rang percentile présente l'avantage d'être une statistique facile à calculer et facilement comprise. Il peut être cependant trompeur lorsque le rang percentile est calculé sur un faible échantillon ou lorsque la distribution est très asymétrique (avec un effet plafond ou plancher). En effet dans ce cas une petite différence de score brut peut être artificiellement associée à une grande différence en rang percentile. Il faut donc interpréter les rangs percentiles avec prudence (et tenir compte de la forme de la distribution comme de la taille de l'échantillon).

(*) Il existe plusieurs méthodes de calcul mais celle proposée est la plus fréquente, prenant en compte les valeurs égales (formule qui correspond aussi aux tableurs les plus courants).

Exemples de calcul du rang percentile

Calculer le rang percentile d'un individu dont le score est le 16^{ème} meilleur score d'un groupe de 80 personnes sans *ex-æquo* :

- le nombre de score inférieur est de 80-16 soit $D = 64$
- $E = 1$
- $r_{100(x)} = 100 * (64 + 0.5) / 80 = 80.625$
=> **le rang percentile est de 81**

Calculer le rang percentile d'un individu dont le score est le 16^{ème} meilleur score d'un groupe de 80 personnes mais il y a *ex-æquo* (4 personnes de rang 16) pour ce score :

- son score est dépassé par 80-16-4 personnes -> $D = 60$
- $E = 4$
- $r_{100(x)} = 100 * (60 + 0.5 * 4) / 80 = 77,5$
=> **le rang percentile est de 78**

2.3. Echelle réduite

Les échelles réduites sont utilisées uniquement lorsque la distribution des notes brutes est normale ou quasi-normale. Les catégories de l'étalonnage vont être fondées sur des écarts à la moyenne.

Principe de construction.

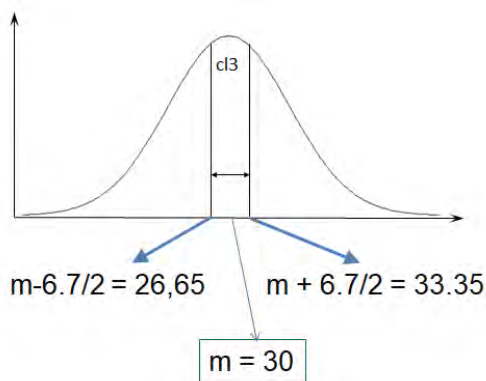
Ces échelles sont faciles à mettre en œuvre. Le nombre de classes est habituellement impair (5, 7, 9, 11) et la classe centrale est centrée sur la moyenne. Les bornes d'une classe sont fixées ensuite en prenant comme **étendue d'une classe**. Cette étendue est déterminée de façon à ce que l'ensemble des classes couvre la distribution des scores et par convention, les valeurs utilisées sont le plus souvent ::

convention	• Échelle en 5 classes : 1σ
	• Échelle en 7 classes : $0,66 \sigma$
	• Échelle en 9 classes : $0,50 \sigma$
	• Échelle en 11 classes : $0,40 \sigma$
	• Échelle en 19 classes : $0,33 \sigma$

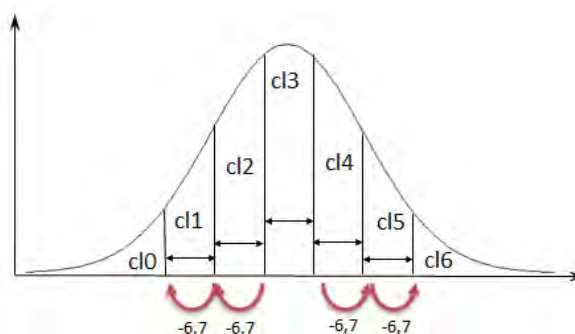
Illustration

Soit une épreuve avec des scores possibles allant de 0 à 60, de moyenne 30 et d'écart-type de 10. L'objectif est de construire un étalonnage en 7 classes. Quatre étapes seront nécessaires :

1. Déterminer l'étendue des classes. Cette étendue, pour une échelle en 7 classes correspond à $0,67 * \sigma$. Cette étendue est donc de $0,67 * 10 = 6,7$
2. Calculer les bornes de la classe centrale (classe 3) qui est centrée sur la moyenne par convention en psychométrie. Les bornes de cette classe sont donc à $6,7/2$ soit 3,35 points de la moyenne de chaque côté (cf. schéma ci-dessous).



3. Il y a 7 classes donc 3 de chaque côté de la moyenne. Il suffit de progresser d'un pas de 6,7 à partir de chaque borne de la classe 3.



4. Cette règle a permis de calculer les bornes théoriques. Les bornes réelles (étalonnage), si les scores bruts sont des entiers vont remplacer ces bornes théoriques dans la table d'étalonnage. La règle est simple, pour la borne inférieure prendre l'entier supérieur de la borne inférieure théorique (exemple $\lceil 13.25 \rceil = 14$) et pour la borne supérieure prendre l'entier inférieur de la borne théorique supérieure (par exemple : $\lfloor 46.75 \rfloor = 46$).

		Bornes Théoriques	Bornes (Etalonnage)
Classes	0	$]-\infty ; 13,25]$	0 - 13
	1	$]13,25 ; 19,95]$	14 - 19
	2	$]19,95 ; 26,65]$	19 - 26
	3	$]26,65 ; 33,35]$	27 - 33
	4	$]33,35 ; 40,05]$	34 - 40
	5	$]40,05 ; 46,75]$	41 - 46
	6	$]46,75 ; +\infty[$	47 - 60

Remarques

- Ce type d'étalonnage présente le grand avantage d'être aisé à établir mais nécessite que la distribution des scores bruts soit normale ou quasi-normale.
- Avant d'utiliser ce type d'échelle il faut réaliser un test de normalité de la distribution (comme le test de [Kolmogorov-Smirnov](#) ou [Lilliefors](#)).
- Si la distribution des données brutes n'est pas normale, les effectifs des diverses catégories varient irrégulièrement et surtout arbitrairement.
- L'étendu d'une classe est fixée par convention et peu varier d'une étude à l'autre mais les valeurs les plus fréquentes sont celles indiquées.

2.4. Echelle normalisée

Ces échelles sont une combinaison des deux types d'étalonnage précédents. Le principe général est de transformer les données en s'appuyant sur les caractéristiques de la courbe normale théorique et non pas sur la moyenne et l'écart-type des données brutes (comme pour les [échelles réduites](#)).

Principe de construction.

- Transformer en échelle réduite une distribution normale théorique de moyenne 0 et d'écart-type 1. Comme pour une échelle réduite les classes seront donc construites autour de la moyenne.
- Déterminer le pourcentage de personnes de l'échantillon qu'il doit y avoir dans chaque classe (ce qui est possible puisque l'on connaît la distribution de la loi normale théorique : $N(\mu=0, \sigma=1)$).
- Appliquer la [règle du quantilage](#) pour déterminer les bornes de chaque classe.

Illustration

Dans une épreuve les scores bruts possibles vont de 20 à 80. Cette épreuve est administrée à 112 personnes représentatives des étudiants de psychologie de deuxième année. Construire, à partir de ces données, une échelle normalisée en 11 classes.

Étape 1. Calcul des bornes pour distribution normale théorique de moyenne 0 et d'écart-type 1. L'étendu des classes par convention est de $0.40 \cdot \sigma$ (11 classes). Les bornes sont (cf. [méthode échelle réduite](#)) :

-2.20 ; -1.80 ; -1.40 ; -1.00 ; -0.60 ; - 0.20 ; +0.20 ; +0.60 ; +1.00 ; +1.40 ; +1.80 ; + 2.20

Étape 2. Calcule les pourcentage et le pourcentage cumulé de personnes se trouvant dans chacune des classes lorsque la distribution est normale. On utilise pour cela une table de la loi normale.

Classe	0	1	2	3	4	5	6	7	8	9	10
%	3,6%	4,5%	7,8%	11,6%	14,6%	15,9%	14,6%	11,6%	7,8%	4,5%	3,6%
% cumulé	3,6%	8,1%	15,9%	27,4%	42,1%	57,9%	72,6%	84,1%	91,9%	96,4%	100,0%

Étape 3. Procéder comme pour un [quantilage](#) mais avec les valeurs du tableau précédent. Les bornes ("pseudo-quantiles") sont calculées de la façon suivante :

$$Q_i = n \cdot \%F_{cum_i}$$

avec Q_i la borne supérieure de la classe i

n le nombre de sujet

$\%F_{cum_i}$ le pourcentage cumulé de la classe i

n = 112

Classe	0	1	2	3	4	5	6	7	8	9	10
%	3,6%	4,5%	7,7%	11,6%	14,6%	16,0%	14,6%	11,6%	7,7%	4,5%	3,6%
% cumulé	3,6%	8,1%	15,9%	27,4%	42,1%	57,9%	72,6%	84,1%	91,9%	96,4%	100,0%
Bornes	4,02	9,04	17,77	30,72	47,12	64,88	81,28	94,23	102,96	107,98	112,00

Étape 4. Surligne dans le tableau des effectifs cumulés la classe avec l'effectif le plus proche de la borne calculée.

Cat.	Eff.	Cum.							
34	1	1	0	↓	58	4	30,72	33	4
39	1	2			59	6	39		
42	2	4			60	4	43		
43	1	4,02			61	4	47		
44	2	7	1	↓	62	5	47,12	52	5
45	1	8			63	4	56		
46	1	9	2	↓	64	4	60	6	
47	3	9,04			65	5	64,88		65
48	1	13			66	2	67		
49	1	14	3	↓	67	5	72	7	
51	1	15			68	4	76		
52	2	17	3	↓	69	8	81,28	84	8
53	2	17,77			71	4	88		
54	2	21			72	7	94,23	95	
55	3	24			73	2	97		
56	4	28	9	↓	74	5	102,96	102	9
57	1	29			75	6	107,98	108	
					76	2	110		
					78	1	111		
					80	1	112	10	

Étape 5. Remplir la table d'étalonnage qui permet ensuite de convertir toutes les notes brutes en notes étalonnées. Toute les notes brutes apparaissent même celles qui ne sont pas observées dans l'échantillon (dans notre exemple le score minimum possible était 20).

Notes étalonnées	% théoriques	%cumulés	Notes Brutes
0	3,59%	3,59%	20-42
1	4,48%	8,08%	43-46
2	7,79%	15,87%	47-52
3	11,6%	27,43%	53-57
4	14,65%	42,07%	58-61
5	15,85%	57,93%	62-65
6	14,65%	72,57%	66-69
7	11,56%	84,13%	70-72
8	7,79%	91,92%	73-74
9	4,48%	96,41%	75
10	3,59%	100%	76-80

Remarques

- **Les deux premières étapes de la procédure ne sont en fait jamais effectuées** . En effet, il existe déjà des étalonnages en 11 classes et les pourcentages de chaque classe sont connues (les mêmes pour tous les étalonnages de ce type) et n'ont pas besoin d'être recalculées (sauf si l'on souhaite prendre une étendue différente de celle habituellement utilisée).
- Ce type d'étalonnage présente le grand avantage d'être aisé à établir et ne nécessite pas que la distribution des scores bruts soit normale.
- Si la distribution des scores bruts est normale une échelle normalisée donne les mêmes classes d'étalonnage qu'une échelle réduite.
- Ce type d'étalonnage est très fréquent. Par exemple cette méthode est utilisée pour construire les notes standards en 19 classes dans les échelles de Weschler.
- Les notes étalonnées se distribuent normalement (c'est donc une normalisation de la distribution par transformation non linéaire des notes brutes).

2.5. Scores z

La note z (le score z ou encore score standard) correspond à l'expression d'un écart à la moyenne exprimé en fraction d'écart-type. Pour une distribution de notes de moyenne m et d'écart-type σ , la note z correspondant au score x se calcule facilement et correspond à :

$$z_x = \frac{x - m}{\sigma}$$

Propriétés

- La moyenne des notes z est égal à 0 et l'écart-type est égal à 1
- Ce score permet de répondre à la question : de combien de fraction d'écart-type s'éloigne-t-on de la moyenne ?
- La transformation en score z est une transformation linéaire. Elle ne change une distribution bimodale ou asymétrique en distribution normale ! La "forme" de la distribution est conservée.
- L'étalonnage a pour objectif de donner un sens à la mesure. Cet étalonnage, pour qu'il ait un sens, ne devrait être utilisé que si les scores bruts se distribuent normalement ou quasi-normalement (il est toujours possible de transformer un score en note z , mais ce score n'a de sens que si la distribution est normale). Malheureusement beaucoup de psychologues semblent avoir oublié cette règle (même dans certains manuels de tests publiés). Si cette règle n'est pas respectée, l'interprétation du score est complexe car il doit prendre en compte l'importance et la nature de l'asymétrie de la distribution.
- Si la distribution initiale suit une loi normale (ce qui devrait être le cas pour que cette transformation soit intéressante) il est facile de connaître les probabilités d'avoir un score supérieur ou inférieur au score z observée en utilisant une [table de la loi normale](#). Un [calculateur](#) est aussi à votre disposition. Certaines de ces [valeurs sont fréquemment utilisées en psychologie](#) et méritent d'être connues. Par exemples:
 - La presque totalité des scores z (99,7%) se trouvent entre -3 et +3.
 - 95% des scores se trouvent entre -1.96 et +1.96.
 - Une note de 1,96 signifie que l'on est à 1,96 écart-type au dessus de la moyenne (et donc que

seul 2,5% des personnes auraient un score plus élevé).

L'intérêt du z score.

Comme pour tous les scores étalonnés les notes z ont du sens contrairement à un score brut. Ils expriment une distance par rapport à la moyenne des scores d'un groupe dans une unité (fraction d'écart-type) comparable quelle que soit la mesure. Il faut être prudent cependant : la distribution des notes brutes doit être normale ou à minima symétrique et unimodale pour que ces comparaisons aient du sens.

Remarques

- Le QI_{standard} aurait pu s'exprimer facilement en note z. En effet l'écart-type du QI est 15, la moyenne 100, donc un QI de 85 correspond à une note z de $(85-100)/15 = -1$
- L'asymétrie même modérée d'une distribution comme une distribution leptokurtique ou platykurtique modérée, impacte le taux de faux positifs (Crawford & Garthwaite, 2005). C'est la raison pour laquelle l'utilisation d'un score z lorsque la distribution n'est pas normale n'est pas recommandée. Des exemples significatifs sont données par Mathilde Muneaux (2018) dans son "Petit Guide de Psychométrie Clinique à l'Usage des Praticiens" pages 61 à 65.

2.6. Scores Standards Normalisés

Lorsque la distribution ne respecte pas la condition de normalité mais ne s'éloigne pas trop de celle-ci, on peut normaliser la distribution assez simplement en calculant **les scores standards normalisés**. Cette procédure permet d'obtenir un score similaire au score z (à partir des effectifs cumulés). Ce score peut ensuite être transformé en score z, ou en toute autre étalonnage supposant une distribution normale.

Cette procédure pour obtenir un score z est facile à mettre en œuvre :

- Étape 1 : établir les effectifs et les effectifs cumulés
- Étape 2 : calculer les fréquences cumulées (on peut aussi utiliser les rangs percentiles)
- Étape 3 : pour chaque score, lire dans une table de la loi normale la valeur z correspondant au rang percentile ou aux fréquences cumulées. Il est aussi possible d'utiliser les fonctions pré-programmées des tableurs ou des fonctions spécifiques avec le logiciel R (R Core Team, 2025).

Avec cette procédure, pour chaque score possible, un score standard normalisé est calculé (c'est une note z). Ce score peut être utilisé pour calculer des scores T comme dans l'exemple suivant.

Exemple (réalisé avec un tableur). *La première colonne correspond aux scores observés, la seconde aux effectifs, la troisième colonne est l'effectif cumulé et enfin la quatrième colonne est le pourcentage cumulé. Les deux dernières colonnes sont respectivement la note standard normalisée (ou score z obtenu en utilisant la fonction d'un tableur (LOI.NORMALE.INVERSE.N) puis le score T en multipliant la note standardisée par 10 puis en ajoutant 50.*

Scores observés	Effectifs	Effectifs cumulés	Fréquences cumulées	Scores standards normalisés	Scores T
0	2	2	0,0167	-2,13	29
1	2	4	0,0333	-1,83	32
2	2	6	0,0500	-1,64	34
3	4	10	0,0833	-1,38	36
4	4	14	0,1167	-1,19	38
5	4	18	0,1500	-1,04	40
6	6	24	0,2000	-0,84	42
7	4	28	0,2333	-0,73	43
8	8	36	0,3000	-0,52	45
9	6	42	0,3500	-0,39	46
10	12	54	0,4500	-0,13	49
11	10	64	0,5333	0,08	51
12	9	73	0,6083	0,27	53
13	8	81	0,6750	0,45	55
14	6	87	0,7250	0,60	56
15	6	93	0,7750	0,76	58
16	8	101	0,8417	1,00	60
17	7	108	0,9000	1,28	63
18	6	114	0,9500	1,64	66
19	4	118	0,9833	2,13	71
20	2	120	1,0000	>3	>80

Si vous n'avez pas de tableur, pour trouver le score standard normalisé d'une note (par exemple la note 17 dans l'exemple précédent), on cherche dans une table de la loi normale, la valeur z correspondant au pourcentage cumulé (ici 0.900). C'est bien entendu la même que celle calculée avec un tableur !

Pour aller plus loin

La présentation faite ici est une présentation classique simple. En fait, il existe plusieurs procédures pour passer du score observé aux scores standards normalisés, les formules de transformation variant sur un simple paramètre. Il s'agit du paramètre c dans la formule ci-dessous (Procédure de Van der Waerden, c=0 ; Blom, c=3/8 ; Tukey, c= 1/3 ; ou enfin procédure Rankit avec c = 1/2).

$$Y_i^t = \Phi^{-1} \left(\frac{r_i - c}{N - 2c + 1} \right)$$

Il peut y avoir aussi des notes non observées dans l'échantillon. Aux extrémité ce n'est pas grave mais entre les bornes minimum et maximum observées, il faut alors fixer des règles de calcul pouvant être légèrement différentes. On peut aussi s'interroger sur la taille de l'échantillon (il est possible qu'il soit insuffisant pour que toutes les notes soient observés). Ce n'est pas le cas dans notre exemple.

Pour en savoir plus, si cela vous intéresse, cf. l'article de [Solomon & Savilowsky](#) de 2009.

2.7. Autres Scores standards

Il existe de nombreux autres étalonnages que les quantiles, les échelles réduites ou les échelles normalisées. Ces étalonnages correspondent le plus souvent à un transformation du score z qui le plus souvent permettent soit d'avoir une distribution qui n'est plus centrée sur 0 ([score T](#) par exemple), soit de catégoriser les scores ([stanine](#), [sten](#)) ou enfin d'avoir une échelle en 100 classes ([scores NCE](#)).

2.7.1 Le score T

Ce score est similaire au [score z](#), mais la moyenne est de 50 et l'écart-type de 10. Donc pour calculer un score T d'une personne, on multiplie son score z par 10 et on ajoute 50.

$$T_i = z_i * 10 + 50$$

donc si x_i est le score qui doit être transformé

$$T_i = [(x_i - m) * 10 / s] + 50$$

(pour une distribution de moyenne m et d'écart-type s)

Remarques :

- Cette transformation doit être utilisée uniquement lorsque les scores (x_i) se distribuent normalement ou quasi-normalement.
- L'interprétation du score est similaire à celle du score z à une transformation linéaire près. Par exemple :

Une note T de 45 signifie que l'on se situe à 1/2 écart-type en dessous de la moyenne. Cela correspond à une note z de -0.5. Une note de 69.6 signifie que l'on est à 1.96 écart-type au dessus de la moyenne et donc que seul 2,5% des personnes ont un score plus élevé (si la distribution est normale !).

2.7.2 Stanine

Le terme « stanine » trouve son origine dans l'expression anglaise standard nine (échelle standardisée en neuf échelons). Cette échelle repose sur une transformation des scores bruts en neuf catégories correspondant à des intervalles de la distribution normale. Elle est bornée entre 1 et 9, avec une moyenne théorique de 5 et un écart-type d'environ 2 ($m=5, s \approx 2, \min=1 ; \max=9$). Cette échelle permet de représenter la position relative d'un individu au sein de la population de référence sur une échelle discrète et symétrique. Pour construire l'échelle en stanines, deux approches peuvent être envisagées selon la nature de la distribution :

Méthode 1 : la distribution est normale ou quasi-normale.

La démarche est identique à celle d'une échelle réduite en 9 classes (l'étendu de la classe est de 0.5 écart-type). La formule suivante peut être utilisée (pour une distribution de moyenne m et d'écart-type s) :

$$S(x_i) = \begin{cases} 1, & \text{si } 2 \times \frac{x_i - m}{s} + 5 < 1, \\ \text{round}\left(2 \times \frac{x_i - m}{s} + 5\right), & \text{si } 1 \leq 2 \times \frac{x_i - m}{s} + 5 \leq 9, \\ 9, & \text{si } 2 \times \frac{x_i - m}{s} + 5 > 9. \end{cases}$$

L'expression $(x_i - m)/s$ dans la formule précédente correspond à une transformation linéaire du score initial en score z. Cette valeur est ensuite doublée puis on ajoute 5. Le score est arrondi et les valeurs supérieures à 9 sont ramenées à 9 et celles inférieures à 1 sont ramenées à 1.

Méthode 2 : la distribution n'est pas une distribution normale.

La procédure est similaire à celle utilisée pour une échelle normalisée. A partir des scores observés auprès de l'échantillon de standardisation, l'étalonnage de l'épreuve va consister à

associer chaque score possible à une classe (un échelon de l'échelle). Les scores sont ordonnés du plus petit au plus grand. Les 4% premiers scores sont associés à la catégorie 1, les 7% suivant à la catégorie 2, etc.. La table suivante vous indique les pourcentages.

STANINE									
Stanine	1	2	3	4	5	6	7	8	9
% théorique	4%	7%	12%	17%	20%	17%	12%	7%	4%
% cumulé	4%	11%	23%	40%	60%	77%	89%	96%	100%
Z-score	$]-\infty ; -1.75[$	$[-1.75 ; -1.25[$	$[-1.25 ; -0.75[$	$[-0.75 ; -0.25[$	$[-0.25 ; +0.25[$	$[+0.25 ; +0.75[$	$[+0.75 ; +1.25[$	$[+1.25 ; +1.75[$	$[+1.75 ; +\infty[$

Interprétation

Avec le tableau, l'interprétation est simple. Par exemple un stanine de 3 signifie que l'on se situe autour d'1 écart-type en dessous de la moyenne (z autour de -1) et 73% des personnes (population de référence) ont un score en stanine supérieur. Un stanine de 6 correspond à une note z de autour de 0.5 et 23% ont un score supérieur.

2.7.3 Sten

Le terme « sten » trouve son origine dans l'expression anglaise "standard ten" (échelle standardisée en dix échelons). Comme pour les stanines, le score sten sur une transformation des scores bruts en dix catégories correspondant à des intervalles de la distribution normale. Elle est bornée entre 1 et 10, avec une moyenne théorique de 5,5 et un écart-type d'environ 2 ($m=5.5, s \approx 2, \min=1 ; \max = 19$). Deux approches peuvent être envisagées selon la nature de la distribution :

Méthode 1 : la distribution est normale ou quasi-normale.

Une formule similaire à celle des stanines peut être utilisée (pour une distribution de moyenne m et d'écart-type s) :

$$Sten(x_i) = \begin{cases} 1, & \text{si } 2 \times \frac{x_i - m}{s} + 5.5 < 1, \\ \text{round}\left(2 \times \frac{x_i - m}{s} + 5.5\right), & \text{si } 1 \leq 2 \times \frac{x_i - m}{s} + 5.5 \leq 10, \\ 10, & \text{si } 2 \times \frac{x_i - m}{s} + 5.5 > 10. \end{cases}$$

L'expression $(x_i - m)/s$ dans la formule précédente correspond à une transformation linéaire du score initial en score z. Cette valeur est ensuite doublée puis on ajoute 5.5. Le score est arrondi et les valeurs supérieures à 10 sont ramenées à 10 et celles inférieures à 1 sont ramenées à 1.

Méthode 2 : la distribution n'est pas une distribution normale.

Pour construire la table d'étalonnage, la procédure est similaire à celle utilisée pour une échelle normalisée. L'ensemble des scores de l'échantillon de standardisation sont ordonnés du plus petit au plus grand. Les 2.3% premiers scores sont associés à la catégorie 1, les scores des 6.7% suivant à la catégorie 2, etc.. La table suivante vous indique les pourcentages.

STEN										
Stens	1	2	3	4	5	6	7	8	9	10
% théorique	2,30%	4,40%	9,20%	15,00%	19,10%	19,10%	15,00%	9,20%	4,40%	2,30%
% cumulé	2,30%	6,70%	15,90%	30,90%	50,00%	69,10%	84,10%	93,30%	97,70%	100,00%
Z-score	$]-\infty ; -2[$	$[-2 ; -1.5[$	$[-1.5 ; -1[$	$[-1 ; -0.5[$	$[-0.5 ; 0[$	$[0 ; 0.5[$	$[0.5 ; 1[$	$[1 ; 1.5[$	$[1.5 ; 2[$	$[2 ; +\infty[$

Interprétation

Avec le tableau, l'interprétation est similaire à celle des stanine. Par exemple un sten de 3 signifie que l'on se situe autour d'1,25 écart-type en dessous de la moyenne (z autour de -1.25) et 15.9% des personnes (population de référence) ont un score en sten inférieur ou égal

2.7.4 Scores NCE

Ce score NCE ("Normal Curve Equivalent") est peu utilisé en France. Il a été développé pour répondre aux besoins des tests éducatifs, plus particulièrement pour le département de l'éducation des États-Unis. Cette échelle a pour propriété d'être globalement une échelle qui varie entre 1 et 100 (moyenne de 50) et qui se veut similaire dans l'interprétation aux rangs percentiles.

La transformation suppose des distributions normales ou quasi-normales. Elle est similaire à celle présentée pour l'échelle en Stanine ou l'échelle en Sten. Pour passer d'un score brut à un score NCE, on utilise la formule suivante :

$$NCE(x_i) = \begin{cases} 1, & \text{si } 21.063 \times \frac{x_i - m}{s} + 50 < 1, \\ \text{round}\left(21.063 \times \frac{x_i - m}{s} + 50\right), & \text{si } 1 \leq 21.063 \times \frac{x_i - m}{s} + 50 \leq 100, \\ 100, & \text{si } 21.063 \times \frac{x_i - m}{s} + 50 > 100. \end{cases}$$

On en déduit que la moyenne de score est de 50 et l'écart-type de 21.063.

Remarques :

- la valeur de l'écart-type (21.063) a été choisie de façon avoir une échelle de 1 à 100, le rang-percentile 99 correspondant à un NCE de 99, le rang percentile 50 correspondant à un score NCE de 50 et le rang percentile de 1 correspondant à un score NCE de 1.
- La correspondance entre le score NCE et le percentile existe uniquement pour les rangs percentiles 1, 50 et 99 (par construction). Un score NCE est une échelle de 1 à 100 mais n'est pas un rang percentile (cf. tableau ci-dessous pour certains rangs percentiles)

score z	-2,33	-1,28	-0,84	-0,52	-0,25	0,00	0,25	0,52	0,84	1,28	2,33
score NCE	1	23	32	39	45	50	55	61	68	77	99
rang-percentile	1%	10%	20%	30%	40%	50%	60%	70%	80%	90%	99%

- Contrairement aux rangs percentiles, cette échelle est une échelle d'intervalle, ce qui permet de moyennier légitimement les scores en **NCE** (ce qui n'est pas le cas pour les scores exprimés en rangs percentiles).

2.8. Un étalonnage particulier : le QI standard

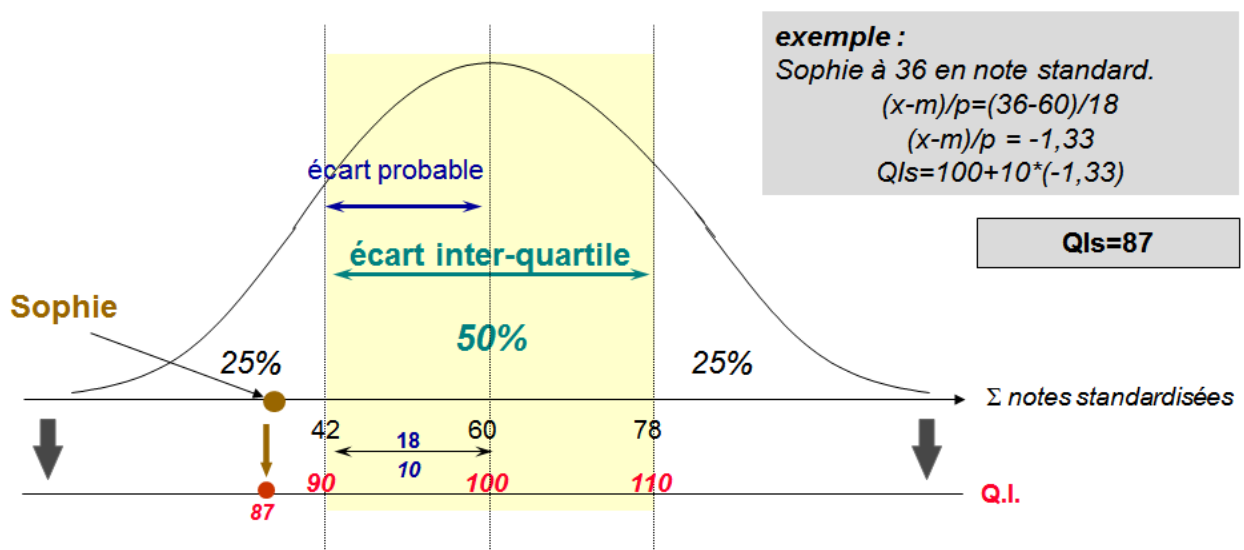
Le quotient intellectuel standard (QIs) est calculé sur la base de plusieurs épreuves (connaissances, raisonnement, attention, mémoire, etc.) et correspond à la somme de scores standardisés ([échelles normalisées](#) en 19 classes) sur chacune de ces épreuves. Une conversion ensuite de ces scores est

réalisée de façon à ce que la distribution de ces scores est une moyenne de 100 et un écart-type 15.

Le QIs est donc un étalonnage particulier proposé par Wechsler (cherchant à respecter les distributions habituelles du [QI classique](#)). La première étape consiste à transformer la distribution des scores bruts sur chacune des épreuves en échelles normalisées (en 19 classes pour les échelles de Wechsler). Ensuite une règle de calcul simple a été fixée, sachant que 50% des scores devaient se trouver entre 90 et 110. L'écart probable ([demi-écart-interquartile](#)) de la somme des notes standardisées est déterminé et permet de calculer le QI en fonction de sa distance à la moyenne exprimée en fraction d'écart-probable.

La moyenne des QI par constructions sera de 100 et l'écart-type de 15. D'autres règles de conversion (donnant les mêmes résultats existent mais c'est celle présentée ici est celle utilisée par Wechsler initialement). Dans les tests, les tables d'étalonnage donnent le résultat des cette conversion en QIs. Le QI mesure donc un écart à la moyenne d'une ou d'un ensemble de performances dans une population ayant passé les mêmes épreuves. Contrairement aux représentations erronées, ce n'est pas à un résultat absolu mais un résultat relatif.

Illustration du mode de transformation d'un score en QIs par Wechsler :



Déterminer un QI standard d'un sujet

- on calcule l'écart qu'il existe entre la note du sujet (x) et la note moyenne (m) en "écart-probable" (p) soit : $(x-m)/p$
- on multiplie cette valeur par 10 et on l'ajoute à 100 : $QIs = 100 + 10 * ((x-m)/p)$

3. Étallonages continus et inférentiels

L'étalonnage est toujours un reflet de la distribution des scores dans une population. La présentation des procédures d'étalonnage paraît simple, mais leur mise en pratique pose plus de problèmes qu'il n'y paraît. En effet :

- 1) Les scores observés peuvent être affectés de façon significative par différents facteurs comme l'âge, le sexe, le niveau d'étude, les professions, etc. Quels facteurs doivent être pris en compte pour déterminer les sous-groupes d'étalonnage (faut-il construire des tables par âge, âge et sexe, par catégorie socioprofessionnelle, etc.) ?
- 2) Quel est le "juste" nombre des sous-groupes (le nombre des tables d'étalonnage) à construire ?

Multiplier les tables d'étalonnage peut conduire à stratifier l'échantillon de standardisation en sous-groupes ayant un nombre de représentants trop restreint. Par exemple, pour une simple variable comme l'âge, dans une épreuve concernant des enfants de 6 à 17 ans, il faudra décider, en fonction de l'importance de l'effet de l'âge, l'opportunité de construire un étalonnage par tranche d'âge d'un an (les 6 ans, les 7 ans, etc.) ou par tranche d'âge de 6 mois, voir 4 mois. Plus le nombre des catégories sera important plus la taille de l'échantillon dans chacune des strates sera potentiellement faible ou très difficile à constituer (pour tenir compte des facteurs à contrôler).

- 3) Lorsque l'on utilise, pour subdiviser l'échantillon de standardisation, une variable comme l'âge la taille de l'intervalle peut avoir des conséquences sur l'interprétation des scores. Si les intervalles sont d'une taille trop importante (relativement à la taille de l'effet) un enfant dont l'âge est à la limite d'une classe pourrait, selon qu'il sera examiné 10 jours avant ou 10 jours après, voir son score brut comparé à des échantillons peu représentatifs pour cet enfant (remarque: en pratique, on devrait toujours regarder, lorsqu'un enfant est à la limite d'une classe d'âge, le score de l'enfant dans les deux groupes d'âge).

Ces quelques questions montrent que construire un étalonnage n'est pas la simple mise en œuvre d'une technique. Les réponses à ces questions sont multiples et on se doit de trouver un compromis entre l'importance de l'effet des facteurs à contrôler (âge, par exemple), la taille de l'échantillon de standardisation et le nombre de groupes de comparaisons (nombre de tables d'étalonnage à construire).

La normalisation continue : une réponse à ces questions

Pour répondre à certains problèmes, les techniques ont évolué et les plus utilisées sont actuellement regroupées sous le nom de «**normalisation continue**» (Lenhard, Lenhard, Suggate et Segerer, 2016, Voncken, Albers et Timmerman, 2016, Zachary et Gorsuch, 1985), ou «**normalisation inférentielle**» (Zhu & Chen, 2011). De façon très résumée, le principe de ces méthodes consiste à modéliser les paramètres de la distribution des scores des tables d'étalonnage par des techniques de régression.

L'intérêt de certaines de ces techniques est qu'elles permettent d'estimer les caractéristiques de chaque groupe (donc chaque table d'étalonnage) en prenant en compte l'ensemble de l'échantillon et non plus simplement le sous-groupe concerné par la table. Elles peuvent parfois permettre d'extrapoler des données ou de s'affranchir des tables d'étalonnage en permettant de calculer pour chaque personne (chaque score observé) un score étalonné en fonction des variables que l'on souhaite contrôler (cet aspect est encore peu développé et suppose pour l'utilisateur final d'utiliser non plus des tables d'étalonnage mais un algorithme de calcul pour convertir le score, algorithme le plus souvent automatisé via une application web ou non).

Ces techniques apportent une contribution importante à la construction des étalonnages mais impliquent de respecter des hypothèses comme la normalité des scores bruts et/ou l'homogénéité des variances. Malheureusement, ces hypothèses sont rarement vérifiées dans les échelles surtout lorsqu'il s'agit d'épreuves développementales (asymétrie droite de la distribution pour les plus jeunes par exemple). Des solutions ont été proposées (normalisation des distributions par des transformations comme celle de Box-Cox) mais restent limitées. Une des solutions prometteuses est probablement celle proposée par Lenhard, Lebard, Suggate et Segerer en 2016 qui est une technique non paramétrique s'appuyant sur les polynômes de Taylor. Elle ne sera pas détaillée ici mais vous pouvez voir un exemple de mise en œuvre dans la batterie FEE évaluant les fonctions exécutives chez l'enfant (Roy et al. 2021).

Pour conclure, quel que soit la technique utilisée, le score étalonné sera exprimé sous forme d'un score T, d'un QI, d'un score z ou autres. L'interprétation pour le psychologue reste identique, seule la façon de

construire l'étalonnage change de façon à ce qu'il soit optimisé. On regrettera cependant que les auteurs, lorsqu'ils utilisent la normalisation continue, donnent trop peu de détails sur le modèle utilisée.

4. Correspondance entre étalonnages

Il est assez facile de passer d'un étalonnage à un autre sous l'hypothèse que la distribution des scores étalonnés est normale. La figure ci-dessous donne quelques correspondances. Ces correspondances sont très faciles à calculer soit même.

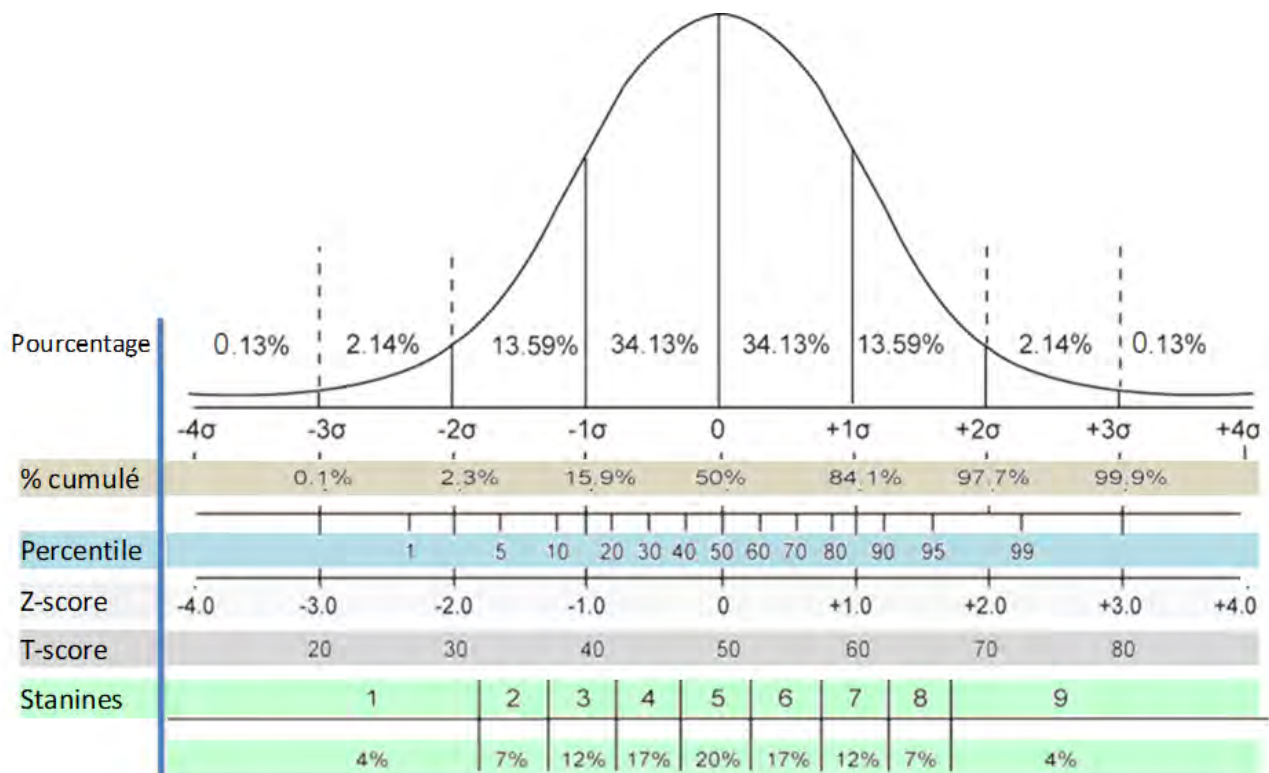


Figure E.1 : Correspondance entre différents type d'étalonnage (pour une distribution normale)

5. Détermination d'un score seuil

Certains tests proposent des valeurs seuils ou critiques permettant de définir un risque (difficulté d'apprentissage de la lecture, pronostic de démence, dépression, etc.). Ces valeurs seuils sont parfois fixées a priori à partir de critères comme un score inférieur ou supérieur à 2 écarts-types à celui observé en moyenne. Cependant, quand un test permet de prédire l'apparition d'une maladie ou des difficultés d'apprentissage, il est possible, d'utiliser les notions de sensibilité et spécificité telles que nous les avons vues dans le chapitre ([Qualités métrologiques - Sensibilité et spécificité](#)).

Pour rappel **la sensibilité** dans ce contexte est la capacité de l'instrument à identifier correctement les personnes présentant la caractéristique étudiée et **la spécificité** est la capacité de l'instrument à identifier correctement les personnes ne portant pas cette caractéristique.

Supposons la construction d'une batterie permettant d'évaluer la mémoire sous toutes ses formes et donnant un score global de mémoire pour les personnes âgées entre 70 et 75 ans qui présentent des plaintes mnésiques (vie quotidienne). L'hypothèse formulée est que les résultats à cette batterie, lorsqu'ils sont élevés (les scores de performances sont inversés), devraient être aussi prédictifs

d'une évolution vers une démence dans les années à venir (exemple fictif). L'épreuve est soumise à un échantillon représentatif de cette population et les résultats sont mis en perspective avec l'évolution des personnes ayant subies ce test. Deux groupes sont constitués : ceux présentant une démence et ceux ne présentant pas de démence. Il est possible de représenter les résultats initiaux à cette batterie sur un graphique en séparant ceux qui présentent un trouble et ceux qui n'en présentent pas :

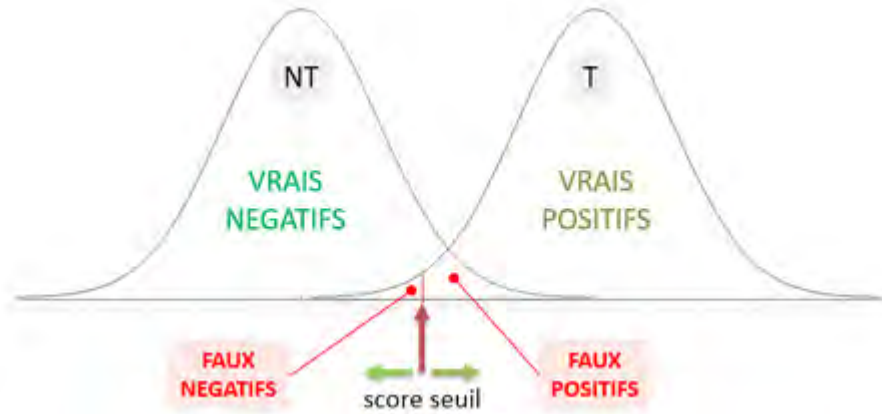


Figure E.2 : Distribution des scores (proportions) des personnes qui présentent un trouble (T) et de ceux qui n'en présentent pas (NT)

Parmi ceux qui présentent des troubles, le score initial peut être faible (sans difficulté particulière) et inversement, parmi ceux qui ne présentent pas de trouble, le score initial peut être élevé. Si les deux courbes sont confondues ou presque confondues (moyenne proche) il ne sera pas possible de trouver une valeur critique. Si ces deux courbes sont suffisamment distinctes, il faudra se fixer une valeur critique en minimisant les faux positifs (FP) et les faux négatifs (FN).

Le graphique précédent permet de comprendre que selon la valeur seuil que l'on prendra fixée, soit on diminue la probabilité de FP (faux positifs) mais on augmente la probabilité d'avoir des FN (faux négatifs), soit on diminue la probabilité des FN mais on augmente celle des FP. Le bon positionnement dépend des risques que l'on veut prendre et de la nature de la décision à prendre. Si, comme dans notre exemple, l'objectif est d'avoir une valeur critique pour identifier les personnes à risque de démences, il peut être préférable d'avoir des FN plutôt que des FP connaissant l'impact du diagnostic dans l'évolution de ces maladies. A l'inverse si, pour une autre recherche avec des enfants, l'objet est d'identifier des possibles troubles d'apprentissage ultérieurs (lors de la scolarisation obligatoire), n'est-il pas à préférable de faire un minimum de FN ?

Ce rapport entre FN et FP et la qualité diagnostic de l'épreuve peut être évaluée au moyen d'une courbe que l'on appelle courbe ROC (Receiver Operating Characteristic). Cette courbe permet de montrer à quel point un test arrive à fait correctement la différence entre deux groupes (par exemple, malades et non malades).

Pour tracer cette courbe, on place en abscisse la valeur de $1 - \text{spécificité}$ et en ordonnée la sensibilité (pour le calcul de ces valeurs, voir [Qualités métrologiques - Sensibilité et spécificité](#)). Le seuil de décision est progressivement modifié de manière à faire varier la spécificité de 1 à 0. Pour chaque valeur du seuil, le couple de coordonnées ($1 - \text{spécificité}$, sensibilité) est reporté sur le graphique. Cette approche permet ainsi de décrire l'évolution conjointe des faux positifs (FP) et des faux négatifs (FN) en fonction du seuil de décision retenu.

Lorsque la courbe ROC se rapproche de la diagonale, la capacité du test à discriminer entre les deux états (présence ou absence de la condition) diminue, traduisant une classification équivalente à un

tirage aléatoire. Dans ce cas, la surface comprise entre la courbe et la diagonale est faible. À l'inverse, un test diagnostique de qualité se caractérise par une courbe ROC située nettement au-dessus de la diagonale, indiquant une meilleure performance discriminante

Le choix du seuil de décision doit être effectué en fonction du compromis acceptable entre la spécificité et la sensibilité, en tenant compte des conséquences associées aux erreurs de classification (faux positifs et faux négatifs). En pratique, le seuil correspondant au point le plus proche de la coordonnée (0,1) est souvent privilégié, car il représente un bon compromis entre la détection correcte des cas positifs et la minimisation des erreurs de classification.

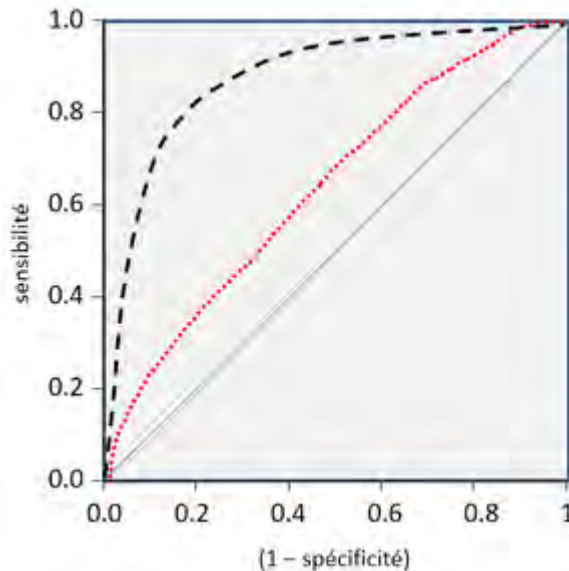


Figure G.3 : Exemples de deux courbes ROC. En rouge (pointillé) une courbe ROC associée à un mauvais test diagnostique et en noir (traitillé) une courbe ROC associée à un meilleur test diagnostique (rem : la spécificité et la sensibilité varient entre 0 et 1).

Pour évaluer la qualité globale des prédictions, il est possible de calculer l'AUC (Area Under the Curve) qui est l'aire sous l'ensemble de la courbe. Les valeurs d'AUC varient de 0 à 1. Un AUC de 0 correspond à un modèle dont 100 % des prédictions sont fausses. À l'inverse, si toutes les prédictions sont correctes, l'AUC est de 1. En règle générale, une AUC comprise entre 0,7 et 0,9 traduit une performance diagnostique satisfaisante ; une AUC supérieure à 0,9 indique une très bonne performance. Une AUC inférieure à 0,5 suggère une inversion du critère de classification (le test classe les cas négatifs comme positifs, et inversement)

Pour calculer cette surface, il existe plusieurs méthodes mais on peut utiliser la formule suivante :

$$AUC = \frac{W_1 - \frac{n_1(n_1 + 1)}{2}}{n_1 n_0}$$

W_1 : Après avoir classé toutes les observations par ordre croissant, W_1 correspond à la somme des rangs des personnes positives.

n_1 et n_0 : ce sont respectivement le nombre de personnes positives et le nombre de personnes négatives.

La présentation qui est faite ici des courbes ROC est très simplifiée. Cet outil est très utilisé dans le domaine médical et plus rarement en psychologie. Pour aller un peu plus loin sur la compréhension des courbes ROC, vous pouvez lire un article introductif (Morin, Morin, Mercier, Moineau, & Codet, 1998) dans le domaine médical ou ces articles montrant une application en psychologie (Lacot et al., 2011, Pinte, Moldovan, 2009).

F - Intervalles de Confiance et comparaison de scores

1. Introduction à la notion d'IC

Le score obtenu à un test ne représente pas nécessairement la performance ou le score véritable de la personne (score vrai), car le score observé dépend également des erreurs de mesure et de divers facteurs aléatoires. Le psychologue doit donc calculer ce que l'on appelle un **intervalle de confiance (IC)**. Cet intervalle de confiance est parfois mal compris et il ne signifie pas que le score vrai se trouve dans l'intervalle de confiance mais indique la plage de valeurs les plus plausibles compte tenu du degré de confiance choisi.

Par exemple, si l'on fixe un degré de confiance de 95 %, cela signifie que si l'on répétait de très nombreuses fois exactement la même mesure (mais ce n'est pas possible), environ 95 % des intervalles de confiance ainsi calculés contiendraient la valeur vraie du score, et 5 % ne la contiendraient pas. Cet intervalle est donc plus informatif que le score observé seul. Si le degré de confiance choisi est plus petit (par ex. 90 %), l'intervalle sera plus étroit pour le même score observé ; inversement, si l'on augmente le degré de confiance (par ex. 99 %), l'intervalle sera plus important.

Pour le praticien, cette notion d'intervalle de confiance est essentielle :

- Il permet de relativiser la note et de prendre conscience de la marge d'erreur (qui est souvent plus importante que l'on croit). L'intervalle de confiance devrait toujours accompagner le score.
- Il permet d'apprécier aussi, ou de rappeler, la qualité de l'épreuve utilisée (précision). En effet, si l'intervalle de confiance (pour une probabilité habituellement utilisée par ce psychologue) est étroit, cela signifie que l'instrument utilisé est précis (fidèle). A l'inverse, si cet intervalle est important, l'instrument est peu précis.

Dans les manuels de tests, les intervalles de confiance sont souvent donnés avec des valeurs usuelles qui sont 68%, 90% ou 95%. On peut aussi les calculer soi-même (cf. intervalle de confiance - [méthode classique](#) et [méthode alternative](#)). Il existe plusieurs méthodes de calcul de cet intervalle de confiance qui seront présentées dans la suite de ce chapitre.

A savoir :

Dans l'approche fréquentiste utilisée pour calculer l'IC, le paramètre inconnu (le score vrai) est considéré comme fixe. La probabilité calculée (degré de confiance) ne concerne donc pas la valeur vraie mais l'intervalle de confiance.

- **Le contre-sens habituel** est donc de croire que l'IC est l'intervalle dans lequel le score vrai a une forte probabilité de se trouver.
- **En pratique, l'intervalle de confiance représente la plage de valeurs du score vrai qui sont les plus cohérentes avec le score observé, selon le niveau de confiance choisi.**
- L'intervalle de confiance à 95 % est construit de manière à ce que, si on pouvait répéter le mesure (sans qu'il y ait un changement du score vrai), environ 95 % des intervalles ainsi obtenus contiendraient le score vrai.

2. Calculer un IC pour un score observé

Dans le cadre de la théorie classique des tests deux méthodes de calcul de l'intervalle de confiance

existe :

- La première, centrée sur le score observé, peut être considérée comme [la méthode classique](#) (la plus fréquente).
- La [seconde méthode](#) centre l'intervalle de confiance non pas sur le score observé mais sur un score vrai estimé.

Ces deux méthodes utilisent [l'erreur standard de mesure](#) (ESM ou SEM en anglais) pour évaluer l'intervalle de confiance et sont encore largement utilisées. Elles supposent cependant que l'ESM est constant quel que soit le score observé, ce qui est faux généralement. En effet l'erreur standard de mesure pourrait doubler aux extrémités de la distribution. Les recommandations du "[Standards for Educational and Psychological Testing](#)" (2014) aux éditeurs de tests indiquent que normalement, on devrait calculer une erreur standard de mesure conditionnelle (C-ESM ou CSEM) pour chaque valeur observée, ou pour des intervalles de valeurs : "*The standard error of measurement, both overall and conditional (if reported), should be provided in units of each reported score.*" (standard 2.13, p.45).

Les méthodes d'estimations de l'erreur standard de mesure conditionnelle (C-SEM) ne sont cependant pas développées dans ce cours. Ces méthodes sont nombreuses et donnent des résultats proches. Pour ceux que cela intéresse, une présentation claire de ces méthodes est celle de [Tong & Kolen \(2005\)](#) dans "Encyclopedia of Statistics in Behavioral Science". Il existe aussi une méthode faisant référence à la théorie de la généralisabilité (pour une introduction, cf. Laveault et Grégoire, 2014) mais cette méthode est essentiellement réservée à l'évaluation dans le cadre des sciences de l'éducation même si sa mise en œuvre peut concerner les tests mentaux.

Remarque : un outil de calcul de l'intervalle de confiance est à votre disposition <[Téléchargement ICI](#)> . Il s'inspire fortement des outils proposés par J. W. Schneider (<https://wjschne.github.io>)

2.1. Erreur standard de mesure et TCT

L'Erreur Standard de Mesure (ESM) plus souvent appelé SEM (notation anglo-saxonne) est un indicateur de l'importance de la variabilité de l'erreur de mesure (le carré de l'erreur standard de mesure est la variance d'erreur observée pour un test dans une population). L'ESM s'exprime en fonction du [coefficient de fidélité](#) (rappel : le coefficient de fidélité représente la part de variance correspondant à des différences vraies, non aléatoires, entre les personnes). Soit s_x l'écart-type des scores dans la population et r_{xx} le coefficient de fidélité, on calcule le ESM à partir de la formule suivante :

$$ESM = \sigma_x \sqrt{(1 - r_{xx})}$$

Dans l'absolu, l'ESM correspond donc à l'écart-type des scores observés sur des mesures parallèles répétées pour une personne ayant une note "vraie" fixe (invariable).

Remarques :

- Plus la fidélité d'un test est bonne, plus l'ESM est petit. L'ESM varie en fonction du coefficient de fidélité.
- L'ESM permet de relativiser l'importance accordée au score (cf. [intervalle de confiance](#))
- L'erreur standard de mesure ne doit pas être confondue avec l'écart-type (racine carrée de la variance) et l'erreur-type qui est, pour un échantillon donné, l'écart-type divisé par la racine carrée du nombre de sujet (cf. tableau ci-dessous)
- Dans certaine traduction française l'ESM est appelé erreur type de mesure.

Variance	$\frac{\sum(x_i - m)^2}{n}$	Une mesure de la dispersion autour d'une valeur (la moyenne)
Écart-Type	$\sqrt{\frac{\sum(x_i - m)^2}{n}}$	Une mesure la dispersion autour d'une valeur (la moyenne)
Erreur-Type	$\frac{\sqrt{\frac{\sum(x_i - m)^2}{n}}}{\sqrt{n}}$	Une mesure standard de l'erreur d'échantillonnage (c'est donc l'écart type de l'estimateur de la moyenne pour un échantillon).
Erreur Standard de Mesure	$\sqrt{1 - r_{xx}} * \sqrt{\frac{\sum(x_i - m)^2}{n}}$	Écart-type de la distribution de l'erreur de mesure.

avec : n le nombre d'individu, m la moyenne et r_{xx} le coefficient de fidélité.

2.2. Erreur standard de mesure et MRI (C-ESm)

Dans le cadre des modèles de réponse à l'item (MRI), l'erreur standard de mesure (ESM) peut être calculée pour chaque valeur du trait latent à partir de la [courbe d'information](#) $I(\theta)$. Elle sera d'autant plus faible que l'information apportée est élevée. Cette erreur de mesure, pour chaque valeur de θ , est égale à 1 sur la racine carrée de l'information apportée :

$$ESM(\theta) = \frac{1}{\sqrt{I(\theta)}}$$

$I(\theta)$ est l'information apportée pour la valeur θ .

Contrairement à la théorie classique des tests, l'erreur de mesure varie donc en fonction du trait latent (on parle d'erreur standard de mesure conditionnelle ou C-ESM). On représente souvent sur un même graphique la courbe d'information et l'erreur standard de mesure (cf. ci-dessous). On notera que dans ce type de représentation graphique, les échelles relatives à ces deux courbes ne sont pas les mêmes.

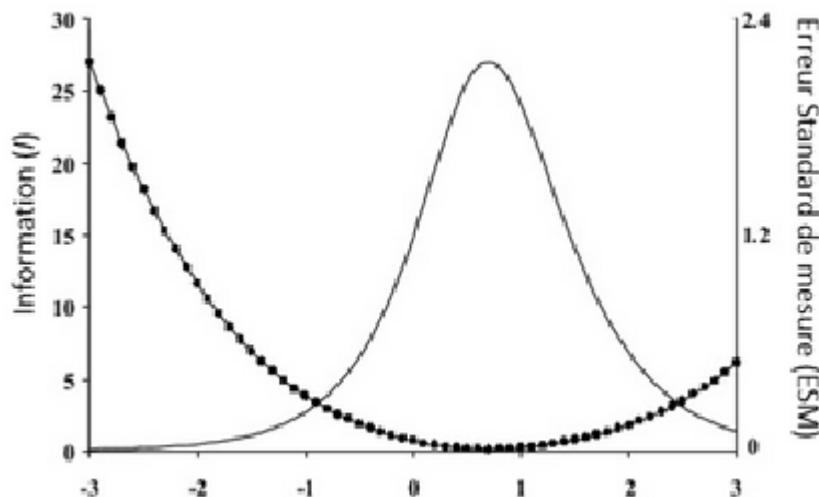


Figure F.1 : C-ESM (courbe en pointillé) et courbe d'information (trait continu)
d'un test pour différentes valeurs de θ (trait latent)

2.3. Méthode classique

Dans la méthode classique de calcul de l'intervalle de confiance, on calcule l'[Erreur Standard de Mesure](#) (ESM) puis l'intervalle de confiance en appliquant la formule suivante :

$$[x - u * ESM ; x + u * ESM]$$

avec X , le score observé et u la valeur lue dans la table de la loi normale

Rappel : *pour un intervalle de confiance est de 95% ; $u = 1,96$
pour un intervalle de confiance est de 90% ; $u = 1,644$
pour un intervalle de confiance est de 70% ; $u = 1,04$
pour un intervalle de confiance est de 68% ; $u = 1$*

Remarques

- De cette formule on peut déduire que plus la fidélité est faible plus l'intervalle de confiance sera important pour une probabilité donnée (ce qui traduit que le score observé peut être très éloigné du score vrai). La connaissance de la fidélité permet donc de relativiser un score numérique unique.
- Dans la méthode classique on centre l'intervalle de confiance sur le score observé. En centrant ainsi l'intervalle de confiance on postule que le score observé est une estimation correcte (sans biais) du score vrai. Sachant que la corrélation entre le score observé et le score vrai n'est jamais parfaite il existe alors nécessairement un phénomène de régression à la moyenne (cf. [glossaire](#)). Cela signifie que les scores supérieurs à la moyenne sont souvent surestimés et les scores inférieurs à la moyenne sont souvent sous estimés. C'est la raison pour laquelle on utilise de plus en plus souvent la méthode recommandée par [Glutting, McDermott et Stanley](#) (1987), méthode qui ne centre plus l'intervalle de confiance sur la note observée et qui utilise non plus l'erreur standard de mesure mais une estimation (ESM_E) du ESM (cf. [méthode alternative](#)).
- Le calcul de l'intervalle de confiance suppose le respect du postulat d'homoscédasticité, c'est-à-dire une variance constante des erreurs de mesure à travers les différents niveaux de score observé. Ce postulat est discutable car l'ESM varie en fonction du niveau de la mesure. Lorsque c'est possible il est donc conseillé de calculer un C-ESM et donc un IC par score observé (très peu d'épreuves appliquent actuellement cette recommandation).

2.4. Méthode corrigée

La méthode classique de calcul de l'intervalle de confiance centre celui-ci sur le score observé. Il existe d'autres méthodes pour calculer l'intervalle de confiance qui tiennent compte du [phénomène de régression à la moyenne](#). Après l'analyse de différentes méthodes, Glutting, McDermott et Stanley (1987) recommandent de centrer l'intervalle de confiance sur une note vraie estimée (x_{ti}) et d'utiliser

un ESM estimé (ESM_E). Les formules deviennent :

$$x_{ti} = \bar{x} + r_{xx}(x_i - \bar{x})$$

$$ESM_E = \sigma_x \sqrt{(1 - r_{xx})} \times r_{xx}$$

$$[x_{ti} - u * ESM_E ; x_{ti} + u * ESM_E]$$

Avec cette méthode, on observe que l'intervalle de confiance est centré sur un score plus proche de la moyenne que pour le score observé et par ailleurs, l'intervalle de confiance est plus petit que celui calculé avec la formule classique. C'est cette méthode de calcul qui est maintenant utilisée dans les échelles de Wechsler.

Remarque. L'erreur standard de mesure estimée (ESM_E) peut être calculée directement à partir de l'erreur standard de mesure (ESM) en appliquant la formule suivante :

$$ESM_E = ESM \times r_{xx}$$

2.5. Exemples de calcul

Pour un test de facteur numérique, le score d'une personne est de 54. Sachant que la fidélité de ce test est de .92, que la moyenne est de 50 et l'écart-type de 10, calculer les bornes de l'intervalle de confiance dans lequel le score vrai à 68% de se trouver en centrant cet intervalle sur le score observé (méthode traditionnelle) puis en utilisant la formule recommandée par de Glutting, McDermott & Stanley.

Borne de l'intervalle ou le score vrai à 68% de se trouver - centré sur le score observé (x=54)

Étape 1 : calcul du ESM

$$s = 10 \quad [\text{écart-type}]$$

$$r_{xx} = .92 \quad [\text{fidélité}]$$

$$ESM = 10 * \sqrt{(1-.92)} \quad [\text{cf. formule}]$$

$$= 10 * \sqrt{0.08}$$

$$= 2.8284$$

Étape 2 : calcul de l'intervalle de confiance

$$IC = [x - u * ESM ; x + u * ESM]$$

$$u = 1 \quad [\text{pour un intervalle avec 68\% de trouver le score vrai}]$$

$$x = 54 \quad [\text{le score observé}]$$

Le score vrai est donc à plus ou moins 1*ESM, donc plus ou moins 2.8284

$$IC = [54 - 2.8284 ; 54 + 2.8284]$$

$$IC = [51.17 ; 56.83]$$

Borne de l'intervalle ou le score vrai à 68% de se trouver - Méthode recommandée par Glutting, McDermott et Stanley.

Étape 1 : calcul du ESM_E

$$s = 10 \quad [\text{écart-type}]$$

$$r_{xx} = .92 \quad [\text{fidélité}]$$

$$ESM_E = 10 * .92 * \sqrt{(1-.92)} \quad [\text{cf. formule}]$$

$$= 9.2 * \sqrt{0.08}$$

$$= 2.6022$$

Étape 2 : calcul du centre de l'intervalle de confiance

$$\begin{aligned} m &= 50 && [\text{moyenne}] \\ x_{ti} &= m + r_{xx} * (x - m) && [\text{cf. formule}] \\ &= 50 + .92 * (54 - 50) \\ &= 53.68 \end{aligned}$$

Étape 3 : calcul du centre de l'intervalle de confiance

$$\begin{aligned} u &= 1 && [\text{pour un intervalle avec 68\% de trouver le score vrai}] \\ x_{ti} &= 53.68 && [\text{le score observé}] \\ \text{Le score vrai est donc à plus ou moins } 1 * ESM_{\xi} &&& \text{soit } 2.6022, \text{ de } x_{ti} \\ \text{IC} &= [53.68 - 2.6022 ; 53.68 + 2.6022] \end{aligned}$$

$$\text{IC} = [51.08 ; 56.28]$$

Remarque : si vous utilisez l'outil mis à disposition <ICI> pour faire ces calculs, vous obtiendrez les mêmes résultats aux arrondis près.

3. Différence entre deux scores

Il est parfois intéressant de savoir si deux scores observés correspondent à une différence réelle ou s'il s'agit d'une simple variation normale compte tenu de l'erreur de mesure. Cette situation se rencontre lorsque l'on veut comparer les scores de 2 personnes (situation rare dans le cadre de la pratique) ou lorsque l'on veut comparer 2 scores d'une même personne à deux épreuves différentes (plus fréquent). Enfin, un cas particulier concerne la comparaison de 2 scores d'une même personne à la même épreuve mais à des moments différents.

Attention : pour que la comparaison entre deux scores soit possible il faut nécessairement que les échelles de mesure utilisées soient similaires. Il est absurde de comparer deux scores exprimés dans des échelles différentes (cf. [les étalonnages](#)).

3.1. Méthode de comparaison

La méthode utilisée est la même que celle présentée dans le cadre du calcul de l'intervalle de confiance d'un score observé. Cette fois on utilise l'erreur standard de mesure de la différence (ESM_{diff}). Dans la théorie classique des tests, les erreurs ne sont pas corrélées. On peut en déduire que l'ESM de la différence (ESM_{diff}) se calcule à partir de l'erreur standard de mesure de l'épreuve 1 (ESM_1) et l'erreur standard de mesure de l'épreuve 2 (ESM_2) :

$$ESM_{diff} = \sqrt{ESM_1^2 + ESM_2^2}$$

Sachant que pour comparer deux scores il est nécessaire que les échelles de mesures soient semblables, la formule précédente peut-être remplacée par une formule équivalente par :

$$ESM_{diff} = s \sqrt{2 - r_{xx} - r_{yy}}$$

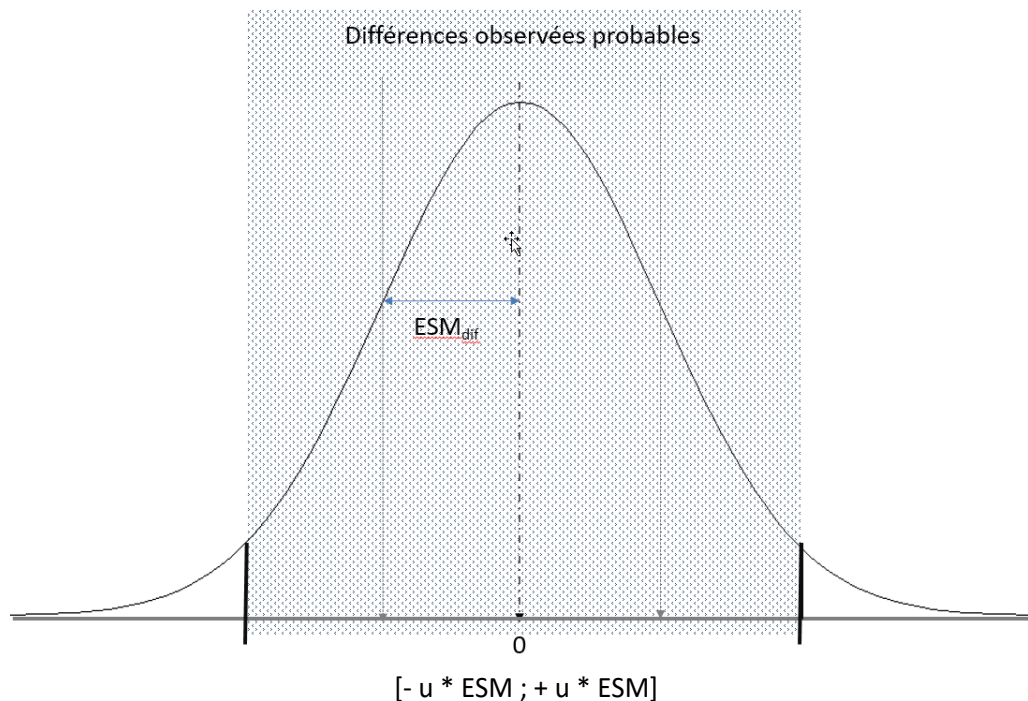
avec :

s = l'écart-type identique des deux épreuves (x et y)

r_{xx} et r_{yy} : la fidélité (r_{xx} et r_{yy}) des deux épreuves x et y

Pour dire que la différence entre 2 scores est suffisamment importante on s'appuie sur le fait que la distribution du score de différence (sous l'hypothèse d'absence de différence) suit une loi normale de

moyenne 0 et ayant pour écart-type ESM_{diff} . Comme pour le calcul de l'intervalle de confiance d'un score observé, on fixe alors un seuil (5% ou encore 1%) et on lit dans une table de la loi normale la valeur u . On multiplie u par le ESM_{diff} pour connaître les deux bornes de l'intervalle de confiance. Si, la différence observée est à l'extérieur de cet intervalle, on peut conclure que la différence entre les deux scores est statistiquement significative, c'est-à-dire qu'elle dépasse ce qu'on pourrait attendre du simple fait de l'erreur de mesure.



avec u la valeur lue dans la table de la loi normale

3.2. Exemple de calcul

Pour un test de facteur numérique, le score de Pierre est de 54 et celui de Lucas de 63. Sachant que la fidélité de ce test est de .92 et l'écart-type de 10, peut-on dire que ces deux scores sont différents (au seuil de 5%) ?

Étape 1 : calcul du ESM

$$\begin{aligned}
 s &= 10 && \text{[écart-type]} \\
 r_{xx} &= .92 && \text{[fidélité]} \\
 ESM &= 10 * \sqrt{(1-.92)} && \text{[cf. formule]} \\
 &= 10 * \sqrt{0.08} \\
 &= 2.8284
 \end{aligned}$$

Étape 2 : calcul du ESM_{diff}

$$\begin{aligned}
 ESM_{diff} &= \sqrt{(2 * ESM^2)} && \text{ ; puisqu'il s'agit ici de la même épreuve} \\
 &= 4 && \text{ ; 3.9858 (mais on simplifie pour l'exemple)}
 \end{aligned}$$

Étape 3 : calcul de l'intervalle I

$$\begin{aligned}
 u &= 1,96 && \text{[pour un intervalle avec 95%]} \\
 I &= [- 1.96 * 4; 1.96 * 4] \\
 I &= \mathbf{[-7.84 ; 7.84]}
 \end{aligned}$$

Étape 4 : Conclusion

La différence entre Pierre et Lucas est de 9 (63-54). Cette différence est supérieure à 7.84, les deux performances sont donc statistiquement (seuil de 5%).

4. Indice de changement fiable

Une nouvelle mode en psychologie et plus particulièrement en neuropsychologie et de calculer, à partir des indices de fidélité un Indice de Changement Fiables (indice RCI en anglais pour "Reliable Change Index"). Cet indice permet de déterminer si la différence observée dans les scores observés d'un individu entre deux passations correspond à une réelle évolution de "score vrai" (non observable) ou est simplement la conséquence de l'erreur de mesure et/ou aux effets de la pratique (c'est-à-dire le fait de passer le test plus d'une fois). Les praticiens peuvent donc calculer le RCI pour améliorer la pratique clinique et obtenir des informations supplémentaires sur les effets d'une prise en charge.

Formule générale du RCI :

$$RCI = \frac{|X_2 - X_1|}{S_{diff}}$$

avec S_{diff} l'écart-type associée à la différence entre deux scores

Pratique (un outils de calcul sera prochainement disponible dans les [outils SCALP](#))

La formule la plus souvent utilisée pour calculer un indice de changement est celle basée sur les travaux de Jacobson et de ses collègues, en particulier Jacobson et Truax (1991) qui ont incorporé une correction de formule fournie par Christensen et Mendoza (1986). Elle nécessite de calculer écart-type associée à la différence observée entre deux passations (S_{diff}) pour calculer un intervalle de confiance.

› Calculs

- calcul de l'erreur standard de mesure (ESM) avec ET comme écart type dans l'échantillon de standardisation et r_{xx} la fidélité du test :

$$ESM = \sigma_x \sqrt{(1 - r_{xx})}$$

- Calcul de l'écart-type associée à la différence entre deux scores

$$S_{diff} = \sqrt{2} * ESM$$

Pour conclure deux méthodes peuvent être utilisées (elles donnent le même résultat)

Méthode 1

- Calcul du RCI (formule ci-dessus)
- Comparer la valeur trouver au valeurs lues dans la table de la loi normale (test est bilatéral). Par exemple le changement est fiable au seuil que l'on se fixe (1%, 5%, 10%) si :

Pour un seuil de 10% : $RCI > = 1.64$

Pour un seuil de 5% : $RCI > = 1.96$

Pour un seuil de 1% : $RCI > = 2.58$

Méthode 2

- Calcul de la valeur absolue de la différence (Δ) entre la passation 2 et la passation 1 : $X_{diff} = |X_2 - X_1|$

→ Calcul d'une valeur seuil pour le score de différence entre la passation 2 et la passation 1

$$\text{Pour un seuil de 10\% (} p < .10 \text{) : } \Delta_{(90\%)} = 1.64 * S_{diff}$$

$$\text{Pour un seuil de 5\% (} p < .05 \text{) : } \Delta_{(90\%)} = 1.96 * S_{diff}$$

$$\text{Pour un seuil de 1\% (} p < .01 \text{) : } \Delta_{(99\%)} = 2.58 * S_{diff}$$

Remarques : en général on arrondit à l'entier le plus proche. Les valeurs (1,64, 1.96 et 2,58) sont les valeurs lues dans la table de la loi normale et le test est bilatéral.

→ Décision : si le X_{diff} est supérieur à la valeur seuil que l'on retient, le changement peut être considéré comme fiable au seuil que l'on s'est fixé. Dans le cas contraire la différence ne peut pas être considérée comme fiable (elle peut être due plus probablement à des facteurs aléatoires)

› Remarques

- Dans ce contexte, le terme de changement fiable signifie que l'on peut affirmer de manière défendable que la différence observée sur les scores est plus importante que celui dû à la seule erreur de mesure.
- Un patient peut avoir un indice de changement fiable significatif, mais rester dans la zone cliniquement significative. Le RCI nous apprend rien sur le niveau de difficulté du patient, il nous renseigne sur l'existence d'un changement de niveau uniquement (présent vs absent).

Critiques, précautions et limites

Cet index, bien qu'il soit de plus en plus utilisé doit être cependant manipulé avec précaution en clinique. Il existe de nombreux articles dans la littérature soit pour l'amender, soit pour proposer des variantes plus ou moins complexes (Wise 2004, Speer 1992, Hageman & Arrindell, 1993). Il faut savoir en particulier (et entre autres) que :

- On doit vérifier l'effet de la pratique pour différent délai de passation. En effet cette formule de calcul du RCI suppose que dans l'intervalle de temps observé entre les deux mesures, il n'existe pas de modification du score ou que cette modification est faible dans l'échantillon de standardisation ou de référence.
- Parmi les nombreux débats concernant cet indice, le plus important repose sur le fait que cet indice fait intervenir le coefficient de fidélité. La question qui est posée et qui n'est pas résolue (Blampied, 2022) concerne la façon dont on estime la fidélité. En effet, même si très souvent la fidélité test-retest est recommandée, il n'est pas toujours évident qu'il s'agisse de la bonne solution. La corrélation r de Pearson du test-retest couramment rapportée est problématique car elle est insensible à l'erreur systématique entre les mesures répétées (Aldridge et al., 2017).
- Cette façon de calculer le RCI s'inscrit dans le cadre de la théorie classique des tests et suppose que l'erreur de mesure (donc la fidélité) ne varie pas en fonction du score observé ce qui est le plus souvent faux. Or lorsque l'on travaille en clinique, les patients ont plus souvent des scores situés aux extrémités de la distribution !

La discussion sur le RCI est donc loin d'être close dans la littérature (Brooks, et al., 2011, Blampied, 2022) et l'utilisation de cet indice doit être fait avec précaution. Il s'agit d'un élément d'information à intégrer dans l'évaluation d'un effet. Il n'existe pas de rituel mécanique (Cohen, 1994) dont le recours

nous dispenserait de la nécessité d'exercer notre jugement. C'est clairement le cas lors de l'utilisation d'un index de changement fiable.

G - Complément : l'échantillonnage

Un test mental est un outil qui permet de positionner ou de comparer la production d'une personne (comportements, performances) à un groupe de référence (population). Il n'est jamais possible d'enregistrer a priori les productions de l'ensemble de la population de référence et on utilise, lors de la construction du test, un ou des sous ensembles de la population (échantillons de personnes) issus du groupe de référence dans lequel on veut utiliser le teste.

L'échantillonnage est donc l'opération par laquelle on sélectionne 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".

Remarques

- Lors de l'analyse critique d'un test, avant de l'utiliser, le psychologue doit donc être capable d'apprécier la qualité de l'échantillonnage (donc d'avoir un avis critique sur la méthode utilisée) et les choix utilisés par les auteurs du test. Cela doit lui permettre de savoir si la mesure est adaptée à la personne à qui il fait passer cette épreuve.
- Les techniques d'échantillonnage sont des techniques non spécifiques à la psychométrie et ont été largement développées dans le cadre des recherches en sociologie, épidémiologie, marketing, etc. Elles sont nombreuses et varient selon la nature de l'information à retenir et la qualité attendue de l'échantillon. L'objectif de ce chapitre est uniquement de sensibiliser aux problèmes d'échantillonnage et de présenter différentes techniques classiques sans être exhaustif sur les méthodes utilisées.

1.1. Échantillon

L'échantillon est un groupe d'individus représentatif de la population ([population parente](#)) pour la mesure effectuée. Si l'échantillonnage est réalisé correctement, les résultats observés sur cet échantillon pour cette "mesure" sont supposés similaires [= représentatifs] à ceux que l'on observerait dans la population parente.

Remarque :

- l'échantillon permettant [d'étalonner](#) un test est parfois appelé **échantillon de standardisation ou échantillon normatif**. Il est préférable d'utiliser le terme échantillon de standardisation », car il désigne le groupe utilisé pour établir les normes d'un test. Le terme « échantillon normatif » est moins précis et moins utilisé.
- Un échantillon n'est pas représentatif en soi : sa représentativité dépend de la méthode d'échantillonnage mais aussi des caractéristiques et de la mesure effectuée. Par exemple si on sait (par des études validées antérieurement) que ce que l'on évalue ne dépend pas de l'activité professionnelle, prendre en compte l'activité professionnelle dans sa méthode d'échantillonnage est inutile. Par contre, le même échantillon, ne sera pas nécessairement utile pour une autre mesure sensible à ce facteur (activité professionnelle), cela dépendra en partie de la méthode utilisée.

1.2. Population parente

La population parente est constituée de l'ensemble des individus sur lesquelles porte l'objet de l'étude (population de référence pour un test). On appelle parfois (très rarement) cette population parente :

population mère.

Application à la construction d'un test : un test est construit pour différencier les individus d'une population donnée et doit permettre (test normatif) de situer un individu par rapport à cette population. Lors de la construction d'un test on va extraire un ou plusieurs [échantillons](#) représentatifs (échantillonnage) afin de [mettre au point](#) le test et étudier ses [qualités métrologiques](#) puis [l'étalonner](#).

1.3. Modèle de la population parente

Un **modèle de la population** parente est une description de cette population à partir de variables censées être en relation avec la ou les mesures à effectuer (le test en construction). Ce "modèle" de la population parente permet par exemple de construire les quotas dans [l'échantillonnage par la méthode des quotas](#) (§4.2) ou de définir des strates dans la [méthode probabiliste par stratification](#) (§4.1).

Exemple : dans la construction d'un test d'intelligence pour des adultes par la méthode des quotas, on construira un modèle de la population française en prenant en compte les catégories socioprofessionnelles, l'âge, le lieu d'habitation et le sexe. Les proportions de chacune de ces catégories simples ou croisées seront recherchées dans les statistiques nationales (le plus souvent celle de [l'Institut National de la statistique et des études économiques](#) - INSEE - pour la France ou [l'office fédéral de la statistique](#) pour la Suisse. Wikipédia donne la liste des sites officiels de statistiques pour différents pays : https://fr.wikipedia.org/wiki/Liste_des_instituts_officiels_de_statistique).

2. Méthodes d'échantillonnage



Il existe de nombreuses techniques d'échantillonnage. Classiquement on distingue deux groupes de méthodes : les [méthodes probabilistes](#) et les [méthodes non probabilistes](#) (dites aussi à choix raisonnés). Il ne sera présenté de façon plus détaillée que 5 méthodes dont 4 sont des méthodes probabilistes.

Les différences entre méthodes probabilistes et non probabilistes sont présentées dans le cadre des [méthodes non probabilistes](#) (cf. chap. D §2.2)

2.1. 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.

2.1.1 Probabiliste stricte

La méthode probabiliste stricte peut-être mis en œuvre de deux façons différentes : le tirage simple au hasard et le tirage systématique.

Tirage simple au hasard : pour sélectionner le groupe représentatif de la population parente (groupe de référence), on tire au sort chaque individu de l'échantillon (tirage au sort sans remise !). La procédure utilisée doit permettre d'assurer que toutes les personnes de la population de référence aient la même probabilité d'être sélectionnées. Cette technique est très difficile à utiliser correctement et est coûteuse lorsque la population parente est très importante puisqu'elle implique de posséder une liste exhaustive de tous les individus afin d'effectuer un tirage au sort réel.

Tirage systématique : Le principe de cette méthode implique de choisir au hasard un point de départ

(un seul tirage au sort), de calculer un taux de sondage et de parcourir la liste des personnes constituant la population parente. Elle est plus simple à mettre en œuvre que le tirage simple au hasard lorsque la liste est longue.

Mise en œuvre : On établit une liste sur laquelle tous les individus de la population sont présents. On tire au hasard une personne dans la liste (par exemple 1234^{ème}). On sélectionne ensuite les individus à intervalle régulier (à partir de la position de la personne tirée au hasard) sur cette liste en la parcourant vers le haut (fin de la liste) et le bas de la liste (début de la liste). L'intervalle (distance en nombre d'individus entre deux personnes sélectionnées) est égal au nombre d'individu de la population divisé par la taille de l'échantillon (l'inverse de ce qu'on appelle habituellement le taux de sondage).

Exemple : si l'on a 2000 individus et que l'on veut un échantillon représentatif de 100 personnes (on dit alors que le taux de sondage est de $100/2000$ soit $1/20^{\text{ème}}$). L'intervalle pour effectuer un tirage systématique est de $2000/100$ soit 20 (ce qui est bien l'inverse du taux de sondage). On sélectionne, à partir d'un individu tiré au sort (par exemple le 1234^{ème}) puis toutes les personnes ayant comme rang dans la liste $1234 + i*20$ et $1234 - j*20$ (i et j prenant initialement la valeur 1 et augmentant d'une unité jusqu'à épuisement de la liste).

2.1.2 Stratification

La méthode probabiliste stricte repose uniquement sur le hasard, hasard qui parfois "fait mal les choses" (biais d'échantillonnage par sur-représentativité d'un sous groupe). Pour contrôler la représentativité de l'échantillon, on peut utiliser la méthode de stratification.

Cette méthode nécessite d'avoir des informations sur chaque individu (par exemple : sexe, âge, profession, etc.) et la fréquence de ces caractères dans la population. On reproduit alors dans l'échantillon les caractéristiques de la population de référence, en tirant au hasard les individus non plus dans la population globale mais dans des strates (sous groupes) définies par les variables retenues pour caractériser la population.

Par exemple, si dans une population il y a 52% de femmes et 48% d'hommes et que l'on veut prendre en compte uniquement cette caractéristique, on échantillonne au hasard (méthode probabiliste stricte) parmi les femmes (première strate) puis parmi les hommes (seconde strate) un nombre d'individus de façon à ce que cette proportion soit respectée dans l'échantillon (52% seront des femmes et 48% des hommes). On peut stratifier un échantillon sur plusieurs caractères considérés conjointement (sexe et habitat et profession par exemple). Les unités (personnes ici) sont ensuite tirées au hasard à l'intérieur des strates ainsi définies.

Remarques :

- Avec cette méthode on a autant de tirage simple au hasard (bases de sondage) que de strates.
- Cette méthode présente un intérêt si le critère de stratification est en relation avec l'objet d'étude (il est par exemple totalement inutile de faire des strates en fonction de la couleur des yeux si l'on construit un test d'intelligence). Les variables prises en compte pour constituer les strates constituent, si l'on prend en compte l'objet d'étude pour constituer les strates, un modèle (partiel) de la population parente.
- Cette méthode est toujours une méthode probabiliste. Chaque individu de la population parente possède la même probabilité de faire partie de l'échantillon.
- Si les variables à la base des strates sont bien choisies, cette méthode permet de diminuer les

risques de biais d'échantillonnage (donc permet en principe, pour le même risque d'erreur, de diminuer la taille de l'échantillon). La qualité des strates détermine en partie la représentativité de l'échantillon.

Exemple pratique :

On souhaite réaliser une étude sur les projets professionnels d'étudiants inscrits dans une université Française (par exemple l'université Savoie Mont Blanc). Pour construire l'échantillon, les auteurs de l'étude observent que les étudiants sont répartis par groupe disciplinaires dans des Unités de Formation et de recherche (UFR) et les projets professionnels peuvent être très différents. Par ailleurs ils pensent qu'un autre facteur pourrait impacter les résultats : le niveau d'étude (premier ou second cycle). Ayant accès à la base de données de l'université, ils vont constituer des strates en fonction de ces deux critères et sélectionner au hasard des étudiants à l'intérieur de ces strates. Le nombre d'étudiants pris au hasard dans chacune des strates sera fixé en fonction de la taille de l'échantillon total souhaité mais en respectant les proportions observées dans chacune des strates pour cette université.



FigureG.1 : Illustration d'un échantillonnage par stratification.

2.1.3 Grappe

Le plus souvent le tirage au sort des participants à une étude parmi une liste exhaustive est impossible car on ne dispose pas de cette liste. L'échantillonnage en grappe (ou en groupe) permet de s'affranchir des difficultés de mise en œuvre de la technique d'échantillonnage probabiliste stricte ou par strate. Dans cette méthode l'unité de sondage n'est plus "la personne". On tire cette fois au hasard des groupes de personnes. Toutes les personnes de ce groupe (la grappe) sont interrogées. Cette méthode permet de prendre des unités de tirage au sort plus importantes (des villes, des écoles, des zones géographiques, etc.).

Remarques

- Contrairement aux autres techniques probabilistes, il n'est donc pas utile d'avoir la liste nominative de tous les membres de la population parente.
- Un des inconvénients de cette méthode est le risque d'homogénéité des grappes. Cette méthode nécessite donc une taille d'échantillon plus importante que l'échantillonnage probabiliste stricte.
- La taille de l'échantillon n'est pas fixée strictement au départ et dépend de la taille de chacune des grappes.

- Les échantillonnages par strate et en grappe peuvent être associés. Un bon exemple en est donné par Reuchlin dans son Précis de Statistique (1976, p. 190).
- Cette méthode est probabiliste mais ne nécessite pas d'avoir une liste exhaustive de la population parente (c'est la seule parmi les méthodes probabilistes).
- Pour cette méthode, la probabilité d'un individu d'être sélectionné dans l'échantillon dépend directement du nombre de grappe (nombre des groupes) et non pas de la taille de l'échantillon.

Cas particulier : les sondages à plusieurs degrés.

Dans le système d'échantillonnage par grappe, on peut effectuer une succession de tirage par grappes de plus en plus petites, incluses dans celles choisies au niveau précédent pour en arriver parfois à tirer au hasard les sujets eux-mêmes dans ces sous-groupes. Par exemple : pour effectuer une étude sur les étudiants français, on peut sélectionner au hasard 30 universités (premier niveau d'échantillonnage en grappe), puis dans chacune de ces universités, toujours au hasard six filières d'enseignement et enfin dans chacune de ces filières, deux groupes de TD.

2.1.4 Les non-réponses

Contrairement à la méthode des quotas et plus généralement à la plupart des méthodes non probabilistes (cf. chap. D §2.2), les méthodes probabilistes engendrent un taux de non-réponses (NR) pouvant être important. Ce taux de non-réponse peut introduire un biais d'échantillonnage si les caractéristiques des répondants et des non-répondants diffèrent et que cette différence a un impact sur la mesure étudiée. **Le taux des non-réponses devrait donc toujours être indiqué** dans un échantillonnage probabiliste.

Remarque : remplacer une non-réponse par la réponse d'une autre personne (même pris au hasard dans la population parente) ne permet pas de supprimer le biais éventuel introduit par les NR ! C'est une erreur classique des "débutants". En effet, les non-réponses peuvent être associées à une caractéristique de la population par ailleurs en rapport avec la mesure que l'on effectue. On se doit donc de donner ce nombre de non-réponse mais aussi quand c'est possible de décrire "la population des non-répondants".

2.2. Échantillonnage non probabiliste

Les méthodes non probabilistes ou les méthodes dites à choix raisonnées sont des méthodes de sélection où la représentativité de l'échantillon est assurée par une démarche raisonnée en utilisant des règles de sélection des individus fixées préalablement. Il existe plusieurs méthodes non probabilistes (à choix raisonné). La méthode la plus utilisée en psychologie est la [méthode des quotas](#) (seule méthode qui est réellement présentée dans ce manuel).

Quelles sont les principales différences entre méthodes probabilistes et méthodes non probabilistes ?

→ Il n'y a plus de hasard au sens strict dans les méthodes non probabilistes.

→ Pour les méthodes non probabilistes, la probabilité qu'a un individu de la population d'appartenir à l'échantillon est inconnue : il est donc impossible d'évaluer la variance d'échantillonnage et donc de mesurer la précision des estimations (ou autrement dit le degré de confiance dans les résultats observés).

→ Il n'existe pas, dans les méthodes non probabilistes, des [non-réponses](#). Lorsqu'un individu ne répond pas il est remplacé par un autre.

2.2.1 Méthode des quotas

La méthode des quotas est utilisée quand il n'existe pas de base de sondage. L'objectif est d'assurer la représentativité de l'échantillon en conformant la structure de l'échantillon aux caractéristiques de la population. Cela suppose des statistiques fiables concernant la population parente. Cette méthode est, d'une certaine façon, proche de la méthode stratifiée mais il n'y a plus de hasard. On choisit les personnes « au gré des rencontres » (et on est libre de choisir telle ou telle personne) mais on s'impose de respecter les proportions de diverses catégories de la population parente. On cherche les sujets représentatifs de la population totale. Par exemple, si dans la population de référence il y a 10% d'agriculteurs et 51% de femmes et 49% d'hommes parmi eux, pour un échantillon de 100 personnes, on devra sélectionner 5 agriculteurs et 5 agricultrices.

Mise en œuvre de la méthode des quotas

1. Construire un modèle de la population parente : on décrit la population à partir de variables supposées être en relation avec l'objet de mesure (par exemple, pour élaborer un test de personnalité, on peut prendre comme variables, le niveau d'étude, la catégorie socioprofessionnelle, l'âge, le sexe, etc.).
2. On décide quelles sont les variables traitées comme des variables simples et celles que l'on va croiser (*i.e.* prendre en compte simultanément plusieurs critères. Par exemple si on prend la variable sexe et la variable vie en couple : les hommes vivant en couple, les hommes célibataires, les femmes vivant en couple et les femmes célibataires).
3. On recherche des statistiques concernant ces variables simples ou croisées (fréquence dans la population parente de l'échantillon que l'on veut constituer).
4. On fixe le nombre des personnes à interroger et on construit les quotas (cf. un exemple de feuille de quota ci-dessous) de façon à ce que dans l'échantillon les proportions observées dans la population de référence soient respectées.
5. On distribue ces feuilles de quotas à des enquêteurs. Ceux-ci, pour chaque personne qu'ils veulent inclure dans l'échantillon, doivent vérifier que leur quota n'est pas dépassé et doivent, lorsqu'ils interrogent une personne, remplir la feuille de quota. Dans l'exemple présenté ci-dessous, les enquêteurs devront pour chaque personne intégrée à l'échantillon, se renseigner sur le nombre d'enfants, l'âge, la CSP et si la personne vit en couple ou non. On décomptera cette personne de la feuille de quota en cochant les modalités qui correspondent. La difficulté pour les enquêteurs est qu'au fur et à mesure de l'enquête, la personne à trouver risque d'avoir des caractéristiques très spécifiques (par exemple, selon les personnes rencontrées, dans l'exemple de feuille de quota donnée ci-dessous, la dernière personne à interroger pourrait être une personne ayant les caractéristiques suivantes : avoir 1 enfant de moins de 18 ans, avoir de plus de 70 ans, être étudiant, être un homme célibataire (compliqué à trouver !).

Exemple de feuille de quotas

Dans cet exemple il y a 5 variables pris en compte pour établir les quotas (Nombre d'enfants, Age, CSP, Sexe, Vie en couple). Ces 5 variables ont donné lieu à 4 quotas : 3 quotas simples (Nombre d'enfants, Age, CSP) et 1 quota croisé (quota qui combine les variables, Sexe et Vie en couple).

Quotas	Nombre d'interviews
Nombre d'enfants (-18ans)	
0	8 : <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
1	9 : <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
2	10 : <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
3 et +	5 : <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
Âges	
18- 30 ans	6 : <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
31-55 ans	12 : <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
56-70 ans	10 : <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
+ 70ans	4 : <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
CSP	
Agriculteur	3 : <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
Artisan/petit commerçant	4 : <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
Prof.lib / Cadre supérieur	3 : <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
Prof. intermed. / employé	8 : <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
Ouvriers	5 : <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
Etudiants	2 : <input type="checkbox"/> <input type="checkbox"/>
Retraité, autre inactif	7 : <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
Caractéristiques	
Femmes vivant en couple	11 : <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
Femmes célibataires	6 : <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
Hommes vivant en couple	12 : <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
Hommes célibataires	3 : <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>

Figure G.1 : Illustration d'une feuille de quota (exemple fictif)

Les contraintes et les avantages de la méthode des quotas :

- On doit connaître les caractéristiques de la population parente (mais on n'a pas besoin d'une liste exhaustive des individus constituant cette population). Le plus souvent ces caractéristiques peuvent être données par les organismes nationaux de statistiques (la liste de ces instituts peut être trouvé sur wikipédia : https://fr.wikipedia.org/wiki/Liste_des_instituts_officiels_de_statistique).
- La difficulté pour trouver des individus participant à l'échantillon augmente au fur et à mesure que l'on avance dans la construction de l'échantillon. Les dernières personnes à interroger sont parfois très difficiles à trouver si l'on veut respecter les caractéristiques de la structure de la population parente. La dernière personne est déterminée de manière unique par les modalités restantes des quotas. Tout le métier de l'enquêteur consiste à ne pas se faire piéger et réaliser correctement ses "fins de quotas".
- La probabilité qu'a un individu de la population d'appartenir à l'échantillon est inconnue : il est donc impossible d'évaluer la variance d'échantillonnage et donc de mesurer la précision des estimations.
- Les quotas doivent être pertinents (modèle de la population parente) notamment en relation réelle ou supposée avec la mesure ou les mesures à effectuer. La qualité du [modèle de la population parente](#) utilisé est essentielle. Un même échantillon ne convient pas pour toutes les mesures. Dans la méthode des quotas l'échantillon est dépendant des mesures (encore plus que dans la méthode [probabiliste par strate](#)).
- Cette méthode présente l'avantage d'être souvent plus rapide et moins coûteuse que les méthodes probabilistes.
- Dans cette méthode, il n'y a pas de non-réponse. Lorsqu'une personne ne veut pas participer, on en cherche une autre !

Remarque : Les psychologues utilisent souvent la méthode des quotas lors de la construction de tests

comme lors de la construction les échelles d'intelligence de Wechsler.

2.2.2 Autres méthodes

En dehors de la méthode des quotas, il existe d'autres méthodes non probabilistes mais ces méthodes peuvent induire des biais plus ou moins importants dans la représentativité de l'échantillon.

Exemples d'échantillonnage non probabilistes :

- Échantillonnage sur la base du volontariat (pas de garantie de représentativité). Cette méthode est utilisée non pas dans la construction des tests mais dans les études lorsque l'on recherche des groupes témoins.
- La méthode des itinéraires : on fixe, à celui qui interview, un itinéraire (du point de vue de l'espace et du temps) à parcourir pour interroger des personnes de façon à avoir la population la plus représentative possible (méthode utilisée en sociologie dans certaines enquêtes).
- Technique boule de neige ("snowball") : on utilise le parrainage ou les amis et collègues des répondants pour construire l'échantillon (permet en fait de sonder ou d'interroger une population spécifique).
- L'échantillonnage dirigé ("purposive or judgmental sample") : on détermine l'échantillon en fonction de l'objet d'étude (on sélectionne des personnes que l'on pense appropriées en fonction d'une expertise dans un domaine objet de l'étude). En psychologie cette méthode est utilisée lorsque l'on sélectionne des groupes extrêmes ou un groupe clinique (ce qui est une définition partiellement différente de celle que l'on trouve dans d'autres domaines).

3. Taille des échantillons

Définir la taille d'un échantillon est complexe et différents facteurs affectent la détermination du nombre de personnes devant appartenir à l'échantillon (taille d'échantillon). Il existe différentes formules et techniques mais toutes ces méthodes montrent que l'on doit prendre en compte :

- Le degré de certitude ou le niveau de confiance que l'on veut avoir (intervalle de confiance) dans les résultats. Plus on souhaite avoir une marge d'erreur faible plus l'échantillon devra être important (si la méthode d'échantillonnage est correcte).
- la prévalence estimée de la variable étudiée (exemple pour une question en oui-non, la fréquence attendue des oui et non) ou pour d'autres types de mesure la dispersion des valeurs autour de l'indice de tendance centrale.
- la méthode d'échantillonnage (plan d'échantillonnage) détermine aussi (toute chose étant égale par ailleurs) la taille d'un échantillon. Par exemple la méthode par stratification assure a priori une meilleure représentativité que d'autres méthodes d'échantillonnage. Dans ce cas, la taille de l'échantillon peut être moins importante qu'avec une autre méthode (pour le même degré de précision).
- la taille de la population parente. Ce facteur qui paraît évident et qui doit être pris en compte lorsque la population de référence a des faibles effectifs, devient de moins en moins important lorsque l'effectif de la population parente devient très important. Son importance est donc relative (cf. pour aller plus loin, ci-dessous).

Remarques

- Si la méthode d'échantillonnage est incorrecte, augmenter de façon importante la taille de

l'échantillon n'apporte aucune garantie sur la validité ou la représentativité de l'échantillon.

Les exemples les plus fameux dans ce domaine concernent les premiers sondages et enquêtes pré-électorales effectués aux USA. Il est rapporté plus particulièrement celui concernant l'élection présidentielle opposant Roosevelt à Landon. Un journal a effectué un sondage auprès de 3 millions de personnes et donnait Landon gagnant, or, Gallup, avec un sondage auprès de 4500 personnes, donnait avec raison Roosevelt gagnant. Le biais du premier sondage était simple à repérer : l'enquête avait été faite par téléphone auprès des abonnés du journal, et les personnes interrogées n'étaient pas représentatives de la population américaine même si elles étaient 665 fois plus nombreuses. Lorsque qu'un auteur met en avant la "grande taille" ou l'importance de son échantillon, il faut s'inquiéter et regarder la méthode d'échantillonnage.

- La qualité de la mesure issue d'un échantillon n'est pas directement proportionnelle à la taille de l'échantillon. Il ne suffit pas de doubler la taille d'un échantillon pour doubler la qualité de la mesure. Très schématiquement, pour multiplier par 2 la qualité de la mesure, il faut par exemple multiplier par 4 la taille de l'échantillon.
- En théorie, les méthodes de calcul de la taille d'échantillon ne s'appliquent que sur les échantillons obtenus par des méthodes probabilistes. En pratique, ces méthodes de calcul sont quand même utilisées pour les méthodes non probabilistes et corrigées (éventuellement). Selon la méthode d'échantillonnage choisie les instituts de sondage multiplient les résultats par un coefficient prenant en compte les caractéristiques du plan d'échantillonnage. Par exemple, pour une méthode par grappe, ils peuvent doubler la taille d'échantillonnage nécessaire pour une méthode probabiliste stricte.
- La taille d'un échantillon d'un échantillon ne dépend que partiellement de la taille de la population parente (population dans laquelle on échantillonne).

Pour aller plus loin

Illustration de l'effet des différents facteurs dans la détermination de la taille d'un échantillon (par simulation)

Cette simulation est présentée uniquement pour illustrer des effets classiques affectant la représentativité d'un échantillon. Nous nous plaçons dans un cadre simple, celui d'une enquête qui cherche à savoir si dans une population on préfère le produit A ou le produit B. Pour calculer la taille de l'échantillon (population parente finie), la formule utilisée est :

$$\pm z_{\alpha} \sqrt{1 - (n/(N - n))} \sqrt{p(1 - p)/n}$$

avec : N = taille de la population parente

n = taille de l'échantillon

p = proportion attendue de choix A dans la population

$1 - \alpha$ = degré de confiance (probabilité)

i = fourchette (intervalle de confiance pour p , +/- $i\%$)

z_{α} = valeur z lu dans la table de la loi normale

Cette formule, va nous permettre d'illustrer les effets des différentes variables (taille de l'échantillon, de la population parente, etc.) sur le degré de confiance dans les résultats. Vous pouvez aussi avec un tableur faire d'autres simulations. En fait, le principe est de faire varier, un paramètre et on regarde l'effet sur un autre paramètre (en laissant les autres paramètres constants).

Simulation 1 : pour une taille de population donnée (N), plus l'échantillon est grand (n), meilleure est la précision (fourchette i diminue)

N	n	p	α	i
100 000	10	50%	5%	30,99%
100 000	100	50%	5%	9,79%
100 000	1 000	50%	5%	3,08%
100 000	2 000	50%	5%	2,17%
100 000	4 000	50%	5%	1,52%
100 000	8 000	50%	5%	1,05%
100 000	16 000	50%	5%	0,71%
100 000	90 000	50%	5%	0,10%

Simulation 2 : la taille de la population parente (N) a une importance toute relative dans le degré de confiance (i.e précision, fourchette i) à taille d'échantillon suffisante ($n=1000$). En fait, avec 1000 individus, la précision du résultat est similaire pour une population parente de 200 000 individus et pour une population parente de 100 000 000 individus !

N	n	p	α	i
10 000	1 000	50%	5%	2,94%
50 000	1 000	50%	5%	3,07%
100 000	1 000	50%	5%	3,08%
200 000	1 000	50%	5%	3,09%
500 000	1 000	50%	5%	3,10%
1 000 000	1 000	50%	5%	3,10%
10 000 000	1 000	50%	5%	3,10%
100 000 000	1 000	50%	5%	3,10%

Simulation 3 : pour un degré de précision fixé (ici $i = 3\%$), la taille de l'échantillon (n) pour une population de référence constante ($N=100\ 000$) dépend de la proportion de choix A (p) dans la population parente. En fait, la taille de l'échantillon est maximum pour $p=50\%$ (donc quand la dispersion est maximum). Comme avant une étude on ne connaît pas p , on pose toujours l'hypothèse $p=50\%$ pour calculer la taille de l'échantillon.

N	p	α	i	n
100 000	1%	5%	3,00%	42
100 000	5%	5%	3,00%	203
100 000	10%	5%	3,00%	384
100 000	15%	5%	3,00%	543
100 000	20%	5%	3,00%	682
100 000	30%	5%	3,00%	894
100 000	40%	5%	3,00%	1022
100 000	45%	5%	3,00%	1054
100 000	50%	5%	3,00%	1064

H - Complément : Statistiques

The statistical method is not psychology, but it is indispensable to the quantitative study of psychological phenomena.

THURSTONE (1935)

Les statistiques peuvent être descriptives ou inférentielles. Les statistiques descriptives permettent de caractériser les distributions de scores, d'évaluer la symétrie, la dispersion et la normalité des données, ainsi que de résumer leur organisation ou leur structure relationnelle. Les statistiques inférentielles, quant à elles, visent à estimer et à tester la significativité des relations observées, qu'il s'agisse de corrélations, de différences entre groupes ou encore de structures factorielles (comme dans les modèles d'équations structurelles).

1. Prérequis Statistiques



Cette partie présente des notions nécessaires (prérequis) pour comprendre ou interpréter des données en psychométrie. Si vos connaissances sont suffisantes, passer votre chemin.

1.1. Les échelles de mesures

De façon générale mesurer c'est **attribuer des nombres aux objets, selon des règles déterminées**. Ces règles vont toujours avoir pour objet d'établir une correspondance entre certaines propriétés des nombres et certaines propriétés des objets. Stevens en 1946 propose de classer les échelles de mesure en fonction des propriétés des nombres qui sont conservées.

	<p><u>Les échelles nominales</u></p> <p>réaliser une partition des observations</p> <p>-----</p>
	<p><u>Les échelles ordinales</u></p> <p>réaliser une partition des observations + définir une relation d'ordre</p> <p>-----</p>
	<p><u>Les échelles d'intervalles</u></p> <p>réaliser une partition des observations + définir une relation d'ordre + distance (point zéro arbitraire)</p> <p>-----</p>
	<p><u>Les échelles de rapport</u></p> <p>réaliser une partition des observations + définir une relation d'ordre + distance (point zéro non arbitraire)</p>

Cette classification est très critiquée par les statisticiens ([Velleman et Wilkinson, 1993](#)) mais il est traditionnel en psychologie (lors de la formation) de distinguer ces quatre grands types de mesures.

1.1.1 Échelle nominale

Une échelle nominale répartie les observations dans un certain nombre de classes disjointes, telles que chaque observation entre dans une seule classe. L'ensemble des classes utilisées constitue l'échelle nominale.

En d'autres termes, on effectue une partition de l'ensemble des observations (application d'une relation d'équivalence) et tous les objets ou les sujets d'une même classe sont considérés comme équivalents. Attention : ce n'est pas un critère statistique qui définit la partition, c'est un critère empirique. Ce critère détermine la signification à attribuer à la mesure.

Exemple d'échelle nominale : lors d'une étude sur l'entretien clinique, toutes les interventions d'un psychologue sont classées en trois catégories : *Interprétations (I)*, *Clarifications (C)* et *Reformulations (R)*. Cette catégorisation constitue une échelle nominale si et seulement si on effectue une partition des interventions c'est-à-dire, si et seulement si, chaque intervention entre dans une seule catégorie I, C, ou R. Dans le cas où il serait impossible de faire entrer les interventions dans une seule catégorie, on devra ajouter d'autres classes à l'échelle ou redéfinir la règle de partition utilisée.

1.1.2 Échelle ordinale

Les échelles ordinales possèdent les propriétés des échelles nominales (effectuer une partition des observations), mais les objets d'une catégorie ne sont pas seulement différents de ceux d'une autre catégorie, il existe entre les catégories de l'échelle une relation d'ordre stricte ou non (*). Pour construire une échelle ordinale, il faut donc :

- Effectuer une partition de l'ensemble des observations (relation d'équivalence).
- Définir une relation d'ordre stricte ou non.

Exemple d'échelle ordinale : Les échelles d'appréciation par lesquelles on demande aux sujets d'exprimer des jugements sur un « objet » (comme Très bon, Bon, Moyen, Mauvais, Très mauvais) sont des échelles ordinales.

 (*) Rappel concernant la relation d'ordre : soit E un ensemble et une relation binaire sur cet ensemble notée « R », cette relation est une relation d'ordre si elle est :

→ Antisymétrique : $\forall x, y \in E \quad (x R y) \text{ et } (y R x) \Rightarrow x = y$

→ Transitive : $\forall x, y, z \in E \quad (x R y) \text{ et } (y R z) \Rightarrow (x R z)$

→ Réflexive : $\forall x \in E \quad x R x$

Lorsque la relation de réflexivité n'est pas respectée et que la relation est antiréflexive, on parle de relation d'ordre strict .

1.1.3 Échelle d'intervalle

Dans l'échelle d'intervalle, la mesure implique, en plus des propriétés des échelles ordinales (partition des observations et relation d'ordre stricte ou non), la notion de distance. L'unité de distance donne la signification à la mesure (par exemple : le temps en millisecondes). Cette unité de distance est stable tout au long de l'échelle, ce qui signifie que l'on peut comparer la différence observée entre deux mesures à la différence observée sur deux autres mesures. Les opérations arithmétiques peuvent s'appliquer sur les nombres représentant les classes. Dans les échelles d'intervalles le point zéro est arbitraire.

Remarque : le problème des psychologues est de définir ce que l'on entend lorsque l'on parle de la distance entre deux mesures et d'unité de mesure. Il est en fait très difficile de faire la preuve expérimentale que l'on a réellement des échelles d'intervalles mais les avantages de ces échelles sont apparus comme suffisamment importants pour que l'on traite un certains nombre d'échelles ordinales comme des échelles d'intervalles sous certaines conditions ou transformations (ces conditions ou transformations ne seront pas détaillées ici).

Exemple d'échelle d'intervalle : un exemple typique est la température mesurée en degrés Celsius. Nous pouvons dire qu'une température de 60 degrés est plus élevée qu'une température de 50 degrés, et qu'une augmentation de 30 à 60 degrés est deux fois plus importante qu'une augmentation de 30 à 45 degrés. Le point zéro est par contre arbitraire et **on ne peut pas dire que 60° Celsius est deux fois plus chaud que 30° Celsius**.

1.1.4 Échelle de rapport

Une échelle de rapport (ou de ratio) est une échelle d'intervalle dans laquelle le point zéro n'est pas arbitraire (comme le temps de réponse ou une mesure de vitesse). Ce type d'échelle est rarement utilisé ou plus exactement les propriétés de ce type d'échelle sont rarement utilisées en psychologie.

Les échelles de rapport représentent des rapports car elles ont une origine absolue (correspondant à l'absence de l'attribut mesuré). Par exemple, la distance a pour origine 0 (absence de distance) et 40 mètres est deux fois plus loin que 20 mètres. Ce n'est pas le cas d'une échelle d'intervalle comme la température exprimée en Celsius ou le 0° est arbitraire. Une température de 40° n'est pas deux fois plus chaude que 20°. Pour connaître le rapport entre ces deux températures, il aurait fallu prendre une mesure absolue de la température en Kelvin (qui est une échelle de rapport) et on peut alors comparer les deux mesures en Kelvin et en faire le rapport (Rappel, la règle de conversion Celsius (t_c) en Kelvin (t_k) : $t_k = t_c + 273.15$)

1.2. Statistiques descriptives



La description d'un ensemble de données repose sur l'utilisation de statistiques descriptives, lesquelles constituent un ensemble de méthodes destinées à résumer, organiser et présenter les informations contenues dans les données observées, sans recourir à une énumération exhaustive de celles-ci. Ces statistiques se déclinent généralement sous deux formes complémentaires :

des représentations graphiques (telles que les histogrammes, diagrammes en bâtons ou courbes), qui permettent une visualisation intuitive de la distribution des valeurs ;

et indices statistiques qui condensent les caractéristiques essentielles de la distribution.

Parmi ces indices, les plus couramment utilisés sont les mesures de tendance centrale (moyenne, médiane, mode), les mesures de dispersion (étendue, variance, écart-type) ainsi que les indicateurs de corrélation, permettant d'apprécier les relations entre variables. Pour les échelles d'intervalle ou de rapport, il est en outre possible de calculer des paramètres de forme de la distribution, tels que le

coefficient d'asymétrie (skewness) et le coefficient d'aplatissement (kurtosis), qui fournissent des informations complémentaires sur la structure des données.

1.2.1 Tendance centrale

L'indice de tendance centrale est un indice résumant l'ensemble des données. Il correspond à la valeur typique de la distribution des valeurs : celle qui "représente" toutes les valeurs (autour de laquelle les données ont tendance à se rassembler). Les indices de tendance centrale que l'on peut utiliser varient en fonction de la [nature des échelles](#). Selon l'échelle, cet indice peut être la valeur la plus fréquente, la valeur dépassée dans 50% des cas, la moyenne arithmétique des valeurs rencontrées, etc. Il existe donc plusieurs indices de tendance centrale et celui que l'on utilisera dépend à la fois de l'échelle mais aussi de ce que l'on veut observer (par exemple : calculer la moyenne arithmétique des salaires d'un pays ou calculer le salaire médian n'apporte pas la même information).

A savoir :

- [Échelle nominale](#) : on utilise le mode (valeur observée la plus fréquente)
- [Échelle ordinale](#) : on peut utiliser le mode (valeur observée la plus fréquente) mais on préfère la médiane (valeur dépassée par 50% des sujets, c'est dire valeur pour laquelle la fréquence cumulée est de 0.50).
- [Échelle d'intervalle](#) : on peut utiliser le mode (valeur observée la plus fréquente) ou la médiane (valeur dépassée par 50% des sujets) mais on préfère souvent la moyenne arithmétique (somme des scores observées divisée par le nombre de scores).

Remarque : lorsque les distributions sont symétriques et uni-modales on a nécessairement le mode, la médiane et la moyenne arithmétique qui sont identiques.

Pour aller plus loin...

La moyenne réfère le plus souvent à la moyenne arithmétique. Il existe cependant d'autres façons de calculer la moyenne. En effet, la moyenne est la valeur que devrait avoir toutes les observations pour que le total reste inchangé. Selon la nature de ces observations, on peut (on doit) utiliser d'autres moyennes (par exemple : moyenne géométrique, harmonique, quadratique, etc.). Vous trouverez facilement des exemples et les formules de calcul de ces moyennes sur internet.

1.2.2 Dispersion

L'indice de dispersion permet de savoir si les valeurs observées sont proches ou relativement éloignées de l'indice de tendance centrale. Cet indice est essentiel puisque, par exemple, savoir que la moyenne des notes observées à un examen est 12 sur 20 est insuffisant. En effet, l'ensemble des notes peut être proche de 12 (compris entre 11,5 et 12,5) ou éloignée de 12 (compris par exemple entre 3 et 19). La meilleure prédiction que l'on peut faire pour une personne dont on ne connaît pas la note sera, pour cet examen la note de 12, mais l'erreur faite (l'écart à la note réelle) sur cette prédiction sera d'autant plus grande que la dispersion des scores est grande. Une forte ou faible variabilité des notes (forte ou faible dispersion) autour de l'indice de tendance centrale est donc une information utile et complémentaire à l'information apportée par l'indice de tendance centrale.

Les indices de dispersion sont multiples et sont associés à l'indicateur de tendance centrale utilisé. Par exemple :

- **associé au mode** : l'indice d'entropie (H , *non présenté ici*). Il exprime le degré de désordre de la distribution des fréquences observées. Plus l'entropie est élevée, plus la distribution est hétérogène.
- **associé à la médiane** : l'écart inter-quartile ou le demi-inter-quartile (différence ou demi-différence entre le premier et le troisième quartile). Pour information on donne aussi parfois l'étendue de la distribution c'est à dire les deux extrêmes. [rappel le $i^{\text{ème}}$ quartile est le score dont la fréquence cumulée est $i*25\%$]. Il est peu sensible aux valeurs extrêmes et constitue un indicateur robuste de variabilité.
- **associé à la moyenne** : la variance (σ^2) ou l'écart-type ($\sigma = \text{racine carrée de la variance}$). La variance est la (moyenne des carrés des écarts à la moyenne ($\sigma^2 = [\sum(x_i - m)]/n$). Si les notes sont toutes identiques la variance est égale à 0.

La variance et l'écart-type sont dépendants de la mesure et de l'unité de mesure. On peut aussi, pour estimer la dispersion indépendamment de l'unité de mesure, calculer ce qu'on appelle le coefficient de variation (CV). Le CV est le rapport de l'écart-type à la moyenne. Il permet la comparaison de distributions de valeurs dont les échelles de mesure ne sont pas comparables. C'est un indice de dispersion relatif contrairement à la variance et l'écart-type qui sont des dispersions absolues.

Remarques :

- La compréhension de la signification de la notion de dispersion est utile quand on met en relation plusieurs variables. En effet, on admet que l'amplitude des différences interindividuelles (mesurées par la dispersion) sont **toujours** dues en partie à l'erreur de mesure (erreur aléatoire = ensemble de facteurs indépendants affectant de façon non prévisible la mesure) mais peuvent être aussi dues à un ou plusieurs facteurs (variables latentes) sous-tendant les comportements et à l'origine de ces différences. Ces facteurs qui sont sources de variations (à la base de la dispersion observée) peuvent être communs à plusieurs épreuves. Ces sources de variations sont donc à la base des covariations entre les scores observés (cf. [l'analyse corrélacionnelle, chap. A §2.4](#)).
- L'indice de dispersion contribue aussi à l'interprétation d'un score observé. Par exemple, si la note obtenue par un enfant est de 10 sur 20 et que la moyenne de la classe est 9, on peut penser que c'est bien. Si l'écart-type observé des notes dans la classe est de 0.30, en fait ce score de 10 est à plus de 3 écarts-types de la moyenne* et donc ce score était très peu probable car les scores devaient être tous proches de 9 (entre 8.4 et 9.6)**. Le score de cet enfant est donc le meilleur ou probablement un des meilleurs de la classe.
A l'inverse si l'écart-type est de 3, la note de 10 était une valeur probable*** (à un tiers d'écart-type de la moyenne). La note de 10 est alors une note dans "la moyenne" de la classe. Cet exemple montre que pour interpréter un score, l'écart à la moyenne n'est pas suffisant, et on doit le mettre en relation avec un indice de dispersion comme l'écart-type (sous l'hypothèse d'une [distribution normale](#) ou quasi-normale). On pourrait prendre aussi un autre indice comme l'écart inter-quartile.

(*) Si une distribution est normale, presque toutes les valeurs observées (99,9%) se situent en général entre -3 et +3 écarts-types.

(**) 94,4% des scores se situent entre -2 et +2 écarts-types de la moyenne (sous l'hypothèse d'une distribution normale ou quasi-normale)

(***) Entre le score minimal et 1 écart-type (donc ici entre 0 et 10), si la distribution est normale, on trouve 84,1% des notes (et 68,3% entre 8 et 10). On peut donc alors dire que cette note était probable,

ou par abus de langage, dans "la moyenne" des notes observées.

1.2.3 Paramètres de formes

Pour caractériser la forme d'une distribution (échelle d'intervalle ou de rapport), c'est-à-dire pour préciser l'allure de la courbe des fréquences, il existe des coefficients permettant d'évaluer l'asymétrie d'une distribution et son aplatissement.

(a) Asymétrie

Une distribution statistique est symétrique si les observations repérées par leurs fréquences sont également dispersées de part et d'autre d'une valeur centrale. Le coefficient d'asymétrie correspond au moment d'ordre 3* de la variable centrée réduite. En pratique, on utilise un estimateur normalisé et non biaisé égal à :

$$G_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^3$$

Avec :

- n le nombre d'observations
- x_i le score observé pour l'observation i
- \bar{x} et s la moyenne et écart-type non biaisés (**)

La valeur de ce coefficient d'asymétrie (skewness) est de 0 pour une distribution normale. Un coefficient négatif traduit une asymétrie avec une queue de distribution plus étendue à gauche. Un coefficient positif traduit une asymétrie avec une queue de distribution plus étendue à droite.

En général, pour les scores observés dans une épreuve cognitive, un coefficient d'asymétrie positif est en relation avec un effet plancher (tâche difficile) et un coefficient d'asymétrie négatif est en relation avec un effet plafond (tâche trop facile).

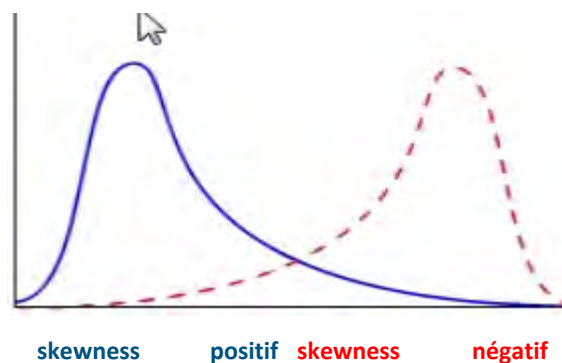


Figure H.1 : Exemples de courbes asymétriques

(*) Pour ceux qui veulent savoir : Un moment d'ordre r est une moyenne des écarts par rapport à un réel "a" élevés à une puissance "r", r étant un entier naturel. La moyenne et la variance sont des moments d'ordre 1 et 2. Le skewness et le kurtosis des moments d'ordre 3 et 4 (facile à comprendre si on inspecte les formules de calcul de ces indices).

(**) Un estimateur non biaisé pour la moyenne et l'écart-type s'obtient en remplaçant n par $n-1$ dans les formules de la moyenne et de l'écart-type

(b) Aplatissement

La mesure d'aplatissement (= degré de voûture) ou Kurtosis (du grec kurtos signifiant courbe ou

arrondi) est une statistique descriptive (moment centré d'ordre 4*) mesurant l'aplatissement de la distribution ou ce qu'on appelle encore son degré de voussure ou parfois sa "kurtose".

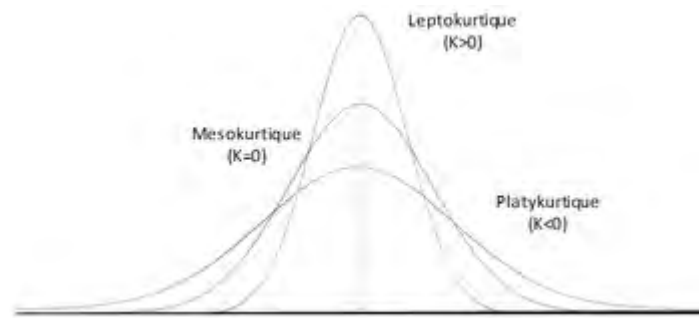


Figure H.2 : Exemples de distributions ayant 3 degrés différents de voussure (kurtose)

Pour une distribution normale, la valeur de ce coefficient (moment centré d'ordre 4) est de 3 pour une distribution normale. En pratique, on utilise le plus souvent un coefficient corrigé G_2 (kurtosis normalisé**). La valeur de ce coefficient est alors de 0 pour une distribution normale (courbe dite alors mésokurtique). Un coefficient d'aplatissement négatif indique une distribution plutôt aplatie (platykurtique) et un coefficient d'aplatissement positif, une distribution "pointue" (leptokurtique). La formule de calcul de ce coefficient d'aplatissement normalisé et non biaisé est :

$$G_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Avec :

- n le nombre d'observations
- x_i le score observé pour l'observation i
- \bar{x} et s la moyenne et écart-type non biaisés (***)

(*) Pour ceux qui veulent savoir : Un moment d'ordre r est une moyenne des écarts par rapport à un réel "a" élevés à une puissance "r", r étant un entier naturel. La moyenne et la variance sont des moments d'ordre 1 et 2. Le skewness et le kurtosis des moments d'ordre 3 et 4 (facile à comprendre si on inspecte les formules de calcul de ces indices).

(**) Le terme excès d'aplatissement dérivé de "kurtosis excess" en anglais est utilisé parfois à la place de kurtosis normalisé mais il est ambigu car un excès d'aplatissement positif est une courbe leptokurtique (distribution pointue) et un excès d'aplatissement négatif à une courbe platykurtique (distribution aplatie).

(***) un estimateur non biaisé pour la moyenne et l'écart-type s'obtient en remplaçant n par $n-1$ dans les formules de la moyenne et de l'écart-type

1.2.4 Corrélation

Une statistique descriptive particulière : la corrélation. Lorsque l'on possède pour chaque sujet d'une population deux mesures (variables dépendantes ou VD), on peut et on doit s'intéresser aux relations entre ces deux variables. Pour les échelles d'intervalles, la question que l'on se pose le plus fréquemment est de savoir si la variance observée sur une VD (c'est à dire l'amplitude des différences interindividuelles) est spécifique à chacun des tests ou s'il existe une part de variance commune à ces deux tests. Cette évaluation de la part commune à un ou plusieurs tests, à la base de l'analyse dimensionnelle et à la base de [l'analyse factorielle](#), est réalisée à l'aide du coefficient de corrélation de Bravais Pearson.

Le coefficient de corrélation est donc une mesure qui évalue la conformité des observations avec un modèle général de relations entre les deux mesures. Ce modèle général est le plus souvent [un modèle linéaire](#) et le coefficient de corrélation associé est le r de Bravais-Pearson pour les échelles d'intervalles ou de rapports. Pour les autres types d'échelles, il n'y a pas de modèle (comme le modèle linéaire) sous-jacent à la mesure des relations entre les deux variables. Pour les [échelles nominales](#) on utilise un indice dérivé du Chi carré* et pour les [échelles ordinales](#), un indice de corrélation identique à celui de [Bravais Pearson](#) et calculé sur les rangs et nommé le coefficient r de Spearman.

Résumé des principaux coefficients évaluant la relation entre deux variables

Coefficient de corrélation	Variable A (échelle)	Variable B (échelle)	Remarques
Bravais Pearson	Intervalle	Intervalle	Coefficient de référence. Relation linéaire.
Spearman	Ordinale	Ordinale	Coefficient équivalent à celui de Bravais Pearson d'un point de vue algébrique mais sur les rangs.
Polychorique	Ordinale	Ordinale	Coefficient utilisée si la distribution des variables latentes sous-jacentes est normale.
Bisérial de point (point-biserial)	Nominale dichotomique	Intervalle	Coefficient équivalent à celui de Bravais Pearson d'un point de vue algébrique. Utilisé pour calculer les corrélations item-test le plus souvent. Si la corrélation de Bravais Pearson est la référence, le coefficient point-bisérial a tendance à surestimer la liaison.
Bisérial	Intervalle dichotomisé	Intervalle	Si la corrélation de Bravais Pearson est la référence, ce coefficient a tendance à sur-estimer la liaison.
Phi	Nominale (dichotomique)	Nominale dichotomique	Coefficient équivalent à celui de Bravais Pearson d'un point de vue algébrique.
Tétrachorique	Intervalle dichotomisé	Intervalle dichotomisé	Peu utilisée. Suppose que les deux variables latentes évaluées se distribuent normalement. Cas particulier de la corrélation polychorique.

Remarques :

- Pour la corrélation de Spearman, le coefficient bisérial de point ou le coefficient Phi, il existe dans tous les manuels des formules de calcul simplifiées. Mais algébriquement, on peut toujours appliquer la formule de Bravais-Pearson, en sachant qu'avec une échelle ordinale, il faut transformer les scores en rangs et pour le coefficient biserial de point comme pour le coefficient Phi, les variables dichotomiques prennent les valeurs 0 et 1. La formule simplifiée était utile à l'époque ou on effectuait encore de nombreux calculs partiellement à la main. Actuellement, ces formules présentent peu d'intérêt (mais sont toujours présentes dans les manuels).
- La signification (importance ou non) des valeurs des coefficients de corrélation varie selon la technique utilisée. Par exemple, pour le même jeu de données, le coefficient de corrélation bisérial est plus élevé que le coefficient bisérial de point (cf. vos cours de statistiques).

(* *Rappel* : le test du Chi carré (χ^2) permet de tester l'indépendance entre deux variables aléatoires (2 mesures dans notre cas) ou de tester l'adéquation d'une série de données à une distribution particulière (famille de loi de probabilité. ou autre)

(a) Corrélation de Bravais-Pearson

La corrélation (r_{xy}) de Bravais-Pearson (ou Pearson) correspond à la covariance divisée par le produit des écarts-types. La covariance (cov_{xy}) est la moyenne des produits des écarts à la moyenne sur chaque mesure. Soit deux mesures (échelles d'intervalles), X et Y, recueillies auprès d'un échantillon de n personnes (on a donc n paires d'observations) :

$$r_{XY} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

et

$$cov_{XY} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Attention : Pour un estimateur non biaisé, utiliser $1/(n-1)$ comme diviseur

Avec :

- x_i et y_i : les valeurs observées pour l'individu i sur les variables X et Y
- \bar{x} et \bar{y} : les moyennes des notes observées pour la variable X et Y

Remarques :

- lorsque la variance d'un test est nulle (donc l'écart-type = 0) la corrélation avec une autre mesure est indéterminée.
- La covariance de x avec x (cov_{xx}) est égale à la variance de x.

Interprétation (r de Bravais-Pearson)

La valeur du r de Bravais-Pearson peut prendre toutes les valeurs réelles comprises dans l'intervalle $[-1 ; +1]$. Plus la valeur absolue de r est proche de 1, plus il y a conformité avec le modèle linéaire. Un indice positif indique que les deux variables « évoluent » dans le même sens. Un indice négatif indique qu'il existe une relation inverse entre les deux mesures (*i.e.*: une augmentation de valeur sur une des variables est associée à une diminution des valeurs sur l'autre variable).

Attention : la corrélation de Bravais-Pearson évalue uniquement la conformité ou non avec le modèle linéaire et une corrélation nulle n'implique pas nécessairement qu'il n'y ait aucune liaison entre les deux mesures.

Ce qu'il faut savoir sur le coefficient de corrélation

- Une corrélation entre deux variables n'implique pas l'existence d'un lien causal(*) entre ces deux variables. Les mesures sont simplement associées selon le modèle de la corrélation utilisé (modèle linéaire pour la corrélation de Bravais-Pearson). C'est au psychologue de faire les hypothèses sur les relations causales éventuelles. La corrélation est purement descriptive en statistique.
- La force de la corrélation est donnée par la valeur absolue de la corrélation. Le sens de la relation entre les variables est donné par le signe de la corrélation.
- Le coefficient de Bravais-Pearson n'est pas modifié si l'on ajoute une même quantité à toutes les valeurs d'une distribution.
- Le coefficient de Bravais-Pearson n'est pas modifié si l'on multiplie par une même quantité (différente de 0) toutes les valeurs d'une distribution.
- Si une mesure discrimine peu les sujets (la mesure est peu sensible), la mise en évidence d'une éventuelle corrélation entre cette mesure et d'autres mesures sera plus difficile.
- De façon similaire, si l'on estime une corrélation sur une population sélectionnée, composée d'individus moins différenciés sur les mesures que ne le sont les sujets de la population générale, le coefficient de corrélation sera plus faible. L'interprétation de l'importance d'une corrélation doit prendre en compte les caractéristiques de l'échantillon de sujets.
- Les erreurs de mesure (sources de variations fortuites pour les valeurs observées) diminuent la valeur du coefficient de corrélation. Plus les sources fortuites de variance sont importantes, plus la part relative de variance explicable sera faible et plus la corrélation est faible. Donc plus la fidélité d'un test est faible, plus sa corrélation avec d'autres mesures sera faible.
- Les moyennes et les corrélations sont des résumés statistiques indépendants les uns des autres.
 - Une bonne corrélation entre deux mesures n'implique pas que les moyennes soient semblables sur les deux mesures. Cela implique que les classements des scores observés sur ces deux mesures sont semblables par rapport à la moyenne.
 - Deux moyennes semblables sur deux mesures n'impliquent pas non plus nécessairement qu'il existe une corrélation entre les mesures.
- Une corrélation de Bravais-Pearson nulle n'implique pas qu'il n'y a pas de liaison entre deux variables, mais signifie qu'il n'existe pas de corrélation linéaire.
- Si une épreuve A corrèle avec une épreuve B et que B corrèle avec une épreuve C, cela n'implique pas nécessairement que A et C corrèle (il n'y a pas de transitivité de la corrélation).

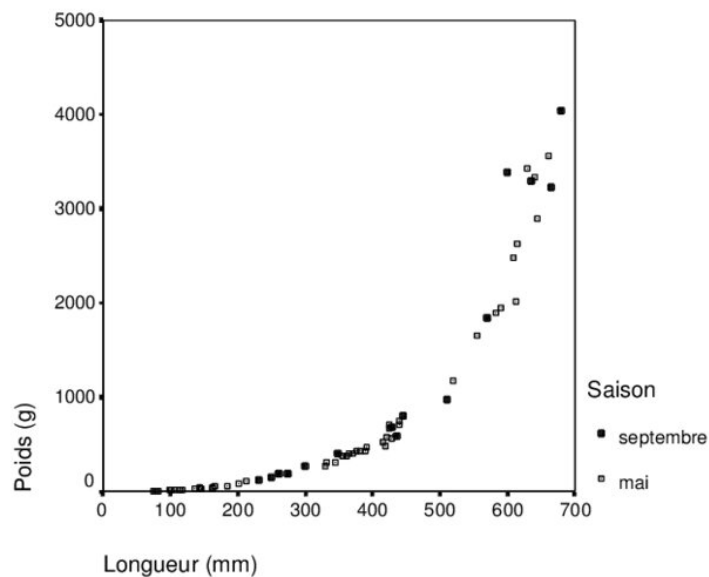
(*) C'est une erreur courante quand on interprète des corrélations. Pour vous dissuader de faire cette erreur, vous pouvez consulter le site "spurious correlations" de Tyler Vigen (www.tylervigen.com) qui nous présente des corrélations parfois supérieures à .95 et où il existe clairement aucune relation de cause à effet (corrélation fallacieuse). Vous serez très surpris de certaines de ces corrélations comme, par exemple, celle de .95 entre le nombre de doctorats obtenus en mathématiques et la quantité d'uranium dans les centrales nucléaires américaines pour la période 1995-2008. ou encore entre le nombre de suicides par pendaison ou strangulation et les dépenses pour la science, espace et technologie entre 1999 et 2009. Quand on recherche des corrélations (au "petit bonheur la chance"), on peut toujours en trouver ! Il faudrait vérifier les sources de ces données mais il n'y a aucune raison de ne pas observer de telles corrélations (des coïncidences artificielles).

Modèle linéaire

Que signifie le terme modèle linéaire ? On parle de modèle linéaire lorsque l'on suppose qu'il existe une relation monotone croissante ou décroissante entre deux variables X et Y (échelles d'intervalles ou de rapports) tel que il est possible de "prédire" le score sur Y à partir de X via une fonction de type $Y = ax + b$ et le score sur X à partir de Y via une fonction de type $X = a'Y + b'$

Par exemple, s'il existe une relation linéaire entre la taille et le poids, nous attendons une corrélation positive, qui traduirait le fait que plus la taille d'une personne est petite, plus cette personne est légère et plus la taille est importante et plus la personne est lourde. Cette formulation serait incomplète si l'on n'y ajoutait pas : pour une différence de taille (ou de poids) donnée, on observe la même différence de poids (de taille) quelle que soit la position relative des sujets dans l'échelle des tailles (dans l'échelle des poids). Bien entendu cette relation ne serait pas parfaite et la corrélation inférieure à 1 sur un échantillon représentatifs de personnes entre 5 et 30 ans par exemple. Il est possible aussi que la relation ne soit pas linéaire et que le modèle linéaire ne soit pas adapté pour tester la relation entre taille et poids (comme c'est l'exemple ci-dessous, qui s'intéresse à la truite Fario de la retenue d'eau de Naussac en Lozère).

Exemple d'une relation monotone croissante mais non linéaire entre la taille et le poids pour la truite Fario



(Source : Argillier, C., Cadic, N., Irz, P., Schlumberger, O., & Proteau, J.-P. (2015).

<https://doi.org/10.13140/RG.2.1.5153.5846>)

Seuil de signification

Suite au calcul d'un coefficient de corrélation, une question fréquente et légitime est : est-ce que le coefficient de corrélation observé est réellement différent de 0 ? Pour répondre à cette question il existe des tables de valeurs significatives de r en fonction de la taille de l'échantillon et du risque (seuil alpha) que l'on se fixe. On peut aussi, calculer la valeur p exacte et l'intervalle de confiance (cf. les cours de statistiques).

Il est aussi possible de tester si deux corrélations observées sont significativement différentes (cf. pour plus de détails, [Rakotomalala, 2025](#))

Attention

- La significativité de la corrélation va dépendre de la taille de l'échantillon (entre autre). Si l'échantillon a un effectif peu important, il peut être insuffisant pour affirmer (pour un risque alpha fixé) que cette corrélation est significativement différente de 0. Par contre si l'échantillon est de taille importante, une valeur de corrélation faible peut-être significative.
- Il ne faut donc pas confondre force de la corrélation et seuil de signification. Une corrélation peut avoir une valeur faible et être significativement différente de 0. Une corrélation plus forte pour un autre échantillon peut être non significativement différente de 0.

Représentation graphique

On peut représenter la relation entre deux variables quantitatives grâce à un nuage de points. Chaque point du graphique indique la position d'un individu selon ses valeurs sur les deux variables considérées, comme la taille et le poids. L'ensemble des points forme ainsi un nuage dont la configuration met en évidence la nature de la relation éventuelle entre ces variables. Dans le cas d'une relation linéaire on s'attend à observer un axe d'allongement du nuage de points, plus ou moins important et en relation avec la valeur de la corrélation linéaire qui existe entre ces deux variables. Le centre du nuage de point correspond à la moyenne sur les deux mesures.

Ci-dessous, sont présentés quelques cas typiques de nuages pour différents degrés de corrélation : (1) Corrélation moyenne à forte négative ($r = -.60$) ; (2) Une corrélation nulle ($r = .00$) ; (3) Une corrélation forte positive ($r = +.93$).

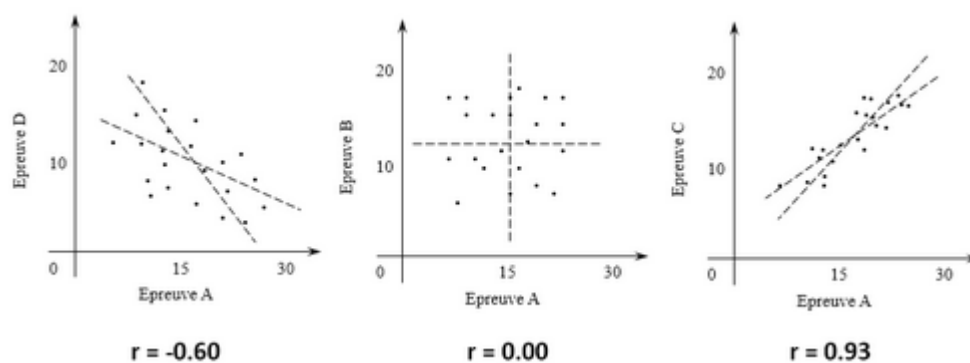


Figure H.3 : Exemples de nuages de points illustrant différents niveaux de corrélations

Les droites correspondent aux droites de régression qui expriment la relation linéaire entre deux variables quantitatives. On distingue la droite de régression de Y sur X (qui permet de prédire Y à partir de X) et la droite de régression de X sur Y (qui permet de prédire X à partir de Y).

Remarque

On ne devrait pas calculer de corrélations sans faire d'analyse graphique. C'est un outil privilégié pour visualiser la nature de la relation (linéaire ou non) mais aussi pour repérer des points "déviants" ou "aberrants" qui conduisent parfois à créer des relations artificielles. Deux exemples extrêmes sont donnés ci-dessous montrant la relation entre des notes scolaires en histoire (en abscisse) et en mathématiques (en ordonnée). Dans le premier graphique (figure B-4 gauche) la corrélation calculée est de .60 mais s'explique par un seul point (le point en rouge). En toute logique, ce point aberrant doit être pris en compte dans l'interprétation des données. Ici, il est probable qu'il faut exclure ce point de l'analyse (il s'agit en fait d'une erreur de saisie ou de transformation des notes, les notes étant sur une échelle de 0 à 20). La corrélation corrigée (suppression de ce point) est alors de .32 (ce qui change de façon significative la valeur de la corrélation). On peut aussi avoir des points aberrants qui inversent la corrélation ou l'annulent (cf. l'exemple de gauche ci-dessous). Il s'agit du même nuage de points mais le point aberrant est différent et la corrélation calculée en prenant en compte ce point est de .09 ! La note d'histoire (40) est probablement ici une erreur de saisie. Il faut donc corriger la saisie ou calculer la corrélation sans tenir compte de ce point.

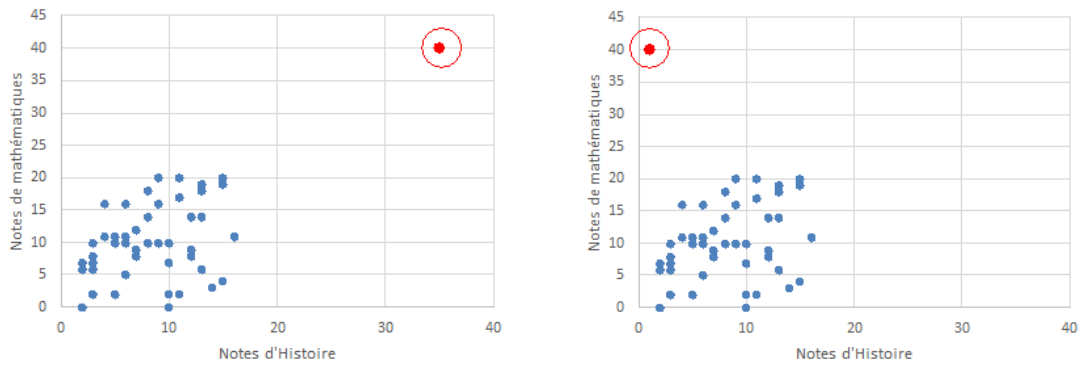


Figure H.4 : Nuages de points identiques traduisant la relation entre les notes observées sur 2 matières (histoire et mathématiques mais avec un point atypique (en rouge et encerclé) différent.

Variance expliquée

Une corrélation linéaire se traduit par une forme de nuage allongée (ellipse) et traduit (entre autre) le fait que l'on peut prédire le score observé sur une variable si on connaît le score observé sur l'autre variable avec moins de chance de se tromper que si on ne le connaissait pas (ce qui traduit la co-variation ou la co-dépendance entre les mesures) .

Par exemple, lorsqu'aucune information n'est disponible sur une variable, la meilleure estimation possible pour une nouvelle observation correspond à la moyenne des valeurs observées de cette variable. Le risque d'erreur associé à cette prédiction est d'autant plus élevé que la variance des observations de A est importante. Ainsi, l'intervalle de confiance lié à cette estimation dépend directement de la variance de A. En revanche, si l'on connaît la corrélation entre A et une autre variable B, et que la valeur de B est disponible pour la nouvelle observation, la prédiction de A peut être affinée. Elle tient alors compte de la relation entre les deux variables, ce qui permet de réduire la part de variance non expliquée. La prédiction devient ainsi plus précise, et la probabilité d'erreur diminue d'autant plus que la corrélation entre A et B est forte.

En fait, lorsque deux tests corrèlent, cela signifie qu'une partie de la variance de chacun des tests est "expliquée" par la variance de l'autre test (variance commune), c'est pourquoi l'erreur de pronostic est plus faible. On peut montrer facilement que **le pourcentage de la variance expliquée est égal au carré de la corrélation linéaire (r de Bravais Pearson)** entre les deux variables multiplié par 100 (le carré de la corrélation linéaire s'appelle le [coefficient de détermination](#)).

Exemples

corrélation	coefficient de détermination	% de variance expliquée
0	0	0
.10 ou -.10	.01	1%
.20 ou -.20	.04	4%
.30 ou -.30	.09	9%
.40 ou -.40	.16	16%
.50 ou -.50	.25	25%
.60 ou -.60	.36	36%
.70 ou -.70	.49	49%
.80 ou -.80	.64	64%
.90 ou -.90	.81	81%
1	1.00	100%

Coefficient de détermination

Dans le cadre des régressions linéaires simples (entre deux variables), le coefficient de détermination (r^2) est le carré du coefficient de [corrélacion](#) linéaire (Bravais-Pearson). Ce coefficient multiplié par 100 rend compte du pourcentage de [variance expliquée](#). Le coefficient de détermination représente donc la fraction de la variance d'une variable « expliquée » par l'autre autre variable. Il permet de juger de la qualité d'une relation.

Attention : Le terme de variance expliquée est un terme technique et en aucune manière cela signifie qu'une variable explique une autre. Cela ne doit pas conduire à en déduire une relation causale.

Correction pour atténuation

La correction pour atténuation (parfois appelé coefficient de désatténuation) dans l'analyse de la force d'une corrélation est une procédure pour tenir compte de l'erreur de mesure.

En effet la corrélation entre deux variables dépend directement de l'erreur de mesure. Plus la variance d'une variable correspond à de la variance d'erreur (*i.e.* plus les différences observées ont pour origine des facteurs aléatoires non contrôlés ou contrôlables), plus la corrélation avec une autre variable sera faible (Spearman, 1904a). La corrélation est une mesure de la part de variance commune à deux tests et seule la variance vraie peut être commune. Si deux mesures sont totalement aléatoires (c-à-d uniquement de l'erreur de mesure) les deux mesures ne corrèleront pas. En fait, la variance commune maximum possible est égale au produit des [fidélités](#) et la corrélation maximum possible est donc égale à la racine carrée du produit des fidélités (le coefficient de fidélité est une évaluation de la part de variance vraie).

La corrélation corrigée pour atténuation (cf. formule ci-dessous), en rapportant la corrélation observée (racine carrée de la variance commune) à la corrélation maximum possible, permet de tenir compte de cette erreur de mesure. On mesure ainsi la relation entre les scores vrais (relation qui nous intéresse directement).

$$r_{xy}^{corr} = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}}$$

avec : r_{xx} est la fidélité du test x

r_{yy} est la fidélité du test y

A savoir

- Cette correction pour atténuation est particulièrement utilisée lors de la recherche de preuves de la [validité des tests](#) (et plus particulièrement [la validité empirique](#) qui s'appuie sur la corrélation observée entre tests et critères).
- La corrélation maximum entre deux mesures est comprise entre les valeurs des deux fidélités. Plus précisément, si deux tests ont une fidélité égale respectivement à .70 et .80, la corrélation entre ces deux tests sera égale au maximum à la racine carrée de .70 * .80 soit $r(\max) = .748$
- Plus la fidélité est forte, plus la corrélation (sans correction pour atténuation) avec un autre test ou un critère pourra être forte. Une fidélité faible minimise à l'inverse les corrélations.
- Une faible corrélation entre deux tests peut donc avoir pour origine une fidélité faible de l'un ou des deux tests.

Corrélations partielles

La corrélation observée entre deux variables peut être artificielle (cf. exemple ci-dessous), masquée ou sur-évaluée en raison d'une ou plusieurs variables confondantes. Selon la nature de/des variable(s) confondante(s), la stratégie d'analyse est différente :

- **La variable confondante est une échelle d'intervalle.**

Le principe est alors de calculer un coefficient de **corrélation partielle** en retirant la variance qui est due à une troisième variable Z (corrélation partielle entre X et Y notée alors $r_{XY.Z}$). Cet indice de corrélation partielle permet par exemple de calculer la corrélation entre deux tests après avoir retiré l'effet de l'âge, c'est à dire après avoir retiré la variance des notes due à l'âge des sujets. La formule de calcul est simple :

$$r_{XY.Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}$$

Lorsque qu'il existe plusieurs variables confondantes qui sont des échelles d'intervalles, la corrélation partielle est alors une **corrélation partielle d'ordre p** ($r_{xy.z_1z_2...z_p}$) et la formule est alors plus complexe (matricielle). Il est souvent préférable si p est supérieur à 3 de passer par des techniques de régression (non présentée dans ce rappel "statistiques").

- **La variable confondante est qualitative**

Une variable qualitative ((qui permet de distinguer différents groupes) conduit à calculer la corrélation pour chaque groupe. On peut ainsi avoir des surprises avec par exemple une corrélation négative entre deux variables x et y, qui devient positive pour chacun des groupes (c'est une des expressions du paradoxe de Simpson*, cf. figure B-5) ou encore des corrélations qui varient selon les groupes et qui sont très différentes de celles observées globalement.

Dans l'exemple donné (figure B-5) les scores sur la variable X augmentent avec l'âge et alors qu'ils diminuent pour Y avec l'âge. La corrélation entre X et Y est négative (nuage de points orienté vers la gauche). Pour les 4 groupes distingués par la variable Z (4 groupes d'âge correspondant aux 4 couleurs dans le nuage de points), les corrélations entre X et Y sont toutes positives ! Cet effet peut sembler paradoxale mais on doit toujours y penser lorsque l'on regroupe des données de différentes études ou à l'inverse dans les études développementales avec des enfants d'âge très différent.

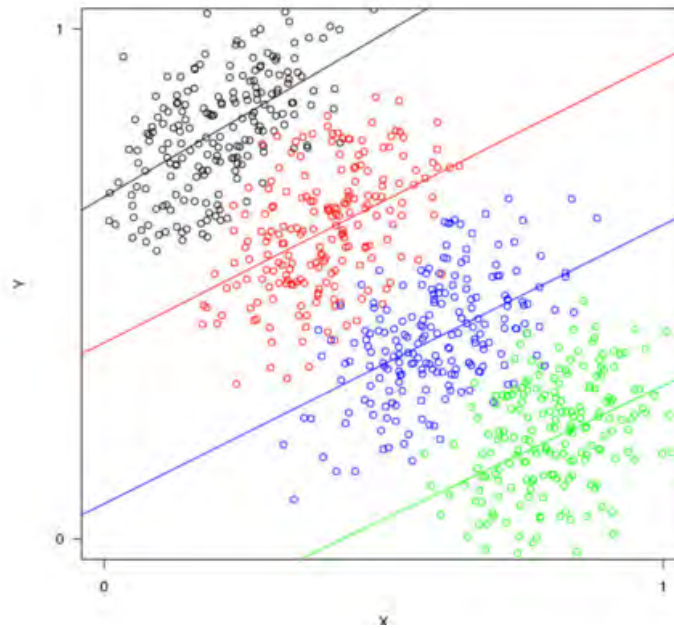


Figure H.5 : Illustration d'une des illustrations du paradoxe de Simpson dans le domaine des corrélations (adapté de Rücker & Schumacher, 2008)

(*) Le paradoxe de Simpson (ou effet de Yule-Simpson) a été décrit initialement par Udny Yule en 1903 puis repris par [Edward Simpson en 1951](#). De façon générale, cet effet correspond à l'inversion d'un effet (fréquence de guérison, corrélation, etc.) observé dans plusieurs groupes lorsque l'on regroupe toutes les données (par exemple une différence de moyennes entre deux conditions est positive dans un premier groupe, positive dans le second groupe mais s'inverse quand on combine les deux groupes). Pour ceux qui veulent mieux comprendre ce paradoxe ou voir des exemples surprenants, cf. https://www.youtube.com/watch?time_continue=11&v=vs_Zzf_vL2I

(b) Corrélation de Spearman

Le rho de Spearman (ρ) est le coefficient de corrélation que l'on utilise lorsque les variables ne sont pas des variables d'intervalle mais des variables ordinales. C'est un test non paramétrique (pas d'hypothèse sur les paramètres). En fait, le principe de ce coefficient est d'appliquer la formule du coefficient de Bravais-Pearson non pas sur les valeurs observées mais sur les rangs (pour chaque variable on remplace le score observé par son rang). Compte tenu de certaines propriétés des rangs (par exemple la moyenne de n scores exprimés en rang est égal à $(n+1)/2$), on donne souvent comme formule de calcul du rho de Spearman une version simplifiée :

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

avec : d_i l'écart entre les rangs (pour chaque paire d'observation)

Remarques :

- Lorsqu'il y a des ex-æquos on affecte comme rang, la moyenne des rangs qui auraient été affectés à chaque sujet s'ils n'avaient pas été ex-æquo (par exemple si entre le rang 4 et 9 on trouve 4 ex-æquos, le rang pour ces quatre observations sera $5+6+7+8$ divisé par 4 soit le rang 6.5). Si le nombre d'ex-æquos est important, il faut toutefois corriger le coefficient de Spearman. Le plus simple est d'utiliser la formule de Bravais-Pearson sur les rangs des scores (en affectant toujours le rang moyen pour les ex-æquos), il n'y a pas besoin de corriger (c'est plus simple !).

- Ce coefficient peut permettre d'évaluer une liaison non linéaire à la différence du Bravais-Pearson, à condition que la liaison soit monotone. Lorsque cette fonction est non monotone, le rho de Spearman est inopérant (comme le coefficient de Bravais Pearson).
- Ce coefficient est plus "robuste" face à des points aberrants. Dans l'exemple concernant le point aberrant proposé plus haut [précédemment](#) la corrélation de Spearman (avec le point aberrant) passe à .36 (bien plus proche de la valeur observée sans ce point aberrant qui est de .32).

1.2.5 Quantiles d'ordre n

Dans une distribution de scores ordonnés (ensemble de notes par exemple), on appelle quantile d'ordre n chacune des $n - 1$ valeurs qui partagent l'étendue des scores en n sous-ensembles d'effectifs égaux. La médiane est un quantile d'ordre 2 (elle partage en 2 ensembles d'effectifs égaux un ensemble de valeurs ordonnées). Les quantiles les plus connus sont les déciles (quantiles d'ordre 10), les quartiles (quantiles d'ordre 4) et les centiles (quantiles d'ordre 100).

De nombreuses méthodes de calcul des quantiles existent. Nous présenterons dans le paragraphe suivant une de ces méthodes.

1.2.6 Centiles - Percentiles

Les centiles (ou percentiles qui est un anglicisme) sont les valeurs d'une variable qui partitionnent la distribution des scores ordonnés en 100 intervalles contenant le même nombre de données (quantiles d'ordre 100). Il y a donc 99 centiles (99 valeurs de la variable correspondant chacune à un centile).

Déterminer les centiles d'une série de valeur :

La méthode qui semble la plus utilisée est la suivante (Nearest Rank method).

1. Trier les valeurs par ordre croissant : $X_1 < X_2 < \dots < X_i < \dots < X_{n-1} < X_n$
2. Le centile P est la valeur du $k_{i\text{ème}}$ élément avec $k = P * n / 100$ (la valeur de k est arrondie au nombre entier supérieur le plus proche).

La méthode recommandée par le National Institute of Standards and Technology (NIST).

1. Le rang k est calculé de la façon suivante : $k = P*(n+1)/100$.
2. La valeur k est ensuite séparée en deux valeurs, sa partie entière (e) et sa partie décimale (d). Le centile est ensuite déterminé selon la règle suivante (avec v_i la $i\text{ème}$ valeur observée dans la série ordonnée) :
 - si $e=0$ alors le centile est v_1 (la première valeur observée)
 - si $e=n$ alors le centile est v_n (la dernière valeur observée)
 - sinon le centile se calcule par interpolation linéaire et est égale à : $v_e + d*(v_{e+1} - v_e)$

Une méthode alternative (toujours recommandée par le NIST) : similaire à la précédente, sauf que la valeur k est égale à $1 + P*(n-1)/100$

Remarques :

- Le terme de centile (percentile) a été utilisé pour la première fois par Francis Galton à la fin du 19^{ème} siècle.
- Les méthodes de calcul des centiles peuvent être différentes d'un logiciel à l'autre et donner des résultats légèrement différents.

- Les centiles doivent être différenciés des [rangs centiles \(ou rangs percentiles\)](#). Un centile est une valeur de la variable pour un rang centile entier (1 à 99), alors que le rang centile est le rang associé à une valeur de la variable.
- Il faut faire attention à l'interprétation des centiles et des rangs centiles car on a tendance à sur-interpréter ou sous-interpréter les valeurs observées. (cf. à ce sujet, Bowman, 2002). Les centiles ne sont pas à équidistance les uns des autres lorsque la distribution n'est pas "rectangulaire" pour une échelle d'intervalle (cf. ci-dessous les rangs centiles associés à la loi normale).

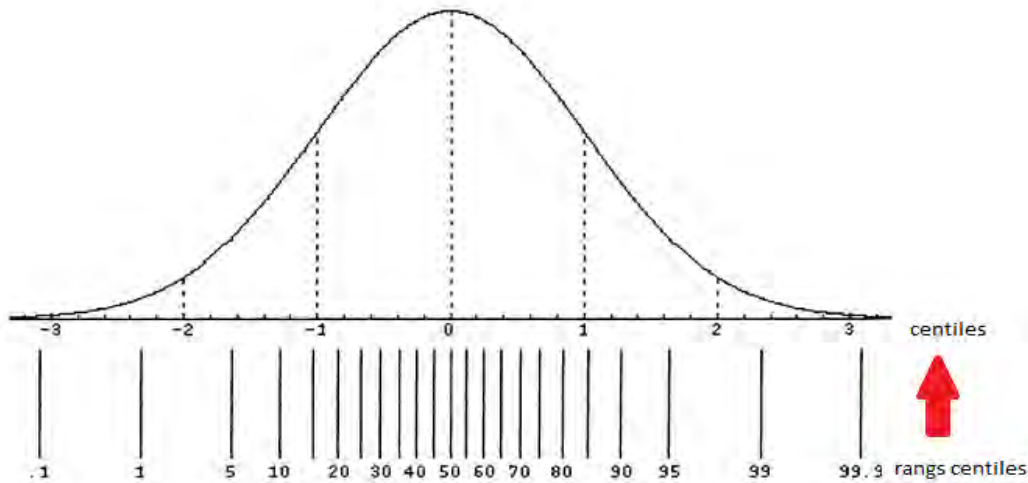


Figure H.6 : Illustration des distances entre centiles pour une distribution normale (d'après Bownam, 2012)

1.3. La loi normale

Lorsqu'une série de mesure subit l'influence de sources aléatoires alors les caractéristiques de cette série répondent à la loi normale (théorème central limite).

La loi normale est la plus connue des lois de probabilité. Sa fonction de densité a une forme simple (courbe en cloche) et est symétrique et presque toutes les valeurs se trouvent entre moins trois écarts-types et plus trois écarts-types de la moyenne (plus de 99%). On notera aussi que 95% des valeurs se trouvent à ± 1.96 écart type de la moyenne).

Définition formelle. $X \sim \mathcal{N}(\mu, \sigma^2)$ est définie sur \mathbb{R} (ensemble des réels) par :

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2}$$

La représentation graphique est la suivante :

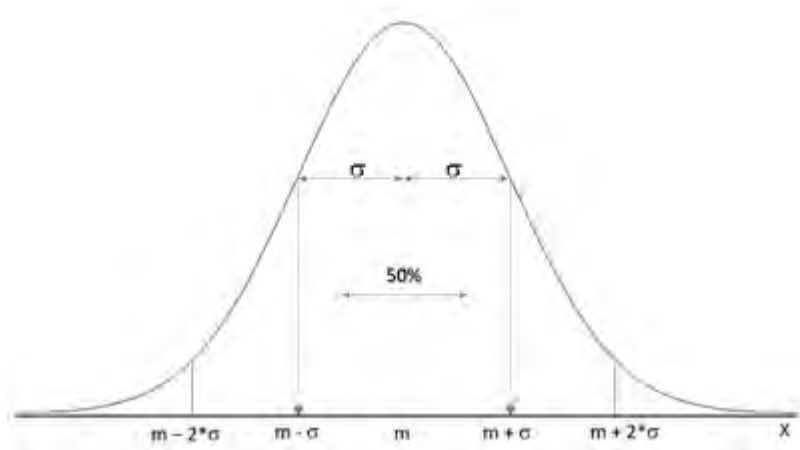


Figure H.7 : Distribution normale et ses 2 paramètres (m et σ)

Remarques

- Cette courbe, caractéristique de la loi normale, est aussi appelée courbe de Gauss en l'honneur de [Karl Friederich Gauss \(1777-1855\)](#). La fonction associée a aussi pour nom loi de Laplace-Gauss (Pierre Simon Laplace, 1749-1827, étant un autre grand mathématicien, astronome, physicien et philosophe).
- Le théorème central limite énonce, dans sa forme générale, que sous certaines conditions, notamment l'existence de moyennes finies et de variances finies et non nulles, la moyenne d'un grand nombre de variables aléatoires indépendantes* et identiquement distribuées tend vers une loi normale, quelle que soit leur distribution d'origine.

(*) Deux variables aléatoires sont dites indépendantes quand le résultat de l'une n'influence pas le résultat de l'autre.

1.3.1 Table de la loi normale

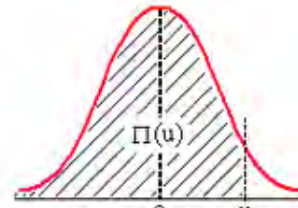
La table de la loi normale centrée réduite présentée ci-après donne pour une valeur u , la densité de probabilité correspondant à $p(x < u)$. La lecture de cette table est donc facile. Cette table ne concerne que les valeurs de u supérieures à 0, mais on peut déduire facilement par symétrie les valeurs de u inférieures à 0. Cette feuille est téléchargeable en [format pdf](#). Il existe aussi un [calculateur SCALP](#).

Exemples.

Valeur de x si $u = 1.23$. On regarde la valeur à l'intersection de la ligne 1.2 et la colonne 0.03 ($1.2 + 0.03 = 1.23$). La valeur dans la table est **0.8907**, il y a donc 89,07% ($0.8907 * 100$) des valeurs de cette distribution qui sont inférieures à 1.23. Si la note d'une personne dans un test de performance est à 1.23 écart type de la moyenne, on pourra donc dire qu'il fait mieux que 89% des personnes de l'échantillon ayant conduit à construire ce test (sous condition que la distribution des scores soit normale !).

Valeur de x si $u = -0.72$, on recherche la valeur dans la table correspondant à +0.72, et on soustrait cette valeur de 1 (on prend le complément). La valeur lue dans la table (ligne 0,7 et colonne 0,02) correspond à 0.7642. On retire donc à 1 cette valeur car u est négatif : $1 - 0.7642 = 0.2358$. Si la note d'une personne dans un test de performance est à -0,72 écart type de la moyenne, on pourra donc dire qu'il fait mieux que 24% des personnes de l'échantillon ayant conduit à construire ce test (sous condition que la distribution des scores soit normale !).

Table de Loi Normale
P(x<u)



	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8254	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

1.3.2 Valeurs fréquemment utilisées

Certaines valeurs ou bornes de la table de la loi normale sont souvent utilisées en psychologie (et sont des repères connus de la plupart des psychologues pour lire des résultats).

- **% de valeurs supérieures à une borne**

- 15% des valeurs sont > 1.04 σ (à + 1σ on trouve 15.87% des valeurs)
- 5% des valeurs sont > 1.644 σ (à + 1,5σ on trouve 5.68% des valeurs)
- 2.5% des valeurs sont > 1.96 σ (à + 2σ on trouve 2.28% des valeurs)
- 0.15% des valeurs sont à > 3 σ

▪ **Par symétrie, % de valeurs inférieures à une borne**

→ 15% des valeurs sont $< -1.04 \sigma$ (à -1σ on trouve 15.87% des valeurs)

→ 5% des valeurs sont $< -1.644 \sigma$ (à $-1,5\sigma$ on trouve 5.68% des valeurs)

→ 2.5% des valeurs sont $< -1.96 \sigma$ (à -2σ on trouve 2.28% des valeurs)

→ 0.15% des valeurs sont à $< -3 \sigma$

▪ **% de valeurs entre deux bornes**

→ 70% des valeurs sont à $\pm 1.04 \sigma$ (à $\pm 1\sigma$ on trouve 68.26% des valeurs)

→ 90% des valeurs à $\pm 1.644 \sigma$ (à $\pm 1,5$ on trouve 88.64% des valeurs)

→ 95% des valeurs sont à $\pm 1.96 \sigma$ (à ± 2 on trouve 95.44% des valeurs)

→ 99.7% des valeurs sont à $\pm 3 \sigma$

Toutes les autres valeurs peuvent être calculées à partir du [calculateur SCALP](#).

2. Introduction à l'analyse factorielle

En France, le terme analyse factorielle désigne une famille de méthodes d'analyse de données qui regroupe, souvent à tort, à la fois l'analyse en composantes principales (ACP) et l'analyse factorielle exploratoire (AFE). L'objectif général de ces techniques est non seulement la réduction de la dimensionnalité (dans le cas de l'ACP), mais aussi l'identification des facteurs d'organisation d'un nuage de points dans un espace à k dimensions, afin de résumer l'information à l'aide d'un nombre restreint de variables. Concrètement, ces méthodes visent à rendre intelligible soit la variance des scores (ACP), soit la structure des corrélations entre variables (AFE), en construisant un ensemble réduit de composantes ou de variables latentes non corrélées. En résumé, l'analyse factorielle regroupe des techniques destinées à révéler les structures sous-jacentes d'un ensemble de données, qu'il s'agisse de composantes principales ou de facteurs latents..

L'accès à ces techniques d'analyse n'est pas toujours aisé même si l'explosion des logiciels, dédiés ou non, a facilité leur mise en œuvre (et parfois une utilisation incorrecte). Nous présenterons ici l'analyse en composantes principales et l'analyse en facteurs communs et spécifiques qui ne sont pas des méthodes inférentielles mais des méthodes descriptives. Elles ne spécifient pas à l'avance quelles variables doivent être associées à tels facteurs ou composantes et elles décrivent les données concernant la population sur laquelle ces données ont été recueillies.

Actuellement, il existe des développements importants en analyse de données et par exemple des techniques dites d'analyse factorielle confirmatoire qui permettent de tester des hypothèses a priori concernant à la fois le nombre de facteurs et l'appartenance de chaque variable à un facteur. On peut aussi, avec des techniques plus complexes (modèles d'équations structurales) tester des relations ou des rapports de causalités multiples entre facteurs (variables latentes non observables). Nous ne ferons qu'aborder ("effleurer") ici les techniques confirmatoires même si actuellement ces méthodes deviennent les outils les plus utilisés dans la construction des tests (sélection des items et études de la validité).

Remarque

Il a été ajouté quelques définitions formelles dans ce cours pour ceux qui auraient un minimum de connaissances en algèbre linéaire. L'objectif cependant est de donner des définitions "simplifiées" des principaux concepts utilisés pour comprendre les résultats de ces techniques d'analyse de données. Il

n'est pas demandé de connaître ou comprendre ses définitions formelles pour ceux qui n'auraient pas les prérequis en algèbre linéaire.

Un peu d'histoire

Les méthodes d'analyse de données (analyse factorielle exploratoire) remontent aux travaux de Spearman (1904) avec le concept de facteur. Le terme d'analyse factorielle reviendrait à Thurstone (1931) et celui d'analyse en composante principale à Hotelling (1933). Depuis, le nombre des méthodes a explosé mais une bonne compréhension de l'ACP et de l'AFE permet facilement d'aller plus loin ensuite. Les techniques d'analyse factorielle ne concernent pas que les échelles ordinales ou d'intervalles. Par exemple, l'analyse factorielle des correspondances concerne les grandes tables de contingence et a été introduite par [Benzecri](#) dans les années 60 (Benzecri, 1982).

2.1. La réduction des données

La réduction des données constitue une réponse pertinente à l'explosion du nombre d'indicateurs lorsque le nombre de mesures (ou de variables) augmente. En statistique descriptive, une mesure peut être résumée par un indicateur de tendance centrale (ex. : la moyenne ou la médiane) et un indicateur de dispersion (ex. : l'écart-type ou l'écart interquartile). Ainsi, pour une dimension mesurée, deux valeurs au minimum sont nécessaires pour résumer l'information. Certains auteurs suggèrent d'ajouter également des indicateurs de forme de la distribution, tels que l'asymétrie ou l'aplatissement.

Lorsqu'on étudie deux variables (par exemple la taille et le poids, les temps de réponse et la qualité des réponses dans une tâche de mémorisation, ou encore les performances verbales et non verbales), cinq valeurs sont alors nécessaires pour résumer les données : un indice de tendance centrale et un indice de dispersion pour chacune des deux variables, ainsi qu'un indice d'association (comme le coefficient de corrélation de Bravais-Pearson) entre elles. Avec trois variables, neuf indicateurs sont requis ; avec dix variables, on atteint soixante-cinq indicateurs !

Le nombre de descripteurs nécessaires pour résumer les données croît donc très rapidement avec le nombre de variables (cf. tableau ci-dessous). Les techniques d'analyse factorielle présentent, entre autres, l'intérêt de permettre une réduction et une synthèse des données afin de les rendre plus intelligibles.

Nombre de VD	Nombre de résumés	
1 variable ->	2	1 tendance centrale et 1 dispersion
2 variables ->	5	2 tendances centrales, 2 dispersions, 1 corrélation
3 variables ->	9	3 tendances centrales, 3 dispersions, 3 corrélations
10 variables ->	65	10 tendances centrales, 10 dispersions, 45 corrélations
n variables ->	$2n+n(n-1)/2$	n tendances centrales, n dispersions, $n*(n-1)/2$ corrélations

2.2. Décomposition linéaire

Le modèle utilisé en analyse factorielle est un modèle linéaire. Par exemple, dans le cas de l'analyse factorielle exploratoire (AFE), on postule que les variables observées sont des combinaisons linéaires

d'un ensemble plus restreint de facteurs latents (ou variables sous-jacentes non observées)

Pour illustrer et comprendre ce principe, imaginons que deux facteurs V et W représentent deux dimensions psychologiques théoriques expliquant les corrélations observées entre cinq mesures (des tests par exemple, X_1 à X_5). La position d'un individu sur chacun des tests peut alors s'exprimer comme une combinaison linéaire de ses positions sur ces deux facteurs communs, d'un **facteur spécifique** propre à chaque test (S_i), et d'une **erreur résiduelle** (ϵ_i). Cette hypothèse se traduit en algèbre par une équation de décomposition :

$$X_{ki} = a_k V_i + b_k W_i + c_k S_{ki} + \epsilon_i$$

avec

X_{ki} , le score observé pour le sujet i dans l'épreuve k .

V_i et W_i sont les scores du sujet i sur les facteurs communs V et W

S_{ki} le score du sujet i pour facteur spécifique à la variable X_k .

a_k, b_k, c_k , les poids de ces facteurs pour l'épreuve k (charges factorielles)

ϵ_i : l'erreur aléatoire de mesure associé au sujet i .

Cette décomposition traduit le fait, si nous prenons en considération les réponses (plus exactement la variance des réponses) au test X_k , que :

- le facteur V explique une partie de la variance X_k (et ce d'autant plus que a_k sera grand)
- le facteur W explique une autre partie de la variance X_k (et ce d'autant plus que b_k sera grand)
- le facteur S_k , explique la variance inexpliquée par V et W et qui est spécifique à l'épreuve X_k

Cette équation revient donc à expliquer, en la décomposant, le score observé au test X_k et donc la variance de X_k . Pour un sujet i , on obtient donc pour 5 épreuves, 5 équations :

$$X_{2i} = a_2 V_i + b_2 W_i + c_2 S_{2i} + \epsilon_i$$

$$X_{3i} = a_3 V_i + b_3 W_i + c_3 S_{3i} + \epsilon_i$$

$$X_{4i} = a_4 V_i + b_4 W_i + c_4 S_{4i} + \epsilon_i$$

$$X_{5i} = a_5 V_i + b_5 W_i + c_5 S_{5i} + \epsilon_i$$

Remarque :

- **Pour l'analyse en composantes principales**, le principe est légèrement différent. On suppose qu'il existe initialement autant de composantes que de variables et ce sont les composantes qui sont des combinaisons linéaires des variables observées.

$$C_k = a_{1k} X_1 + a_{2k} X_2 + \dots + a_{pk} X_p$$

Dans l'interprétation des résultats, seules certaines composantes extraites, celles qui expliquent le plus de variance, seront cependant pris en compte (voir : [fixer le nombre de facteurs](#)).

2.3. Analyse en Composantes Principales (ACP)

L'objectif, lorsque l'on réalise une ACP, est de réduire les données, c'est-à-dire avoir une méthode pour obtenir un nombre réduit de composantes non corrélées. En terme clair c'est une technique d'analyse de données qui consiste à transformer des variables corrélées entres elles en nouvelles variables (composantes) non corrélées. Il faut savoir que :

- Le nombre de composantes extrait est initialement identique au nombre des variables initiales (et explique toute la variance du nuage des points dans l'espace à n dimensions défini par les variables initiales) mais, en pratique, on interprétera uniquement les premières composantes qui sont par construction (cf. plus loin) les plus explicatives (qui rendent compte d'une part significative de la variance).
- Lors de l'extraction les composantes (première étape de l'analyse) sont définies comme indépendantes les unes des autres (« orthogonales »). La position d'un individu sur une composante (*i.e.* un facteur) n'implique en rien sa position sur une autre composante.
- Une hypothèse complémentaire est ajoutée pour permettre de résoudre le système : la première composante doit expliquer le plus de variance possible (*i.e.* doit être au plus près de tous les points du nuage de points. La seconde (orthogonale à la première) doit expliquer le plus de la variance non expliquée, la troisième composante le plus de variance non expliquée par les deux premières, etc.
- Enfin, le plus souvent, l'analyse est faite sur des variables centrées-réduites (note z). En effet, si les variables n'étaient pas réduites et qu'une des variables avait une variance plus importante que les autres (quantitativement), la première composante aurait naturellement tendance à expliquer cette variable (cf. ci-dessus). Les réduire (ramener la variance à 1) fait que toutes les variables ont le même poids dans l'analyse.

En pratique, pour effectuer une ACP, on doit successivement (démarche générale simplifiée et ce sont ces éléments que nous allons reprendre dans les parties suivantes) :

- construire ou sélectionner une batterie d'épreuves ou de mesures.
- sélectionner la population sur laquelle on administre ces épreuves.
- calculer la corrélation entre les scores pour toutes les paires de tests ; on obtient ainsi une [matrice de corrélations](#).
- effectuer la première étape de l'ACP (via un logiciel d'analyse) et on s'intéresse plus particulièrement au tableau des [valeurs propres](#) mais aussi l'évolution des [communautés](#) en fonction du nombre de composantes que l'on pourrait retenir.
- décider du nombre de composantes (facteurs) à retenir.
- vérifier que les épreuves sont bien expliquées par ce système de facteurs (le pourcentage de variance cumulée expliqué par les facteurs doit être proche des [communautés](#) observées avec les facteurs retenus).
- décider si on va procéder à une [rotation des facteurs obtenus](#). Il s'agit de passer de facteurs initiaux à de nouveaux facteurs plus aisément interprétables.
- interpréter les facteurs/composantes.

2.3.1 Matrices des corrélations

La matrice des corrélations est tout simplement la matrice des coefficients de corrélation (de Bravais-Pearson pour l'ACP). Comme le montre l'exemple suivant les valeurs au-dessus et au-dessous de la diagonale sont donc identiques puisque la corrélation entre un test A et B est évidemment la même que celle observée entre B et A.

Exemple d'une matrice de corrélation pour l'ACP

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
V1	1.00	.41	.16	.003	.61	.07	.20	.13	.24	.09
V2	.41	1.00	.20	.06	.36	-.01	.15	-.03	.09	.05
V3	.16	.20	1.00	-.01	-.15	-.09	.06	.04	.11	.24
V4	.003	.06	-.01	1.00	.23	-.03	.41	.10	-.04	-.06
V5	.51	.36	-.15	.23	1.00	.01	.20	.12	.09	.03
V6	.07	-.01	-.09	-.03	.01	1.00	.23	.11	.31	.00
V7	.20	.15	.36	.41	.20	.28	1.00	.07	.18	.03
V8	.13	-.03	.04	.10	.12	.11	.07	1.00	.34	.05
V9	.24	.09	.11	-.04	.09	.31	.18	.34	1.00	.01
V10	.00	-.05	.24	-.06	.03	.09	.03	.05	.01	1.00

Pourquoi cette matrice est aussi une matrice de variance-covariance ?

Les analyses en composantes principales effectuées en psychologie sont le plus souvent des ACP normées (on effectue les analyses sur les variables centrées et réduites de façon à ce que chaque variable ait le même poids dans l'analyse). La corrélation étant la covariance divisée par le produit des écart-types, la covariance est donc égale à la corrélation lorsque les variables sont centrées et réduites. Dans la diagonale se trouve la valeur 1 qui correspondent à la variance de chaque variable. Lorsque les variables sont centrées réduites la matrice de corrélation est donc identique à la matrice de variance-covariance.

A vérifier avant de commencer une analyse

- Si tous les coefficients de corrélation d'une matrice de corrélations sont faibles (proches de 0) il n'y a absolument aucun intérêt à procéder à une ACP car pour que celle-ci ait un sens il faut qu'il existe suffisamment de corrélations significatives entre les variables. A l'extrême, la matrice pourrait être une matrice d'identité (matrice dont le [déterminant](#) serait de 1). Pour savoir si on a affaire à une matrice de ce type, il existe le test sphéricité de Bartlett (non présenté ici). Quand il est significatif, on rejette l'hypothèse d'identité.
- A l'inverse, il ne faut pas non plus, dans cette matrice, qu'il y ait des variables parfaitement corrélées (condition dite de « [singularité](#) ») ou qu'une variable soit parfaitement corrélée avec une combinaison de plusieurs variables. Pour savoir si la matrice est "singulière", on peut calculer le « déterminant » de la matrice. Ce déterminant ne devrait pas être inférieur à 0.00001.

Exemple : pour la matrice présentée, le déterminant est 0.20 (la matrice n'est pas singulière).

- Le déterminant et le test de sphéricité de Bartlett nous aident à vérifier si une matrice de corrélation possède les propriétés nécessaires pour effectuer une ACP. Il est également important d'examiner chacune des variables. En effet si une variable corrèle avec aucune autre il est recommandé de retirer cette variable de l'analyse.

L'examen des variables peut être réalisé par le calcul d'un indice le KMO (Kaiser-Meyer-Olkin) pour chacune des variables et pour la matrice globale. Il nous renseigne sur la qualité des corrélations (mesure d'adéquation de l'échantillon ou en anglais Measure of Sampling Adequacy). Cet indice prend

des valeurs entre 0 et 1 et sa valeur devrait être égale ou supérieure à .50 [on accepte la gradation suivante : inacceptable en dessous de .50, médiocre entre .50 et .60, moyen entre .60 et .70, bien entre .70 et .80, très bien entre .80 et .90 et excellent au delà de .90].

Remarque : pour les termes techniques comme [singularité](#), [déterminant](#), etc. cf. le glossaire.

2.3.2 Saturations

L'AFC va consister à extraire des composantes (appelées parfois facteurs dans les logiciels ou articles). Plus une composante contribue "à expliquer" une variable observée, plus la corrélation entre cette composante (les scores sur la composante) et la variable observée sera élevée. Cette corrélation entre une variable et une composante correspond à ce qu'on appelle la saturation.

Exemple d'un tableau de saturation

Un tableau de saturation est une matrice dans laquelle pour chaque variable (en ligne) on indique la saturation observée avec [les composantes qui ont été extraites](#). On trouve parfois dans ce tableau les [valeurs propres](#) sur la dernière ligne et les [communautés](#) (h^2) dans la dernière colonne. Dans l'exemple suivant, 4 composantes (F1, F2, F3, F4) sont extraites et les saturations, valeurs propres et communautés sont reportés dans cette table des saturations.

	F1	F2	F3	F4	h^2
Variable 1	.766	-.244	.273	.215	.76
Variable 2	.559	-.432	.248	.019	.56
Variable 3	.177	.078	.640	-.565	.77
Variable 4	.327	-.144	-.610	-.525	.77
Variable 5	.712	-.404	-.114	.260	.75
Variable 6	.301	.613	-.136	.127	.50
Variable 7	.564	.151	-.422	-.446	.72
Variable 8	.352	.475	-.027	.163	.38
Variable 9	.483	.578	.120	.247	.64
Variable 10	.133	.245	.451	-.395	.45
Valeurs propres	2.32	1.45	1.37	1.16	6.3

A savoir

- Les saturations varient (comme les corrélations) entre -1 et +1. Plus la valeur absolue de la corrélation est élevée plus la variable contribue à la composante.
- Pour une variable donnée et une composante, plus la valeur absolue de la saturation est élevée, plus la composante est "proche" ("similaire") de la variable considérée.
- La part de variance observée expliquée par une composante correspond au carré de la saturation de cette variable par cette composante. Par exemple, dans le tableau précédents, la part de variance expliquée de la variable 6 par C1 est de .301 X .301 soit 0.09. Le pourcentage de variance de la variable 6 expliqué par la première composante (premier facteur) est donc de 9%.
- Le tableau des saturations et la connaissance que l'on a des variables empiriques (variables observées) permettront d'analyser la signification des facteurs extraits (cf. [interprétation des](#)

[résultats](#)).

2.3.3 Valeurs propres et vecteurs propres

L'ensemble des [saturations](#) des variables pour une composante correspondent à un [vecteur propre](#). La **valeur propre*** (ou "eigenvalue") est la somme des carrés de ces saturations. Elle représente la quantité de variance du [nuage de points](#) expliquée par cette composante.

Le rapport de la valeur propre au nombre de variables soumises à l'analyse donne le pourcentage de variance expliquée par la composante (taux d'inertie).

Notes :

- Les valeurs propres peuvent prendre des valeurs comprise entre 0 et la [quantité de variance à expliquer](#). En ACP (telle qu'elle est utilisée en psychologie) la quantité de variance à expliquer est égale au nombre des variables (car les variables sont centrées réduites pour l'analyse et chaque variable à une variance de 1).
- Avant toute [rotation](#), la valeur propre de la première composante est toujours la plus élevée (elle rend compte du maximum de variance), ensuite vient la valeur propre de la seconde composante (celui-ci rend compte du maximum de variance restant à expliquer), puis la valeur propre de la troisième composante, etc.). Cette propriété est la conséquence des contraintes fixées lors de la méthode d'extraction des facteurs.
- Sachant qu'en ACP on interprète les n premières composantes, la quantité de variance expliquée par ces n premières composantes ensemble est égale à la somme de leur valeur propre. Le pourcentage de variance expliquée par le système de facteur est donc cette somme des valeurs propres divisée par la [trace de la matrice de variance covariance](#) (soit le nombre des variables en ACP, cf. glossaire) le tout multiplié par 100. Cette valeur qui est aussi égale à la somme des [communautés](#) des composantes que l'on retiendra (cf. plus loin) et traduit l'importance du système de facteurs retenus.

Exemple

Le tableau suivant reprend la table des saturations. Dans ce tableau, la colonne en gris est le vecteur propre correspondant à la composante F2 et la cellule en bleu est la valeur propre (ici égale à 1.45). Le nombre des variables étant égal à 10, la variance du nuage de points est de 10 (car il y a 10 variables de variance égale à 1 dans l'analyse) et la composante F2 explique donc 14,5% de la variance totale ($1,45 \cdot 100 / 10$).

	F1	F2	F3	F4	h ²
Variable 1	.766	-.244	.273	.215	.76
Variable 2	.559	-.432	.248	.019	.56
Variable 3	.177	.078	.640	-.565	.77
Variable 4	.327	-.144	-.610	-.525	.77
Variable 5	.712	-.404	-.114	.260	.75
Variable 6	.301	.613	-.136	.127	.50
Variable 7	.564	.151	-.422	-.446	.72
Variable 8	.352	.475	-.027	.163	.38
Variable 9	.483	.578	.120	.247	.64
Variable 10	.133	.245	.451	-.395	.45
Valeurs propres	2.32	1.45	1.37	1.16	6.30

(*) *Définition formelle.* En mathématique, la notion de valeur propre s'applique à des applications linéaires d'un espace vectoriel dans lui-même (endomorphisme). Un scalaire λ est une valeur propre d'une matrice carrée $U \times U$ s'il existe un vecteur x (appelé alors vecteur propre) non nul tel que $u(x) = \lambda x$.

2.3.4 Communautés

Dans une ACP, seul un nombre restreint de composantes est retenu pour l'interprétation. La communauté (h²), indique pour chaque variable la quantité de variance de la variable expliquée par les composantes retenues. Sachant que les variables dans l'analyse ont une variance de 1, la communauté multipliée par 100 correspond au pourcentage de variance de la variable expliquée par l'ensemble des composantes retenues.

Notes :

- La communauté d'une variable est la somme des carrés des [saturations](#) entre cette variable et chacune des composantes extraites (rappel : le carré de la saturation est la quantité de variance de la variable expliquée par la composante). Elle ne peut donc être négative, les valeurs possibles sont dans l'intervalle [0 ; 1].
- En ACP, si le nombre de composantes extraites (pris en compte) est égal au nombre des variables, la communauté est de 1 (l'ensemble des composantes explique 100% de la variance)
- L'idéal consiste à avoir des communautés globalement similaires les unes des autres (toutes les variables doivent être suffisamment expliquées par les composantes extraites). Si une ou plusieurs variables sont peu ou pas expliquées cela peut signifier que l'on n'a pas [extrait](#) assez de facteurs pour expliquer les variables ou que ces variables corrèlent avec aucune autre variable et [n'auraient pas du être intégrées dans l'analyse](#).

Exemple

Le tableau suivant reprend la table des saturations. Dans ce tableau, la colonne en gris correspond aux communautés observées pour chacune des variables si on extrait 4 composantes. La variable 5 est expliquée à 75% ($100 \cdot .75$). La variable la moins bien expliquée par ce système est la variable 8 (remarque, cette variable est celle qui corrèlait le moins bien avec les autres variables, cf. [la matrice des corrélations](#)).

	F1	F2	F3	F4	h ²
Variable 1	.766	-.244	.273	.215	.76
Variable 2	.559	-.432	.248	.019	.56
Variable 3	.177	.078	.640	-.565	.77
Variable 4	.327	-.144	-.610	-.525	.77
Variable 5	.712	-.404	-.114	.260	.75
Variable 6	.301	.613	-.136	.127	.50
Variable 7	.564	.151	-.422	-.446	.72
Variable 8	.352	.475	-.027	.163	.38
Variable 9	.483	.578	.120	.247	.64
Variable 10	.133	.245	.451	-.395	.45
Valeurs propres	2.32	1.45	1.37	1.16	6.30

On remarquera que la somme des communautés est égale à la somme des [valeurs propres](#). Cette valeur correspond à la quantité de variance expliquée par les facteurs extraits.

2.3.5 Nombre des composantes

Un des points les plus délicats de l'ACP (mais aussi pour l'analyse factorielle exploratoire) est de fixer le nombre de composantes à retenir dans l'analyse (on dit parfois le **nombre des facteurs/composantes à extraire**). Pour fixer ce nombre on doit apprécier la perte d'information induite par le fait que l'on réduit le nombre de dimensions. Par exemple si on a 15 variables, ne retenir que 4 facteurs supposent que le nuage de points dans cet espace à 4 dimensions n'est pas trop éloigné du nuage initial et que toutes les variables sont suffisamment "expliquées" par les facteurs/composantes. Il faudra prendre en compte (d'une façon ou d'une autre) :

- la qualité de représentation du nuage dans ce sous-espace factoriel (exprimé en pourcentage de variance expliqué) ;
- la qualité de la représentation qu'apporte chaque composante ([valeur propre](#));
- la qualité de la représentation de chacune des variables ([communauté](#)).

Règles pour définir le nombre des facteurs extraits

Soit la table des valeurs propres suivantes (indiquant pour chacun des 10 composantes, la valeur propre, le [pourcentage de variance expliqué](#) par le facteur (taux d'inertie) et le pourcentage cumulé de variance expliquée. Comment déterminer le nombre de composantes à retenir ?

Facteurs	Valeur propre	% Variance expliquée	% cumulé
F1	2,32	23,2%	23,2%
F2	1,45	14,5%	37,7%
F3	1,37	13,7%	51,4%
F4	1,17	11,7%	63,1%
F5	0,75	7,5%	70,6%
F6	0,62	6,2%	76,8%
F7	0,61	6,1%	82,9%
F8	0,59	5,9%	88,8%
F9	0,57	5,7%	94,5%
F10	0,55	5,5%	100,0%

Il n'existe pas une seule méthode mais plusieurs qui ne sont pas toujours convergentes et qui ne sont pas toutes recommandées. Les méthodes les plus utilisées sont :

- **le critère de Kayser (ou Kayser -Guttman)**

Ce critère simple est souvent évoqué (et utilisé) est imparfait et **ne devrait plus être utilisé**. On ne retient que les facteurs dont la [valeur propre](#) est supérieure à 1. Dans l'exemple précédent on ne retient que les 4 premières composantes. Cette méthode n'est pas une méthode recommandée.

- **Le test d'accumulation de variance ou scree-test (un autre nom est parfois donné "test du coude").**

Le scree-test (test d'accumulation de variance de Cattell, 1966) consiste à regarder comment évoluent les valeurs propres en fonction de leur ordre d'extraction. Le terme « scree » fait référence à l'accumulation de dépôts rocheux au pied d'une montagne créant ainsi un petit promontoire à l'endroit où le dénivelé de la montagne se transforme en une pente plus douce. On ne retient justement que les composantes qui précèdent le passage à cette pente douce.

Dans la figure suivante, représentant l'évolution des valeurs propres pour les composantes extraites (du premier au 10^{ème}), le changement de pente s'effectue avec la 5^{ème} composante, on devrait donc ne retenir que les 4 premières. Cette technique est souvent utilisée. Facile à mettre en œuvre elle devrait cependant être utilisée en complément d'autres techniques.

Remarque : le graphique des valeurs propres s'appelle aussi parfois en français "l'éboulis des valeurs propres".

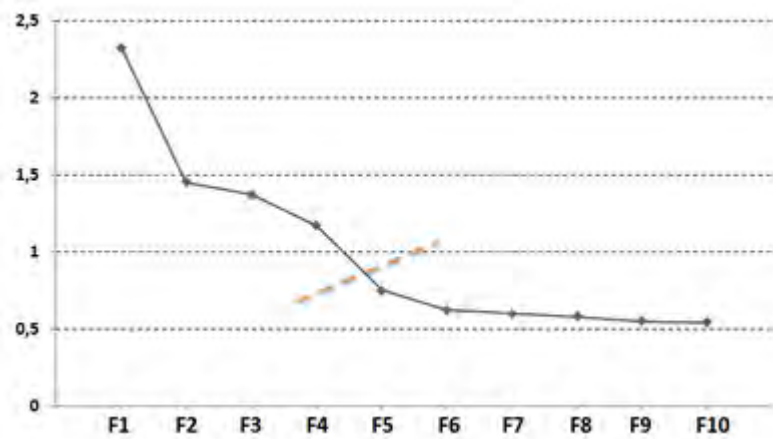


Figure H.1 : Scree test (évolution des valeurs propres pour les facteurs 1 à 10)

- **L'analyse parallèle**

Cette approche, en complément du test d'accumulation précédent a été proposée par [Horn](#) (1965). Cette méthode s'appuie sur le fait que, même en partant de données générées au hasard, il est possible d'observer une composante pouvant expliquer une proportion de variance supérieure à 1. L'analyse parallèle consiste donc à réaliser une série importante (1000 ou plus) d'ACP sur une matrice de corrélations générée au hasard mais comportant le même nombre de variables et le même nombre de participants que l'étude principale. La série des valeurs propres observée sur les données de l'étude est comparée à celle issue des valeurs propres calculées sur les données aléatoires (il existe plusieurs programmes, faciles à trouver sur le web, permettant de calculer ces valeurs). On ne conserve que les composantes dont la variance est significativement supérieure à celle obtenue avec la matrice de corrélations générée au hasard. La figure suivante illustre ce processus de décision. On ne retient que les 4 premiers facteurs. Cette technique fait partie des techniques recommandées.

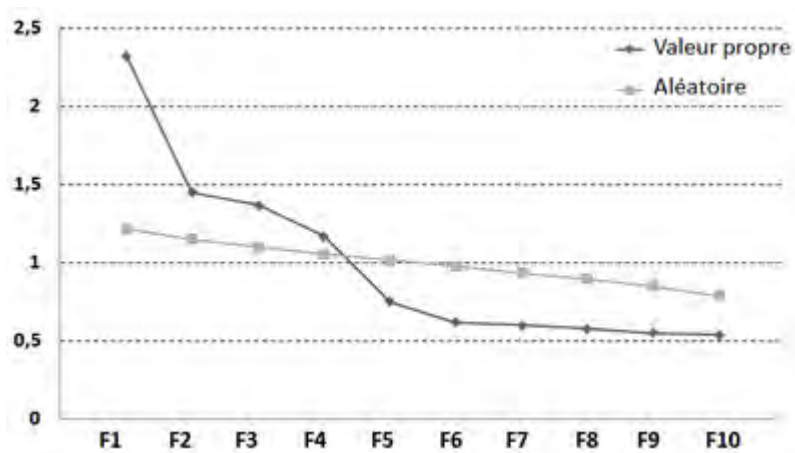


Figure H.2 : Évolution des valeurs propres et analyse parallèle

- **La qualité de représentation du nuage de points**

Très ambigu, ce critère consiste à retenir les facteurs de façon à expliquer au moins un certain pourcentage de variance. Selon la nature des mesures et de leur [fidélité](#) la valeur de ce critère peut varier. Ici, si on fixe le critère à 70% de variance expliquée, il faudrait retenir 5 composantes. Cette

méthode peut être utilisée en complément des méthodes précédentes et considérée comme un "regard" sur la qualité de la représentation retenus dans l'analyse. A elle seule, elle n'est pas recommandée pour déterminer le nombre de composantes à retenir.

- **Autres critères pouvant être utilisés mais beaucoup moins courants** (cf aussi Velicer, W., Eaton, C., & Fava, J., 2000)
 - **VSS (Very Simple Structure Critérium)**. Proposé par Revelle et Rocklin en 1979, le principe de cette méthode est de recalculer la matrice de corrélation initiale en ne gardant pour chaque variable, que la saturation la plus élevée (dans certaines variantes on garde les deux saturations les plus élevées), toutes les autres saturations étant fixées à 0. La valeur de VSS est un test d'ajustement de cette matrice recalculée à la matrice originale de corrélations (prend des valeurs entre 0 et 1). Cette valeur est calculée pour des solutions allant de 1 composante au nombre maximum de composantes. Elle tend vers une valeur optimale associée au nombre de facteurs à retenir. Peu utilisée, cette méthode convient peu pour des structures factorielles complexes.
 - **Comparative Data (CD)** introduite par Ruscio et Roche (2012) est une extension de la méthode parallèle de Horn. Elle consiste à prendre en compte la structure factorielle dans la génération de données aléatoires.
 - **Minimum Average Partial (MAP)**. Introduite par Vélicer en 1976, cette méthode consiste à recalculer la matrice des corrélations en retirant les k premières composantes (1, puis 2, etc.). Pour chaque valeur de k , on calcule moyenne des carrés des corrélations se trouvant en dehors de la diagonale. Cette moyenne va diminuer puis, à partir d'une certaine valeur de k , réaugmenter. Le nombre de composantes à retenir correspond à la valeur la plus basse observée.

Contrôle de la pertinence du nombre des facteurs sélectionnés.

- Le nombre des composantes retenues doit permettre d'expliquer globalement un pourcentage de variance suffisant (varie selon les domaines mais si on n'explique que 30% de la variance on peut s'inquiéter de la représentativité des composantes)
- La communauté correspond à la quantité de variance d'un test expliqué par les n premières composantes. Chaque communauté devrait avoir une valeur proche (plus ou moins) du pourcentage de variance cumulée expliqué par les composantes retenues (divisé par 100). Si une variable est clairement peu expliquée cela signifie soit que le nombre des composantes sélectionnées n'est pas suffisant, soit que cette variable corrèle peu avec les autres variables et devrait être exclue de l'analyse.

Pour ceux qui veulent aller plus loin

Ils existent de nombreux articles sur la façon de déterminer le nombre de facteurs. L'analyse parallèle de Horn est celle qui semble la plus appropriée (parmi les méthodes simples). Cependant, dans des simulations récentes, la méthode CD (comparative data) est préférable. Cette méthode est plus complexe à mettre en œuvre mais Ruscio (auteur de la méthode avec Roche) a déposé un script sous R permettant de déterminer ce nombre de facteurs (<http://ruscio.pages.tcnj.edu/quantitative-methods-program-code/>). Vous pouvez toujours, pour ceux qui connaissent R, en profiter pour voir comment on simule des données et la méthode utilisée.*

() R est un système d'analyse statistiques et un langage dérive de S. Il est distribué librement sous les*

termes de la GNU General Public Licence et est disponible pour plusieurs environnements (Windows, Linux, MacIntosh).

2.3.6 Rotation



La table initiale des saturations est souvent difficilement interprétable car les facteurs extraits répondent à une règle d'extraction simple : la première composante explique le plus de variance, la seconde est orthogonale à la première et explique le plus de variance restante, etc. Pour repérer les groupes de variables et donner un sens au système de composantes retenues, on effectue une rotation qui vise à rendre interprétable la table des saturations. On parle de rotation car il s'agit de faire tourner dans un espace vectoriel (les variables sont des vecteurs) les axes représentant les composantes(*).

L'objectif d'une rotation est toujours de simplifier la lecture de la table des saturation. Simplifier la lecture implique dans chaque rangée de la table de saturation que l'on trouve un maximum de saturation proche de 0 et des saturations en valeur absolue élevées. De nombreuses solutions sont possibles et le choix d'une rotation dépend des hypothèses de recherche. Pour simplifier il existe deux groupes de rotations : les rotations orthogonales et les rotations obliques.

- **Rotations orthogonales** : ces rotations maintiennent l'orthogonalité entre les composantes. On utilise ces rotations lorsque l'on suppose que les composantes sont indépendants les uns des autres.
- **Rotations obliques** : Les rotations obliques sont utilisées lorsque les composantes ne sont pas supposés indépendantes (il existe des corrélations entre elles). Dans le cadre strict de l'analyse en composantes principales (ACP), ces rotations ne devraient pas être utilisées, car elles contredisent l'hypothèse fondamentale d'orthogonalité des composantes.

VARIMAX - Une rotation orthogonale fréquemment utilisée

Cette rotation orthogonale permet d'obtenir une structure simple dans laquelle le nombre de variables corrélées avec un axe factoriel (composante) est maximisé. En effet, le but d'une rotation VARIMAX est de rechercher une structure simple : on fait tourner les axes de façon à augmenter le nombre de saturations fortes et faibles sur chacun des facteurs. Autrement dit, on recherche un système d'axe minimisant au maximum le nombre des saturations moyennes.

(*) En fait chaque variable est un vecteur ayant comme coordonnées les saturations observées sur chacune des composantes (cf. : [représentation graphique](#)). Ce système définit donc la base de cet espace vectoriel. Faire une rotation, revient, sans changer de position les variables dans cet espace, à rechercher une nouvelle base qui à la même origine mas des axes factoriels différents (plus facilement interprétables). En fait c'est comme si pour situer un objet dans une pièce, on prenait comme référence, non plus le coin droit de la pièce formé par les intersections des murs entre eux et du sol, mais à partir du même point, l'axe haut-bas, et l'axe nord-sur et l'axe est-ouest. Les coordonnées de l'objet changent mais il reste au même endroit.

Exemple de rotation VARIMAX

Table des saturations avant rotation

	F1	F2	F3	F4	h ²
Variable 1	.766	-.244	.273	.215	.76
Variable 2	.559	-.432	.248	.019	.56
Variable 3	.177	.078	.640	-.565	.77
Variable 4	.327	-.144	-.610	-.525	.77
Variable 5	.712	-.404	-.114	.260	.75
Variable 6	.301	.613	-.136	.127	.50
Variable 7	.564	.151	-.422	-.446	.72
Variable 8	.352	.475	-.027	.163	.38
Variable 9	.483	.578	.120	.247	.64
Variable 10	.133	.245	.451	-.395	.45
Valeurs propres	2.32	1.45	1.37	1.16	6.30

Table des saturations après rotation VARIMAX

	F'1	F'2	F'3	F'4	h ²
Variable 1	.845	.201	-.026	.117	.76
Variable 2	.724	-.109	.045	.155	.56
Variable 3	.099	-.064	.011	.869	.77
Variable 4	.064	-.095	.869	-.080	.77
Variable 5	.809	.080	.185	-.234	.75
Variable 6	-.089	.688	.144	-.044	.50
Variable 7	.160	.274	.779	.099	.72
Variable 8	.057	.608	.039	.025	.38
Variable 9	.171	.772	-.065	.116	.64
Variable 10	-.018	.117	-.001	.658	.45
Valeurs propres	1.97	1.61	1.42	1.30	6.30

On peut remarquer que cette rotation (comme toutes les rotations orthogonales) entraîne une redistribution de la variance expliquée par chaque facteur (les valeurs propres changent) mais la rotation ne modifie pas les communautés et donc la variance totale expliquée.

2.3.7 Représentation graphique

La représentation graphique consiste à représenter dans l'espace des facteurs (composantes), les variables. Elle est donc la "transcription" graphique partielle du [tableau des saturations](#). Le principe de cette représentation est simple mais les représentations graphiques permettent de visualiser au plus 2 ou 3 facteurs (pour les espaces à plus de 3 dimensions, la représentation graphique simple est impossible). Dans ce cas (le plus fréquent), pour interpréter les résultats, on utilise donc directement la table des saturations et très peu les représentations graphiques .

ILLUSTRATION SIMPLIFIEE

Soit une table de saturation simple avec deux facteurs et 4 variables (adapté de Reuchlin, 1976) :

Variables	F1	F2	h^2
Math.....	.70	.40	0,65
Sciences.....	.68	.26	0,53
Français.....	.56	-.58	0,65
Latin.....	.60	-.30	0,45
Valeur Propre	1.63	0.65	2.28
	(41%)	(16%)	(57%)

La représentation graphique va consister à représenter dans l'espace des facteurs (axes F1 et axes F2) les variables par des vecteurs ayant pour coordonnées les saturations.

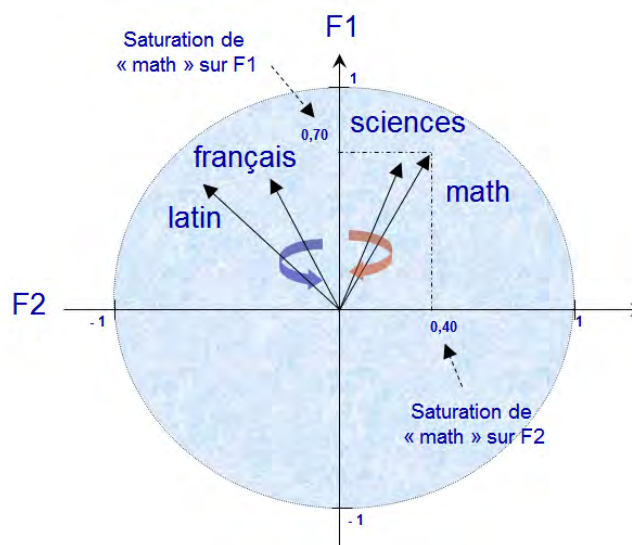


Figure H.8. Représentation graphique des variables dans l'espace des 2 premiers facteurs avant rotation.

Cette représentation permet de repérer qu'il y a deux groupes de vecteurs mais pour lire correctement le graphique il faut avoir compris que :

- Plus un vecteur est proche d'un axe, plus il est expliqué (associé) à la composante correspondant à cet axe.
- La longueur du vecteur (qui correspond à la norme du vecteur) est en relation avec la quantité de variance de la variable (vecteur) expliqué par les 2 composantes. Plus le vecteur est grand plus il est expliqué par les deux composantes. En fait le carré de la norme du vecteur est la quantité de variance expliquée par les facteurs (composantes) puisque c'est la somme des carrés des saturations (théorème de Pythagore tout simplement !)
- Cette norme (longueur du vecteur) ne peut dépasser 1 (une variable ne peut être expliquée à plus de 100%) et tous les vecteurs s'inscrivent dans un cercle de rayon 1 (cercle bleu sur le graphique)

Remarque : l'exemple précédent est une représentation avant [rotation](#). La rotation VARIMAX va consister à rechercher deux nouveaux axes orthogonaux qui passent au plus près (pour chacun d'eux) d'un des groupes de vecteurs.

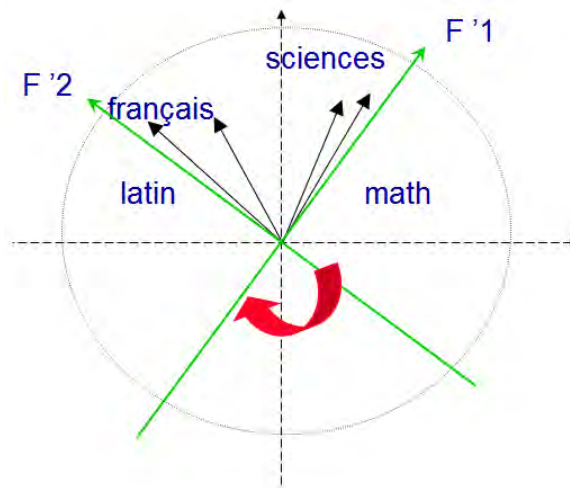


Figure H.4. Représentation graphique des variables dans l'espace des 2 premiers facteurs après rotation VARIMAX

Une rotation oblique aurait conduit à faire passer les axes au "centre" de chaque groupe de vecteur. Dans ce cas les axes n'étaient plus orthogonaux et les facteurs corrélaient entre eux. Cette solution aurait traduit le fait que toutes les variables après rotation VARIMAX se situent dans le même quadrant, et qu'il existe une corrélation entre toutes ces variables.

2.3.8 Interprétation des résultats

Interpréter les résultats consiste à donner un nom aux composantes extraites et identifier leur importance dans le système. Pour cela on examine la table des saturations (après [rotation](#) le plus souvent). On cherche pour chaque composante ce qui semble commun aux variables ayant les coefficients de saturation les plus élevés et qui n'est pas commun aux autres variables (on repère les groupes de variables associés à chaque facteur après rotation). S'il existe un pattern simple, la rotation permet, par les contrastes introduits entre les saturations, de faciliter l'interprétation.

Méthode.

1. La première étape est d'identifier pour chaque variable la saturation la plus élevée (en valeur absolue) significative. Il arrive que plusieurs saturations significatives élevées existent. Une variable qui a des saturations fortes (significatives) sur plusieurs composantes mérite parfois d'être exclue de la matrice (elle mesure plusieurs dimensions ou sous-dimensions). Ceci implique que l'analyse devra être exécutée de nouveau sans cette variable.
2. Une fois que les saturations ont été bien identifiées, l'interprétation (on parle d'étiquetage des facteurs ou des composantes) est réalisé à partir des variables qui ont une saturation significative pour chaque facteur. Pour identifier le facteur, on cherche ce qui est commun aux variables ayant des saturations élevées sur cette composante et qui les distingue des autres variables. L'interprétation s'appuie donc nécessairement sur une bonne connaissance des variables soumises à l'ACP. Le nom donné au facteur (composante) provient de ce que l'on sait sur chacune des variables.

Exemple

Table des saturations après rotation Varimax

	F'1	F'2	F'3	F'4	h ²
Variable 1	.845	.201	-.026	.117	.76
Variable 2	.724	-.109	.045	.155	.56
Variable 3	.099	-.064	.011	.869	.77
Variable 4	.064	-.095	.869	-.080	.77
Variable 5	.809	.080	.185	-.234	.75
Variable 6	-.089	.688	.144	-.044	.50
Variable 7	.160	.274	.779	.099	.72
Variable 8	.057	.608	.039	.025	.38
Variable 9	.171	.772	-.065	.116	.64
Variable 10	-.018	.117	-.001	.658	.45
Valeurs propres	1.97	1.61	1.42	1.30	

Après rotation, on observe que les sous-tests 1, 2 et 5 sont saturés par F'1, les sous-tests 6, 8 et 9 par F'2, les sous-tests 4 et 7 par F'3, et les sous-tests 3 et 10 par F'4. La connaissance des variables (qu'est-ce qui est commun à un groupe de variable et qui n'est pas commun aux autres variables) permettra d'interpréter les facteurs (= donner un nom à ces composantes).

2.4. AFE

L'analyse factorielle exploratoire (AFE) permet de mettre en évidence la structure latente (les facteurs) expliquant les corrélations entre les variables. Cette méthode fait l'hypothèse que les variables observées sont les résultantes de deux types de facteurs, ceux communs à plusieurs variables et ceux spécifiques à chacune des variables.

Il existe plusieurs méthodes d'extraction des facteurs mais l'objectif est toujours de maximiser la reproduction de la matrice de corrélations originale. Cette méthode postule que les variables observées sont des [combinaisons linéaires](#) de variables sous-jacentes que l'on appellent facteurs ou encore variables latentes (selon la méthode utilisée et le contexte de la recherche).

La matrice de variances-covariances soumise à l'AFE va avoir pour originalité que dans la diagonale de la matrice, on trouvera la communauté (variance des variables expliquée par les facteurs retenus) et non plus 1 (comme pour l'ACP). En effet on ne souhaite plus expliquer la totalité de la variance mais la variance qui est commune à plusieurs variables (au moins 2). Dans une AFE la quantité de variance à expliquer (trace de la matrice de variances-covariances) n'est donc plus égale au nombre des variables.

Pour aller plus loin...

Pour une introduction plus détaillée mais accessible vous pouvez consulter l'article de Watkins (2018) qui rappelle les règles d'usage de l'AFE mais aussi les principes essentiels.

2.4.1 Les étapes d'une AFE

Les sorties logiciels et l'interprétation sont similaires à celle de l'ACP. On peut cependant souligner que dans l'AFE les [rotations obliques](#) sont justifiées. Pour résumer, les étapes d'une AFE sont les suivantes :

1. Calculer la matrice de variances-covariances (cf. aussi "[matrice des corrélations](#)" dans le chapitre

ACP).

2. Sélectionnez le nombre de facteurs à extraire de façon à rendre compte des covariances avec le moins de facteurs possibles. Pour déterminer le nombre de facteurs, il existe de nombreuses méthodes qui sont similaires à celles utilisées en ACP (cf. "[Nombre des composantes](#)" dans le chapitre précédent.
3. Extraire les facteurs. Il existe plusieurs méthodes d'extraction (cf. [sous-chapitre suivant](#)). La meilleure méthode est généralement celle du Maximum de Vraisemblance.
4. Appliquer une rotation pour rendre le plus intelligible possible la solution trouvée. Il existe comme pour l'ACP de nombreuses rotations possibles, toujours classées en deux grandes catégories : les rotations orthogonales, qui produisent des facteurs non corrélés et les rotations obliques qui conduisent à des facteurs corrélés.
5. Interpréter la structure factorielle. Rappel : la force de la relation entre facteurs et variables est exprimée par le coefficient de saturation.

Remarque : L'analyse factorielle ne consiste pas simplement à identifier les variables latentes à l'origine des différences interindividuelles sur des variables observées. On peut aussi calculer, les scores factoriels des personnes (scores théoriques sur ces variables latentes), scores qui peuvent devenir de nouvelles variables pour des analyses ultérieures.

(a) Méthodes d'extraction en AFE

Les méthodes d'extraction sont des méthodes itératives de façon à reproduire au mieux la matrice de corrélation initiale. Ces méthodes sont multiples et sont souvent sources de difficultés (choix de méthodes) pour les utilisateurs de ces techniques. Les principales méthodes (les plus fréquentes) sont :

- **Les méthodes des moindres carrés (pondérées ou non).** Méthode d'extraction de facteur qui minimise la somme des carrés des différences entre la matrice de corrélations observée et celle reconstituée.
- **La méthode du maximum de vraisemblance (maximum likelihood estimation ou MLE dans les ouvrages anglo-saxons).** Méthode d'extraction de facteurs qui fournit les estimations de paramètres les plus susceptibles d'avoir généré la matrice de corrélations observée si l'échantillon est issu d'une distribution normale multivariée et le modèle est celui où chaque variable latente sature chaque variable observée. Un algorithme itératif est utilisé.
- **Méthode par factorisation en axes principaux (PFA).** Cette méthode cherche à maximiser les communautés. Puisque l'utilisateur ne connaît pas, par définition, la valeur des communautés avant d'avoir fait l'analyse, un exemple d'algorithme pour effectuer une AFE est :
 - (1) Remplacer la diagonale de la matrice par une estimation de la communauté de chaque test.
Le plus souvent, on reporte dans la diagonale la corrélation multiple entre la variable de cette colonne et les autres variables ;
 - (2) Extraire les facteurs de cette matrice
 - (3) On calcule les communautés (pourcentage de variance expliquée par ces facteurs) ;

- (4) Si ces communautés sont différentes des valeurs initiales (à un degré de précision prédéterminé = critère de convergence), on remplace dans la diagonale de la matrice des corrélations les valeurs des communautés qu'on avait estimées par ces nouvelles valeurs qui viennent d'être calculées. Puis on recommence les étapes 2,3, et 4.

Cette technique illustre qu'en AFE on ne cherche plus à expliquer la variance totale mais uniquement les facteurs communs qui expliquent les corrélations entre variables. Cette technique est automatiquement mise en œuvre par les logiciels.

A Savoir

L'analyse en composante principale (ACP et non AFE) est souvent proposée par défaut dans de nombreux logiciels ce qui entraîne des confusions entre AFE et ACP. L'ACP fut longtemps la solution préférée car la méthode d'extraction des composantes (méthode de calcul) était plus simple et demandait moins de ressources informatiques. Cependant, l'AFE doit être préférée dans le champs de la psychométrie, car cette méthode ne prend en compte que la variance partagée dans la solution factorielle et non la la variance spécifique (à un item ou une épreuve) ou l'erreur de mesure.

(b) Un exemple d'AFE



2.5. En résumé (à savoir)

Pour résumer, la pratique de l'analyse factorielle (ACP et AFE) comme l'analyse critique de résultats demande une expertise minimum. Avec Tabachnik et Fidell (2013) on peut résumer les points à vérifier systématiquement :

- Les variables présentes dans une analyse factorielle doivent avoir une sensibilité suffisante (doivent discriminer les positions des individus).
- Pour qu'une solution factorielle soit prise en considération stable, il faut un nombre suffisant d'observations. La règle veut qu'il y ait un minimum de 5 observations par variable (ce qui est vraiment un minimum).
- Les variables utilisées pour l'analyse devraient se distribuer normalement. Toutefois, on peut "transgresser" cette règle (en mode exploratoire) en utilisant des procédures d'extraction* qui prennent en compte les caractéristiques ces distributions. On peut aussi effectuer des transformations normalisant les distributions.

- La relation entre les variables est supposée linéaire.
- La matrice de corrélation ne doit pas être singulière (une variable ne peut pas être une combinaison linéaire d'une ou plusieurs autres variables). Lorsqu'une variable est trop fortement corrélée avec une ou plusieurs autres variables on peut avoir un problème de calcul de la solution factorielle (cas Heywood) avec des saturations qui deviennent supérieures à 1 (ce qui est théoriquement impossible) !
- Certains ensembles de variables doivent être corrélés entre eux (l'indice Kaiser-Meyer-Olkin [\[KMO\]](#) doit être suffisant et devrait être supérieur à .60).
- La solution factorielle doit expliquer une proportion suffisante de la variance (sinon la perte d'information est trop importante).
- Toutes les variables doivent faire partie de la solution factorielle (elles doivent avoir au moins une saturation supérieure à .20 ou .30 sur un des facteurs retenus dans l'AFE).
- Après rotation, un facteur doit saturer suffisamment (supérieure à .20 ou .30) plus d'une variable. On doit en général avoir au moins deux variables, sinon 3 qui ont des saturations suffisantes dans chaque facteur.
- Dans l'interprétation des données, on doit connaître (et prendre en compte) les caractéristiques des variables mais aussi celle de la population. L'analyse factorielle exploratoire reste une statistique descriptive.
- Une structure factorielle peut être différente pour différentes populations. Comme pour les corrélations (paradoxe de Simpson**), on ne doit pas regrouper dans une analyse des populations trop différentes.

Comparatif analyse en composantes principales et Analyse Factorielle Exploratoire

Aspect	ACP	AFE
Objectif	Résumer les données, maximiser la variance expliquée	Identifier des facteurs latents expliquant les corrélations
Variance prise en compte	Toute la variance (100%)	Uniquement la variance à expliquer par les facteurs.
Méthode de calcul	Décomposition en valeurs propres	Estimation par moindres carrés, maximum de vraisemblance, etc.
Résultat	Composantes principales	Facteurs + charges factorielles
Rotations	Rotations orthogonales	Rotations obliques ou orthogonales
Usage	Analyse exploratoire descriptive	Psychométrie, sciences sociales, tests, questionnaires

(*) La méthode du maximum de vraisemblance (ML pour maximum likelihood en anglais) est sensible aux déviations à la normalité des distributions. Pour des échelles ordinales (type likert) ou lorsque les distributions ne sont pas normales, on peut utiliser par exemple la méthode des moindres carrés non pondérés (ULS = Unweighted Least Square en anglais) qui minimise les résidus.

(**) Le paradoxe de Simpson est un paradoxe statistique décrit en 1951 par Edward Simpson (mais aussi par George U. Yule en 1903) dans lequel un résultat observé sur plusieurs groupes s'inverse lorsque les groupes sont combinés. Ce paradoxe est souvent rencontré en sciences sociales (et souvent oublié !). On trouve de nombreux exemples de ce paradoxe sur le web.

2.6. Usage - avertissements

L'analyse factorielle est un outil qui permet de condenser et de décrire des données. Utilisée dans un but purement exploratoire pendant de nombreuses années, des techniques plus récentes permettent de renouer avec les visées des premiers factoralistes qui étaient d'utiliser l'analyse factorielle (au sens générique du terme) pour tester ou éprouver des hypothèses structurales limitées. Ces nouvelles techniques (analyse confirmatoire et/ou de façon encore plus large la modélisation par équations structurales) viennent remplacer l'analyse factorielle (au sens général du terme) ou plus souvent viennent en complément de ces techniques lors de l'élaboration des tests et leur validation.

De façon générale, il faut aussi souligner que contrairement à une idée très répandue, ces statistiques ne sont pas des procédures « presse boutons » même si elles sont informatisées via des logiciels qui parfois prennent la place de l'utilisateur. Ces techniques n'apportent pas des réponses « automatiques » aux questions que l'on se pose et nécessitent tout au long de la démarche une activité et des décisions de la part de l'utilisateur, décisions qui ne relèvent pas uniquement de contraintes formelles (les résultats et l'interprétation vont dépendre de : l'échantillonnage des sujets, l'échantillonnage des variables, la technique d'analyse, AFE ou ACP, du nombre de facteurs à interpréter, du type de rotation, etc.). Par exemple :

(a) Nombre de sujets : l'utilisateur doit savoir que pour effectuer une analyse factorielle il faut un échantillon suffisamment important de façon à avoir de bonnes estimations des corrélations entre variables, c'est à dire des estimations les plus proches possibles des valeurs réelles des corrélations si celles-ci étaient effectuées sur toute la population de référence. Le nombre de variables que l'on peut soumettre à l'analyse factorielle détermine aussi la taille de l'échantillon. Il doit y avoir au moins 5 à 10 fois plus de sujets que de variables (Gorsuch, 1974).

(b) Choix des variables : Les facilités qu'apportent l'informatique conduisent souvent l'apprenti chercheur ou le chercheur à multiplier les variables en espérant en savoir d'autant plus. Ceci conduit le plus souvent à des structures de facteurs incompréhensibles, d'autant plus que l'interprétation relève aussi de notre connaissance de chaque variable. Comme pour toute méthode de recherche, des hypothèses explicites doivent guider le choix des variables et il y a peu de sens à faire des analyses factorielles sur n'importe quel ensemble d'observations.

(c) Interprétation des facteurs : les problèmes posés par l'interprétation des facteurs concernent d'une part, leur signification (activité d'interprétation) et, d'autre part le degré de généralité de ces facteurs. Du point de vue formel, on peut toujours trouver plusieurs systèmes de facteurs pour le même ensemble de données, et les facteurs ne sont donc que des entités mathématiques. Néanmoins, même si différentes condensations des mesures sont possibles, le psychologue ne pourra pas, dans le cadre d'une démarche scientifique, démontrer ou faire apparaître n'importe quel système de facteurs. Un certain nombre d'hypothèses doivent être à la base d'une démarche factorielle : choix des variables, choix de la population, choix de la technique d'analyse, choix du type de rotations éventuelles, etc. Ces précautions sont parfois oubliées.

Pour une présentation et une utilisation de l'analyse factorielle cf. l'article de [Beavers et col. en 2013](#).

3. Analyse factorielle des correspondances

L'analyse factorielle des correspondances (AFC, ou CA pour "Correspondence Analysis" en anglais)

introduite par Benzecri dans les années 1960 concerne le plus souvent le traitement des tableaux de données comme les tableaux de contingence. L'AFC est une ACP mais la métrique utilisée est celle du Chi2. Elle permet d'explorer la structure de variables cette fois catégorielles (et non plus uniquement des variables quantitatives comme pour les techniques précédentes). Cette technique est plus utilisée en sociologie qu'en psychologie et est par ailleurs souvent associée à des outils de classification.

Pour résumer de façon simplifiée, lors de l'analyse d'un tableau de contingence, une question typique est de savoir comment certains éléments lignes sont associés à certains éléments colonnes. Il faut savoir par ailleurs que cette méthode accorde une importance plus grande aux lignes de somme marginale élevée (lié à la métrique du CHI2). Il existe des techniques permettant, si nécessaire, d'équilibrer la contribution de chaque ligne.

Cette technique est peu utilisée en psychologie (sinon dans le domaine de la psychologie de la santé) et encore moins dans le champs de la psychométrie. Elle n'est donc pas présentée dans ce manuel.

4. Analyse factorielle confirmatoire

Dans le prolongement de l'AFE, l'analyse factorielle confirmatoire s'intéresse aux variables latentes qui sous tendent l'organisation des différences inter-individuelles observées sur un ensemble d'épreuve. Ce qu'elle apporte en plus de l'analyse factorielle exploratoire est qu'elle permet surtout de tester l'adéquation des données à un modèle théorique (démarche hypothético-déductive). Les hypothèses concernant les variables latentes sont formulées a priori.

4.1. Principe général

L'analyse factorielle confirmatoire est un cas particulier de la modélisation par équations structurales (Structural Equation Modelling [SEM]). Dans ce type d'approche, on fixe a priori un modèle qui va préciser le nombre de facteurs, les relations éventuelles entre ces facteurs, les relations entre ces facteurs et les variables observées, les termes d'erreurs attachés à chaque variable observée et les corrélations éventuelles entre eux. La figure suivante (Hx) présente un exemple de modèles (à trois facteurs) et 9 variables manifestes. Les facteurs ε_1 à ε_9 sont des facteurs spécifiques, non corrélés entre eux.

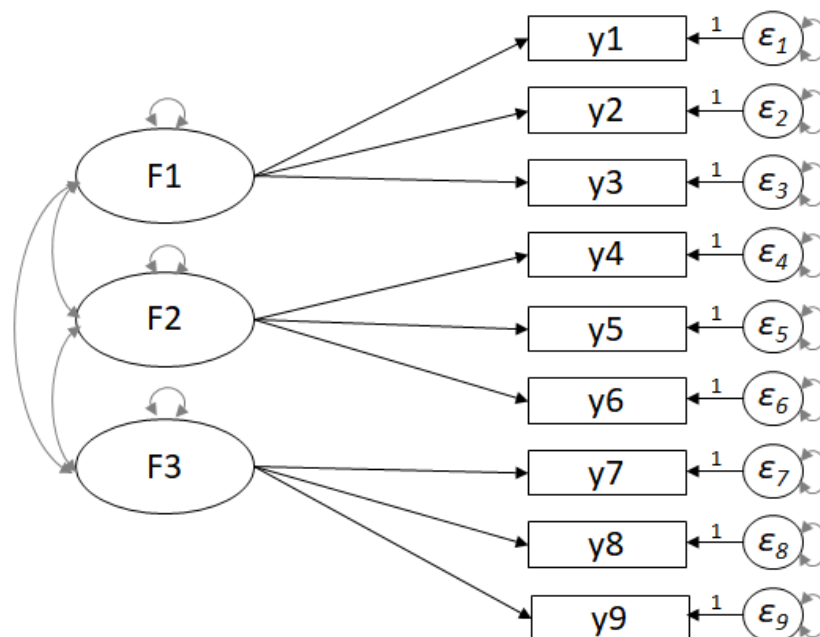


figure H.9 : Exemple de modèle à 3 facteurs.

(Les facteurs (F1 à F3) sont représentés par des ellipses ou pour les facteurs spécifiques par des petits cercles. Les variables manifestes (y_1 à y_9) sont représentées par des rectangles. Les double flèches traduisent les relations entre facteurs (covariances). Enfin, les variances des facteurs comme des facteurs spécifiques sont représentés par des doubles flèches courbes sur les facteurs).

Démarche générale pour mettre en œuvre une analyse factorielle confirmatoire :

La démarche générale (présenter très succinctement ici) consiste à spécifier un modèle (comme celui donné ci-dessus) en introduisant éventuellement en plus de la configuration générale des contraintes sur l'égalité de certains paramètres du modèle. On vérifie ensuite que le modèle est identifié (*i.e.* le nombre de paramètres à estimer est bien inférieur ou égal au nombre de variances et covariances de la matrice de données). On estime ensuite les paramètres (par exemple par la méthode du maximum de vraisemblance). Il existe plusieurs types d'estimateur que nous ne présentons pas ici.

4.2. Un bon modèle ?

Les indicateurs que l'on peut utiliser pour s'assurer que le modèle est un "bon modèle". Les plus utilisés sont :

⇒ les indices d'ajustement (goodness of fit)

- A priori, le premier indicateur à prendre en compte est le χ^2 qui permet de calculer l'écart entre la matrice de covariance observée et la matrice de covariance estimée. Cet écart doit être minimum (χ^2 non significatif) mais plus le nombre des observations est important plus on risque de rejeter à tort le modèle.
- Le GFI (« Goodness of Fit Index ») et l'AGFI (Adjusted GFI). Le GFI est un indicateur de la part relative de la variance-covariance expliquée par le modèle. Cet indicateur varie théoriquement entre 0 et 1 et devrait être supérieur à .90. L'AGFI est le GFI version ajustée du GFI qui prend en compte la complexité du modèle. Il pénalise les modèles avec beaucoup de paramètres à estimer.
- Le SRMR standardisé (Standardized Root Mean Residual). Il mesure la différence moyenne entre les covariances observées et les covariances prédites par le modèle (une fois standardisées). Donc il nous renseigne sur la façon dont le modèle reproduit les corrélations observées. C'est la racine carrée de la moyenne de la somme des carrés des résidus de chaque cellule de la matrice. Sa valeur doit être inférieure à .05 (.10 pour certains). Plus il est petit mieux c'est.

⇒ Indice de non centralité

- Le RMSEA (Root Mean Square Error of Approximation). Indicateur de la qualité d'ajustement du modèle. On considère qu'une valeur égale ou inférieure à .05 est le gage d'un bon ajustement. Il est souvent considéré comme acceptable à partir de 0.08. Plus il est faible meilleur est l'ajustement.

⇒ Indices incrémentaux (ces indices évaluent ce qu'apporte le modèle par rapport à un modèle de base pris en référence.)

- CFI (Bentler comparative fit index). Compare le modèle étudié au cas d'indépendance entre variables manifestes. Il mesure donc l'amélioration de l'ajustement du modèle par rapport à

un modèle où toutes les variables sont non corrélées. Il doit être supérieur à .90.

- Le TLI (Tucker-Lewis Index) : compare un modèle au modèle où toutes les variables sont supposées indépendantes (comme le CFI), mais pénalise les modèles trop complexes (c'est là sa différence avec le CFI). Il doit être supérieur à .90.

⇒ Indice de parcimonie (permet de comparer aussi des modèles différents)

- AIC (Akaike Information Critérium). Ce critère est utilisé lors de la comparaison de modèles (en AFC mais plus généralement en modélisation par équations structurelles). On doit privilégier le modèle donc l'AIC est le plus petit. En effet ce critère prend en compte non seulement la qualité de l'ajustement mais aussi la complexité du modèle (il pénalise les modèles ayant un grand nombre de paramètres). Attention : l'AIC ne donne pas d'information sur la qualité du modèle. Il sert uniquement à comparer des modèles entre eux

4.3. Pour conclure

Ce que permet l'analyse factorielle confirmatoire :

- La démarche, contrairement à l'AFE, n'est plus de rechercher les variables latentes hypothétiques sources de l'organisation des différences interindividuelles observées mais de tester un modèle particulier.
- L'analyse factorielle confirmatoire permet aussi de comparer différents modèles et de rechercher celui qui rend le mieux compte des données observées.
- Permet d'évaluer de tester la compatibilité des données avec des structures plus complexes (modèles factoriels à plusieurs niveau de deuxième voir de troisième ordre ou plus, modèle bifactoriel, etc.).

A savoir cependant :

- Il est plus difficile de maîtriser cette technique que l'AFE.
- Ce type d'analyse nécessite souvent des échantillons plus importants (c'est une technique inférentielle et non exploratoire).
- En pratique, on combine souvent analyse exploratoire et confirmatoire. L'analyse confirmatoire permet de valider la version finale d'un test ou d'un questionnaire.

I - Brèves sur des auteurs

ANASTASI, Anne (1908-2001). Psychologue américaine, Anne Anastasi est née en 1908 à New York. Précoce, elle entre au Collège Barnard à l'âge de 15 ans puis intègre l'université de Columbia où elle obtient son doctorat en deux ans (à l'âge de 21 ans). Les temps économiques difficiles ont marqué sa carrière académique (crise de 29) et elle débute comme instructrice au Collège Barnard. En 1939, elle occupe le poste de présidente du Département de psychologie du Queens College, où elle est restée jusqu'en 1947. Elle intègre alors l'Université Fordham où elle demeure jusqu'à son départ à la retraite en 1979. Anne Anastasi a reçu de nombreux prix dont la médaille d'or de la Fondation américaine de psychologie et le prix de l'Association américaine pour la recherche en éducation. Elle a été présidente de l'APA en 1972, première femme en plus de 50 ans à le faire. Trois de ses livres sont devenus des classiques dans le domaine de la psychologie différentielle. Son livre "Psychological Testing", plusieurs fois réédité (7 rééditions augmentées !), est mondialement reconnu comme l'un des textes de psychologie les plus importants du vingtième siècle et il reste encore un classique de la psychométrie. Très grande dame de la psychologie, elle cherchait toujours à s'exprimer clairement et simplement ("*when I wrote I was going to take tough things and make them simple*"). Figure de la psychométrie, elle décède à 92 ans (source : <http://www.apa.org/about/governance/president/bio-anne-anastasi.aspx>)

BAYES, Thomas Bayes (1702-1761). Mathématicien britannique et pasteur de l'Église presbytérienne, il est membre de la Royal Society. Paradoxalement il ne publie aucun article mathématique. Ses découvertes en probabilités ont cependant été publiées à titre posthume par un de ses amis (R. Price). On lui doit ainsi une loi importante des probabilités (formule de Bayes). Ce théorème est fondamental dans le développement d'un courant de la statistique moderne (statistique Bayésienne) qui se développe dans de nombreux domaines.

BINET, Alfred (1857-1911) : Alfred Binet est dans un premier temps avocat au barreau de Paris. Il démissionne après 6 ans de pratique et commence des études de médecine (inachevées) puis suit des cours de psychophysiologie et de clinique psychiatrique et enfin des études de sciences naturelles à la Sorbonne. Théodule Ribot (professeur au collège de France) le pousse à poursuivre en psychologie et il travaille à l'Hôpital de la Salpêtrière. Il devient, suite à différentes rencontres, directeur adjoint (1892) du Laboratoire de psychologie physiologique de la Sorbonne puis directeur de ce laboratoire (1895). Il fonde avec Henri Beaunis la revue L'Année psychologique en 1894. A la demande du gouvernement Français, il travaille avec T. Simon sur l'évaluation des enfants et publie la première échelle métrique de l'intelligence (élaborée conjointement avec Théodore Simon). Ce psychologue et pédagogue, est actuellement surtout connu pour avoir introduit cette première échelle moderne d'intelligence. Il a l'intuition qu'une des solutions est d'évaluer l'intelligence par les "connaissances banales" acquises et typiques d'un âge de développement. Cette idée que beaucoup de chercheurs qualifieront de géniale le conduit à introduire la notion d'âge mental. Il n'est cependant pas l'auteur de la notion de QI qui sera introduite en 1912 par W. Stern.

BENZECRI, Jean-Paul (1932-). Statisticien, né à Oran, ancien élève de l'École normale supérieure, il effectue une thèse sous la direction de Carton en 1955 (thèse en topologie). Il est le fondateur de l'école française d'analyse des données et est connu surtout pour le développement de

l'Analyse Factorielle des Correspondances. Enseignant chercheur à Rennes, il fut ensuite professeur à la Faculté des sciences de Paris puis à l'Université Pierre-et-Marie-Curie. En relation dans sa jeunesse avec le groupe MultiDimensional Scaling (aux Etats-Unis), ses travaux recevent un accueil "peu enthousiaste" dans le monde anglo-saxon suite, semble-t-il, à une visite en 1965 aux laboratoires de la *Bell Telephone*. Les développements de l'AFC par Benzecri sont partiellement oubliés et il semble faire l'objet d'un certain ostracisme. En réaction, on décrit un isolement et un travail réflexif à l'origine d'un ouvrage important sur l'histoire et la préhistoire de l'analyse des données (Benzecri, 1982).

CATTELL, James McKeen (1860 - 1944). Premier professeur en psychologie des États-Unis (université de Pennsylvanie) et éditeur de journaux scientifiques. Il effectue son PhD avec Wundt (en Allemagne), avec qui il collabore aussi pour élaborer des méthodes d'études scientifiques de l'intelligence (sa thèse a pour titre "Psychometrische Untersuchungen", soit en français "investigation psychométrique"). Après un séjour en Angleterre à Cambridge il prend en 1889 un poste de professeur aux Etats Unis en Pensylvanie puis en 1891 à l'université de Columbia. Il devient aussi président de l'American Psychological Association (APA) en 1895 et participe à la fondation de différents journaux (comme *Psychological Review*). C'est Cattell qui contribua à ce que la psychologie devienne une discipline scientifique légitime. Au moment de sa mort, le *New York Times* lui rendit hommage. On considère qu'il est l'inventeur du terme "*test mental*". Ses intérêts l'on porté vers les différences interindividuelles mais aussi vers l'analyse des temps de réaction, la psychophysique et la psychométrie. Parmi ses exemples de travaux où il fut un des pionniers, on peut citer ceux sur la mémoire et le témoignage.

CATTELL, Raymond (1905-1998). A ne pas confondre avec James McKeen Cattell. Né à Londres, après des études de chimie puis de psychologie, il travaille d'abord avec Spearman. Il rejoint Thorndike puis Allford aux États-Unis avant de diriger un laboratoire de recherche dans l'Illinois. Auteur important et influent du 20^{ème} siècle en psychologie, il est connu pour ses travaux sur l'intelligence et la personnalité. Concernant l'intelligence, il théorise et décrit avec Horn, deux principales formes d'intelligence (intelligence cristallisée et intelligence fluide). Dans le domaine de la personnalité, il est à l'origine du 16 PF, la personnalité étant décrite à partir de 16 facteurs. En 1997, il refuse la médaille du Lifetime Achievement de l'APA suite aux réserves formulées à son encontre à propos de propos eugéniques qu'il aurait tenu (propos qu'il dément, cf. la [lettre ouverte](#) qu'il publie).

CRONBACH Lee Joseph (1916-2001). Psychologue américain, il obtient un doctorat de psychologie de l'éducation en 1940. Il est instituteur puis occupe des emplois dans différentes universités : State College of Washington, l'université de Chicago, l'université de l'Illinois. Il terminera sa carrière à l'université Stanford où il est professeur à partir de 1964. Il a été président de l'Association américaine de psychologie (1957) et président de l'American Educational Research Association (1964-1965). Cronbach est surtout connu pour ses travaux sur la fidélité et sa contribution à l'amélioration de la modélisation psychométrique et au développement de la théorie de la généralisabilité. Une des mesures de la fidélité porte son nom : le coefficient alpha de Cronbach. On oublie souvent que ses travaux ne se limitent pas à cet aspect. Il s'est interrogé sur la mesure des variables décrivant les interactions pédagogiques, la nature du processus d'apprentissage (enseignement), l'évaluation des programmes éducatifs.

EBBINGHAUS, Hermann (1850-1909). Psychologue associationniste allemand il est considéré comme un

des pères de la psychologie expérimentale de l'apprentissage et des travaux sur la mémoire. Il est en effet le premier à avoir mis en place des paradigmes expérimentaux pour l'étude de la mémoire et l'apprentissage de liste de mots ou de syllabes sans signification basés sur l'auto-observation (met par exemple en évidence l'effet de récence et de primauté).

ESQUIROL, Jean-Étienne Dominique (1745-1826) : psychiatre français mais aussi homme ayant eu une influence importante dans la mise en place du secteur psychiatrique en France. Successeur de Pinel à la Salle Pétrière, il fit voter en 1838 la loi obligeant chaque département français à se doter d'un hôpital spécialisé (CHS dans la terminologie actuelle). Il enrichit la nosologie de Pinel en développant les concepts de monomanie et d'hallucination.

GALTON, Francis (1822-1911) : Homme de science britannique avant d'être psychologue, il fut anthropologue, explorateur (inventeur du sac de couchage), géographe, inventeur, météorologue, psychométricien et statisticien. Il fait partie des pères fondateurs de la psychologie différentielle. Cousin de Charles Darwin (et très influencé par la théorie de l'évolution de ce dernier), fortuné, il a été avant tout un touche-à-tout intuitif (il est à l'origine par exemple du mot « anticyclone », il découvre les ultrasons, travaille sur la surimpression en photographie, etc.).

Ses travaux en psychologie sont connues pour deux raisons. La première est l'importance qu'il va attribuer à l'étude de la transmission héréditaire des caractères. La seconde est le rôle qu'il a joué dans le développement des statistiques et de la psychométrie (introduit ou développe les notions d'étalonnage, de régression, de corrélation, initie l'analyse factorielle, etc.). Avec son disciple Karl Pearson, il fonde un journal (*Biometrika*) et le premier laboratoire d'anthropométrie (école biométrique). Il sera aussi à l'origine du mot et du courant eugénique (minimisant trop le rôle du milieu dans les caractères psychologiques).

GAUSS Carl Friedrich (1777-1855) : un des plus grands mathématiciens de tous les temps (dit le prince des mathématiciens) mais aussi astronome (calcul de la trajectoire de comètes) et physicien (optique, magnétisme, géodésie). Ses travaux touchent à de très nombreux domaines : il développe par exemple la méthode des moindres carrés de Legendre, envisage la possibilité de géométries non euclidiennes, travail sur la convergence des séries, explore des conjectures sur les nombres premiers, découvre le moyen de dessiner un polygone à 17 cotés avec un compas et une règle, etc. Si ces travaux concernent toutes les branches des mathématiques, les résultats les plus remarquables sont cependant obtenus en théorie des nombres et en géométrie. Son nom est souvent associé à des théorèmes ou des fonctions (exemple le plus connus étant les fonctions gaussiennes et la courbe de Gauss). Il cesse de travailler professionnellement en 1840 et se consacre au magnétisme terrestre jusqu'à la fin de sa vie.

GUTTMAN LOUIS (1916-1987). Docteur en sociologie de l'Université du Minnesota, Louis Guttman est un des psychométriciens les plus influents du 20^{ème} siècle (sa thèse portait sur l'analyse factorielle et ses développements algébriques). Il enseigne dans diverses universités américaines (Cornell, Harvard et Ann Arbor). En 1954, il est nommé professeur à l'université hébraïque de Jérusalem (*professor of social and psychological assesment*). Il fut président de la Psychometric Society. Il est connu pour le développement de l'analyse d'échelle dans la mesure des opinions et des attitudes (à l'origine de ce qu'on appelle les échelles de Guttman). Il est aussi à la source de nombreux travaux importants comme ceux concernant l'intelligence et l'analyse de matrices de corrélations (cf. Radex de Guttman). Il publie en psychologie dans

des revues comme Psychometrika et ces articles sont encore souvent cités en statistiques. En 1971, Guttman, est présenté par la revue Science comme un des 62 chercheurs en sciences sociales les plus influents du début du XX^{ème} siècle.

HORN John Leonard (1928-2006). S'engage comme militaire sans avoir terminé le lycée. Il obtient un GED (Graduate Equivalency Degree) et accède à l'université (Denver) et entreprend des études de chimie et psychologie. Il effectue son master en Australie et rencontre R.B. Cattell, avec qui il décide de réaliser une thèse qui sera consacrée à l'Intelligence fluide et cristallisée (première étude empirique dans ce domaine). Chargé de cours à Berkeley (Californie) il devient professeur à l'université de Denver en 1970. Fortement impliqué dans les mouvements afro-américain (concernant les droits civiques) il encourage fortement ses étudiants à la pensée critique. Esprit critique et positif, il sera un des moteurs du développement de la théorie gf-gc. Il recevra plusieurs prix scientifiques (Research Career Development Award ; Annual Prize for Distinguished Publications in Multivariate Psychology ; Lifetime Achievement Award). Horn sera aussi président de la NAACP (National Association for the Advancement of Colored People) et de l'ACLU (American Civil Liberties Union). [source : McArdle, 2007].

KRAEPELIN, Emil (1856 -1926) : psychiatre allemand considéré comme le fondateur de la psychiatrie scientifique moderne. Élève de Wilhelm Wundt (fondateur de la psychologie expérimentale), il a initié une classification des maladies mentales fondée sur des critères cliniques objectifs.

LIKERT, Rensis (1903 - 1981) : psychologue américain il est connu pour avoir donné son nom aux [échelles de Likert](#). Après un Bachelor en Sociologie, il obtient un doctorat en 1932. C'est au cours de ce doctorat qu'il a conçu une échelle d'enquête pour mesurer des attitudes. Il fut l'un des fondateurs du "Michigan Institute for Social Research" et a consacré la plupart de ses travaux à la recherche sur les organisations et le management.

OTIS Arthur Sinton (1886-1964). Élève de L. Termann (Université de Stanford) Arthur Otis est un des pionniers dans l'élaboration des tests d'intelligence. Il développa les premiers tests collectifs à choix multiple pour l'Armée américaine (test alpha et bêta). Ils seront administrés à 1,7 millions de recrues de l'armée. Après la guerre, il revient brièvement à Stanford et rejoint la World Book Company de Yonkers en 1921 (poste qu'il occupe pendant 25 ans). Il contribue à fixer les normes des développeurs de test et apporte par ailleurs de nombreuses contributions à la méthode statistique. Pendant la Seconde Guerre mondiale, il est consultant psychologique auprès du Bureau de l'aéronautique de la Marine. Arthur Otis a écrit sur la géométrie, la comptabilité, l'arithmétique, l'aéronautique, la congestion routière, la fiscalité et les habitudes de vote. Il était membre de nombreuses sociétés savantes (Association américaine pour l'avancement des sciences, 'American Psychological Association, 'Académie des sciences de New York, 'American Educational Research Association, 'Académie nationale d'économie et de sciences politiques, Académie d'économie mondiale). Il est cependant surtout connu en dehors de la psychologie pour son ouvrage controversé sur la théorie de la relativité (Light Velocity and Relativity, 1963) écrit juste avant sa mort.

RORSCHACH, Hermann (1884-1922) : Psychiatre Suisse, membre de la société suisse de psychanalyse. Sa thèse dirigée par Eugen Bleuler porte sur les hallucinations. Il est essentiellement connu pour l'élaboration d'un test projectif (avec Konrad Gehrung) voulant évaluer les caractéristiques d'une personne à partir de ces réactions à des tâches d'encre (ce test serait inspiré d'un recueil de poèmes inspirés par des taches d'encre mais aussi d'un jeu d'enfant dit klecksographie qui

consiste à déposer une goutte d'encre sur une feuille et de la plier pour obtenir des tâches représentant des formes interprétables).

REUHLIN, Maurice (1920 - 2015) : Né à Marseille en 1920, ce psychologue français a été élève de l'Ecole normale d'instituteurs d'Aix, puis instituteur. Après une formation de conseiller d'orientation il poursuit ses études Doctorat (1962). Enseignant et chercheur au CNRS il est un des psychologues différentialistes qui marque la seconde moitié du 20ème (occupe le premier poste de professeur de psychologie différentielle créé en 1968 à Paris V). Il s'est aussi intéressé à l'histoire de la psychologie. Il est connu pour l'introduction du concept de vicariance. Un de ces manuels (Psychologie) a été pendant très longtemps un des classiques incontournables pour tous les étudiants de psychologie (en langue française).

SEGUIN, Edouard (1812, 1880). Pédagogue Français, il sera surnommé « l'instituteur des idiots ». Jeune médecin, il s'intéresse aux maladies mentales et travaille avec Jean-Marc Itard et Jean Étienne Esquirol. Il est le premier à ouvrir en France (en 1840) une école destinée aux enfants ayant un retard mental. Il publia en 1846 "*Traitement moral, hygiène et éducation des idiots*" qui devint un ouvrage classique en psychologie. Il émigre aux États-Unis en 1852 et il ouvrira plusieurs institutions pour enfants avec retard mental. Il utilisera tardivement des formes encastrables pour entraîner les enfants déficients sur le plan cognitif, formes reprises dans des tests dont celui le Wechsler-Bellevue pour mesurer l'intelligence.

SPEARMAN, Charles Edward (1863-1945). Psychologue anglais, élève de Wundt, connu à la fois pour ses travaux sur l'intelligence (postule l'existence d'un facteur générale d'intelligence sous-tendant toutes les activités intellectuelles) et pour les développements statistiques qu'il proposa (analyse factorielle et corrélation). Il débute ses recherches tardivement (commence son PhD à 34 ans). Il est considéré comme le père de la théorie classique des tests (Boake, 2002) et des indices dérivés comme la fidélité. Il reste un des chercheurs les plus connus (avec Cattell, Binet, Galton) dans le domaine de la mesure de l'intelligence.

STERN, William (1871 -1938). Psychologue allemand, c'est lui qui est à l'initiative du QI classique (en 1912, soit un an après la mort de Binet). Il fait ses études à Berlin et il aura Ebbinghaus comme enseignant. En 1911, il introduit le terme de psychologie différentielle (à la place de psychologie individuelle) et publie le premier ouvrage allemand de psychologie différentielle. Professeur à Hambourg, mais évincé par le régime d'Hitler en 1933, il émigre au Pays Bas puis aux États-Unis (Université de Duke).

STEVENS, Stanley Smith (1906-1973), psychologue américain, théoricien de la mesure, il est le fondateur du laboratoire de psychoacoustique d'Harvard. Il s'intéresse et développe particulièrement à une branche de la psychophysique moderne qui concerne la mesure directe des sensations. Il est à l'origine de nombreux travaux dans son domaine et donna son nom par exemple à une loi de puissance reliant la grandeur physique d'un stimulus et l'intensité perçue, dite loi de Stevens. Il a joué aussi un rôle essentiel dans le développement de définitions opérationnelles comme celle d'échelle de mesure. En 2002, un article de "Review of General Psychology survey" classe Stevens comme le 52ème auteur de psychologie du 20ème siècle le plus cité.

SIMON, Théodore (1873-1961). Médecin aliéniste français, fondateur de la première école d'infirmière en psychiatrie à Maison-Blanche (Neuilly sur Marne), il est surtout connu pour sa contribution à la première échelle d'évaluation de l'intelligence (avec Binet), échelle qui porte les deux

noms : échelle d'intelligence de Binet-Simon.

THURSTONE Louis Leon (1887 - 1955). Psychologue représentant de l'école Américaine de psychologie, il commence sa carrière comme ingénieur électricien. Il invente un projecteur cinématographique et est brièvement assistant de Thomas Edison. Ces travaux l'orientent vers la mesure des sensations et la psychologie. Il obtient son PhD à Chicago en 1917 et consacre le reste de sa carrière à la psychologie. En opposition à Spearman, il développe des nouvelles techniques d'analyse et isole des aptitudes mentales primaires (compréhension verbale, fluidité verbale, raisonnement, etc.). Il est connu essentiellement pour ses travaux sur l'analyse factorielle appliquée à l'étude de l'intelligence mais il développe aussi des recherches reconnues en psychophysique et en psychologie sociale.

WECHSLER, David (1896-1981). Psychologue américain d'origine roumaine (né en Roumanie, il émigre aux États-Unis à l'âge de 6 ans). Il est à l'origine des tests d'intelligence normalisés les plus utilisés (les échelles de Wechsler). Il termine sa formation à Londres dans les années 20, période au cours de laquelle il travaille avec le psychologue Charles Spearman et le mathématicien Karl Pearson. Psychologue à partir de 1932 à l'hôpital psychiatrique de Bellevue à New-York, il développe une première batterie pour mesurer l'intelligence des adultes. Il rejette la notion d'âge mental ou d'âge de référence à la base du QI. Il introduit la notion de QI standard (somme de scores normalisés) pour les adultes comme pour les enfants. Il choisira comme norme une moyenne de 100 et d'un écart type de 15 pour obtenir des résultats proches de ceux obtenus avec les tests de QI classique. Ces échelles étaient composites (tâches variées) car il estimait que l'intelligence était une capacité globale composite : *"Les attributs et les facteurs de l'intelligence, comme les particules élémentaires en physique, ont à la fois des caractéristiques individuelles et collectives qui semblent se comporter différemment lorsqu'ils sont isolés et lorsqu'ils opèrent de concert"* (Wechsler, 1975).

WUNDT, Wilhem (1832-1920). Après des études de médecine, il devient en 1857 l'assistant de Hermann Van Helmholtz. Il est connu surtout pour ses travaux sur la perception et fait partie des premiers chercheurs qui pensent que les phénomènes mentaux peuvent être l'objet de sciences. La méthode expérimentale doit permettre d'isoler et de mesurer des phénomènes complexes. Imprégné d'associationnisme, il crée le premier laboratoire de psychologie expérimentale (en 1879) à Leipzig. C'est dans ce laboratoire que viendront se former de nombreux chercheurs à la méthode expérimentale. On peut citer C. Spearman (réalise sa thèse avec Wundt), J.McK Cattell (réalise sa thèse avec Wundt), Titchener (réalise sa thèse avec Wundt), Stanley Hall, etc.

YERKES, Robert (1876-1956). Psychologue et primatologue américain, il fait ses études à Harvard et obtient son doctorat en 1902. Pionnier de la psychologie comparée, il fonde un laboratoire de biologie-primatologie dont il sera directeur entre 1929 et 1941 (en Floride). Pendant la première guerre mondiale, il développe avec d'autres psychologues le test Alpha (pour ceux qui savent lire) et le test Beta (figural pour ceux qui ne savent pas lire). Ces tests servaient à la sélection des officiers ou des soldats (plus de 1,75 millions de passations). Premiers tests collectifs, ces deux premières versions de tests mentaux vont contribuer à une expansion des tests d'intelligence et vont contribuer au développement des tests à choix multiples. Dans la continuité de ces travaux, un des collègues de Yerkes est à l'origine du Scholastic Aptitude Test dont la première version date de 1926. La distinction entre les tests Alpha et Beta va inspirer aussi Wechsler dans la construction de sa première échelle d'intelligence. Intéressé par

l'intelligence et l'apprentissage, il est aussi connue pour avoir formulé avec Dodson, la loi Yerkes-Dodson qui décrit une relation entre niveau d'éveil (arousal) et performance. Cette loi, courbe en U inversée (courbe en "cloche"), dit que le niveau de performance varie avec le niveau d'éveil mais passerait par un optimal. Si le niveau d'éveil devient trop fort la performance est affectée.

WITMER, Lighter (1867 - 1956). Psychologue américain dont la place n'est probablement pas justifié dans ce manuel mais il est un auteur souvent oublié. Witner et non Daniel Lagache a introduit et participé au développement du terme "psychologie clinique". Il a créé la première "clinique psychologique" au monde à l'Université de Pennsylvanie en 1896, ainsi que le premier journal de psychologie clinique et la première école hospitalière clinique en 1907. Il a contribué à de nombreuses branches de la psychologie et au développement de l'éducation spécialisée (pour plus de détail sur l'auteur, cf. Thomas 2009)

Pour aller plus loin et connaître les auteurs ayant influencé le développement des conceptions sur l'intelligence

<http://www.intelltheory.com/map.shtml>

J - Glossaire

A-B

ADAPTATIF (test) : cf. TEST ADAPTATIF

AFFINE : cf. fonction affine

APPLATISSEMENT (Coefficient) : cf. KURTOSIS.

APTITUDE : cf. TESTS D'APTITUDE.

ASYMPTOTE : (origine grec qui associe le privatif "a" et "symptôsis" qui signifie rencontre). En mathématique, ligne droite qui s'approche indéfiniment d'une courbe à une distance de plus en plus petite sans jamais la couper.

ASYMETRIE (coefficient) : le coefficient d'asymétrie (skewness) est un des paramètres de forme de la distribution concernant l'écart par rapport à la symétrie. Un coefficient d'asymétrie négatif correspond à une distribution plus étirée à gauche et inversement pour les coefficients positifs.

BARYCENTRE : pour un ensemble fini de points d'un espace à n dimensions ($n \geq 1$), le barycentre est le point obtenu comme la moyenne arithmétique des positions de chacun de ces points sur ces dimensions auxquels on peut éventuellement affecter des coefficients de pondération. Par exemple, dans un plan, pour 2 variables (x en abscisse et y en ordonnée), l'abscisse du barycentre sera la moyenne pondérée des abscisses et l'ordonnée la moyenne pondérée des ordonnées. Lorsque ces coefficients de pondération sont égaux, le barycentre est appelé isobarycentre.

BIAIS DE REPONSE : on parle de biais de réponse lorsque la réponse à un item a tendance à être déterminée par des éléments externes à ce que l'item est censé mesurer. (cf. aussi BIAS des TESTS)

BIAIS des TESTS : Selon les standards de la construction des tests, le terme biais fait référence à une erreur de mesure non aléatoire introduite lors de la construction d'un test et entraînant des scores systématiquement inférieurs ou supérieurs pour des groupes de personnes. Attention, ce n'est pas parce qu'il existe des différences en fonction des catégories socioprofessionnelles qu'un test d'intelligence est biaisé. Il est biaisé si les différences observées ne sont pas de l'ordre de celles attendues (en supposant que l'on connaisse a priori l'ordre de grandeur éventuelle de ces différences).

BOX-COX (transformation ou transformée de Box-Cox) : méthode de transformation des scores permettant de normaliser une distribution. Cette méthode non linéaire est très utilisée en statistiques. Son nom fait référence à deux auteurs qui ont proposé cette transformation en 1964 : George Box et David Roxbee Cox.

C

CHI-CARRE - TEST : Statistique permettant de déterminer si la différence entre deux distributions de fréquences est attribuable à l'erreur d'échantillonnage (le hasard) ou est suffisamment grande pour être significative. Cette statistique suit la loi du CHI-CARRE.

CHI-CARRE -LOI : la loi du Chi carré est une loi utilisée en statistique inférentielle. Sa distribution est asymétrique (asymétrie gauche) et dépend d'un seul paramètre k . C'est la somme des carrés de k lois normales centrées réduites indépendantes (k étant le degré de liberté de cette loi). Elle est souvent utilisée pour les tests statistiques basés sur la somme des carrés des écarts (exemple le test du Chi-carré).

COEFFICIENT DE FIDELITE : Noté r_{xx} , ce coefficient correspond au carré de l'index de fidélité. Il varie entre 0 et 1 et renseigne sur la proportion de variance vraie. Plus il est proche de 1, plus l'erreur de mesure est faible (cf. [chapitre E §6.3](#)).

COHERENCE INTERNE : cf. CONSISTANCE INTERNE.

COMPOSITE (SCORE) : cf. SCORE COMPOSITE.

COMMUNAUTÉ : En analyse factorielle, la communauté (h^2), indique pour chaque variable la quantité de variance de la variable expliquée par les composantes (en analyse en composantes principales) ou facteurs (en analyse factorielle exploratoire) retenues (en général les n premiers facteurs ou les n premières composantes). La valeur de la communauté d'une variable correspond à la somme des carrés des saturations entre la variable et les facteurs (ou composantes).

CONGENERIQUE : cf. MODELE DE MESURE CONGENERIQUE.

CONSISTANCE INTERNE : on parle de consistance interne lorsque chacun des items mesure un même construit.

CONSTRUIT : désigne un objet mental (dérivé d'une démarche scientifique) destiné à représenter quelque chose qui n'est pas concret, n'a pas en soi de grandeur et n'a de réalité que celle créée par l'opération de mesure (par exemple l'intelligence, l'extraversion, le neuroticisme sont des construits). Un construit est, pour résumer, une entité non observable dont l'existence est inférée à partir d'observations. Un des objectifs de la psychométrie est de mesurer des construits. Ces construits sont des variables latentes (sources des différences interindividuelles observées dans des tâches) et correspondent à des dimensions théoriques hypothétiques.

COTE Z : cf. SCORE Z.

COURBE ROC (Receiver Operating Characteristic) : courbe traduisant l'efficacité d'un seuil de classification binaire (présence/absence). Ces courbes furent inventées pendant la seconde guerre mondiale pour montrer la séparation entre des signaux radars et le bruit de fond (indicateur de la relation entre la probabilité d'une détection et la probabilité d'une fausse alerte). Ces courbes utilisées en psychologie (méthode des tests) ou dans le domaine médical permettent la détermination de la valeur seuil optimale mais aussi la comparaison de plusieurs tests. On utilise le terme anglais courbe ROC qui pourrait être traduit par courbe Caractéristique du Fonctionnement ou d'Efficacité d'un Récepteur.

COVARIANCE : La covariance entre deux variables peut être considérée comme une extension de la notion de variance puisque la covariance est la moyenne des carrés des distances à l'isobarycentre du nuage de points (nuage défini par les scores observés sur chacune des variables). La valeur de la covariance dépend donc de la relation qui existe entre les variables mais aussi de la variance sur chaque variable (et donc de l'échelle de mesure).

D

DETERMINANT D'UNE MATRICE : la définition est souvent calculatoire (comment calcule-t-on le déterminant d'une matrice). Simple pour les matrices carrées de 2 par 2, le calcul du déterminant d'une matrice est complexe et ne sera pas présenté ici. Le calcul du déterminant permet de savoir si une matrice est inversible (c'est-à-dire si on peut calculer une matrice B qui multipliée par la première donne la matrice d'identité I : $AB = BA = I$). Le déterminant est une valeur numérique qui peut prendre n'importe quelle valeur entre 0 et 1. Un déterminant de 0 indique que la matrice est singulière (non inversible). D'un point de vue géométrique, si les lignes de la matrice sont des vecteurs, le déterminant correspond au volume du parallélépipède engendré par ces vecteurs.

DECILE : quantile d'ordre 10 soit 9 valeurs qui partagent l'étendue des scores brutes ordonnés en n sous-ensembles d'effectifs contenant chacun 10% des scores observés.

DIAGRAMME DE VENN : Façon de représenter des relations simples (introduit par J. Venn en 1880). Par exemple, concernant les relations entre deux distributions de notes, on utilise le diagramme de Venn en représentant les variances partagées et non partagées par un ensemble de cercles (ou ovales) qui se chevauchent.

DIAGONALE PRINCIPALE D'UNE MATRICE : La diagonale principale d'une matrice est l'ensemble des éléments situés de l'angle en haut à gauche jusqu'à l'angle en bas à droite, c'est-à-dire les éléments où la ligne et la colonne ont le même indice (a_{ii}). Rappel : dans une matrice carrée (n lignes et n colonnes), les colonnes de la matrices comme les lignes sont numérotées de 1 à n .

DIFFICULTE (paramètre de difficulté ou indice de difficulté) : dans la TCT, ce paramètre noté p (pour puissance) est le pourcentage de réussite de l'item. Dans les modèles de réponse à l'item, par convention, ce paramètre pour un item correspond au niveau d'aptitude (θ) qui est réussi à 50% (cf. [chap. E §5.3](#))

DIMENSION : une dimension en psychologie représente la variation d'un trait ou processus psychologique en intensité.

DIMENSION THEORIQUE : on parle de dimension théorique lorsque l'on fait référence une "construction conceptuelle" pour rendre compte de différences interindividuelles particulières (intelligence, extraversion/introversion, anxiété, etc.). Elle n'est pas directement mesurée, mais inférée à partir d'un cadre théorique. Ce terme est à distinguer de [variable latente](#) qui est un concept statistique même s'ils sont proches. La dimension théorique est un concept abstrait non observable et inféré. La variable latente est une représentation statistique que l'on interprète et que l'on peut associer à une dimension théorique.

DIMENSION OPERATIONNELLE : Une dimension opérationnelle est une différenciation d'individus qui résulte d'une opération de mesure. Cette dimension opérationnelle est supposée être la manifestation d'une dimension théorique (exemples : le QI mesuré au WISC-IV, le score à l'échelle de féminité du test CPI, etc). Les dimensions opérationnelles sont censées refléter les dimensions théoriques, mais la relation n'est jamais simple et dépend de la validité des outils de mesure utilisés mais aussi de la validité des hypothèses théoriques.

DIMENSIONNALITE : nombre de traits latent sous tendant la réponse à un item (cf. aussi unidimensionnalité).

DISTRIBUTION CENTREE : On dit qu'une distribution est centrée si son espérance (sa moyenne) est nulle. Pour centrer une distribution il suffit de retirer à chaque valeur de la distribution la moyenne de la distribution (moyenne calculée avant de la centrer).

DISTRIBUTION NORMALE : Le terme de distribution *Normale* été utilisé pour la première fois par Galton en 1889. Il s'agit d'une distribution associée à la loi normale qui présente les caractéristiques suivantes : (i) la distribution est symétrique (ii) la moyenne, le mode et la médiane sont identiques (iii) sa fonction de répartition est connue et 68% des observations sont à plus ou moins un écart-type de la moyenne. Elle sert de base pour déterminer l'asymétrie et l'aplatissement (kurtosis) d'une distribution.

DISTRIBUTION NORMALE CENTREE REDUITE : Distribution normale dont la moyenne est 0 et l'écart-type 1.

DISTRIBUTION REDUITE : On dit qu'une distribution est réduite si son écart-type est égal à 1. Pour

réduire une distribution, il suffit de diviser toutes les valeurs par l'écart-type de la distribution avant réduction.

E

ECART-TYPE : L'écart-type est une mesure de la dispersion des valeurs d'une distribution autour de leur moyenne. C'est la racine carrée de la moyenne non pas des écarts à la moyenne (qui serait égale à 0) mais des carrés des écarts à la moyenne : $\sigma = \sqrt{(x_i - m_x)^2}$.

ECHANTILLON : sous ensemble d'individu représentatif d'une population et obtenu par une méthode d'échantillonnage.

ECHANTILLON NORMATIF : non donné à l'échantillon ou aux échantillons qui permettent de construire l'étalonnage d'une épreuve. On préfère en général le terme d'échantillon de standardisation.

ECHANTILLON DE STANDARDISATION : non donné à l'échantillon qui sert à "standardiser" le test.

ECHANTILLONAGE : sélection d'un sous ensemble d'individu d'une population par une méthode permettant d'assurer que ce sous ensemble soit représentatif de la population pour la variable mesurée (cf. cours pour plus de précision sur les méthodes d'échantillonnage, [chapitre. D](#))

ERREUR ALEATOIRE (ou ERREUR NON-SYSTEMATIQUE) : écart entre score vrai et score observé dont la cause est un ensemble de facteurs inconnus qui font que parfois la mesure sera légèrement supérieure à la valeur réelle et parfois légèrement inférieure. Cette erreur aléatoire est celle qui est associée à la notion de fidélité.

ERREUR STANDARD DE MESURE (ESM ou SEM en anglais) : Écart-type de la distribution de l'erreur de mesure. Ne doit pas être confondu avec l'erreur type (cf. [chapitre F §1.1](#)). On utilise parfois comme sigle SEM (de l'anglais Standard Error of measurement). Dans l'absolu correspond à l'écart-type des scores observés sur des mesures parallèles répétées pour une personne ayant une note "vraie" fixe et invariable.

ERREUR STANDARD DE MESURE CONDITIONNELLE (ESM-C ou C-SEM en anglais). Terme utilisé lorsque l'on calcule l'erreur de mesure standard pour chaque valeur du trait latent et non plus la même valeur quel que soit la position d'une personne sur le trait latent.

ERREUR SYSTEMATIQUE : déviation constante, négative ou positive introduit par l'instrument de mesure. Cette erreur n'est pas aléatoire et n'est donc pas évaluée par la fidélité. La fidélité d'une épreuve peut-être bonne mais l'erreur systématique de mesure importante (cf. [chapitre E §6.3.1](#)).

ERREUR TYPE : mesure standard de l'erreur d'échantillonnage (correspond à l'écart type de l'estimateur de la moyenne pour un échantillon).

ESPERANCE MATHÉMATIQUE (E(X)) : L'espérance mathématique d'une variable aléatoire est la valeur moyenne que l'on s'attend à trouver si l'on répète un grand nombre de fois la même expérience aléatoire. Par exemple si une variable X prend les valeurs $x_1, x_2, x_3, \dots, x_n$ avec les probabilités $p_1, p_2, p_3, \dots, p_n$, $E(X)$ est la somme des x_j pondéré par leur probabilité.

ESTIMATEUR : un estimateur est une statistique permettant d'évaluer un paramètre relatif à une distribution (la moyenne, la variance, etc.) à partir d'un échantillon de données.

ETALONNAGE : procédé qui consiste à établir des catégories ordonnées de références ou des classes ordonnées à partir des notes brutes de l'échantillon (cf. [Chapitre G](#))

ETALONNAGE (tables) : tables accompagnant le manuel d'un test et permettant de convertir les scores brutes en scores standard, percentiles ou autres échelles dont les propriétés sont connues.

L'étalonnage permet l'interprétation des scores observés (par comparaison à une norme ou un échantillon normatif).

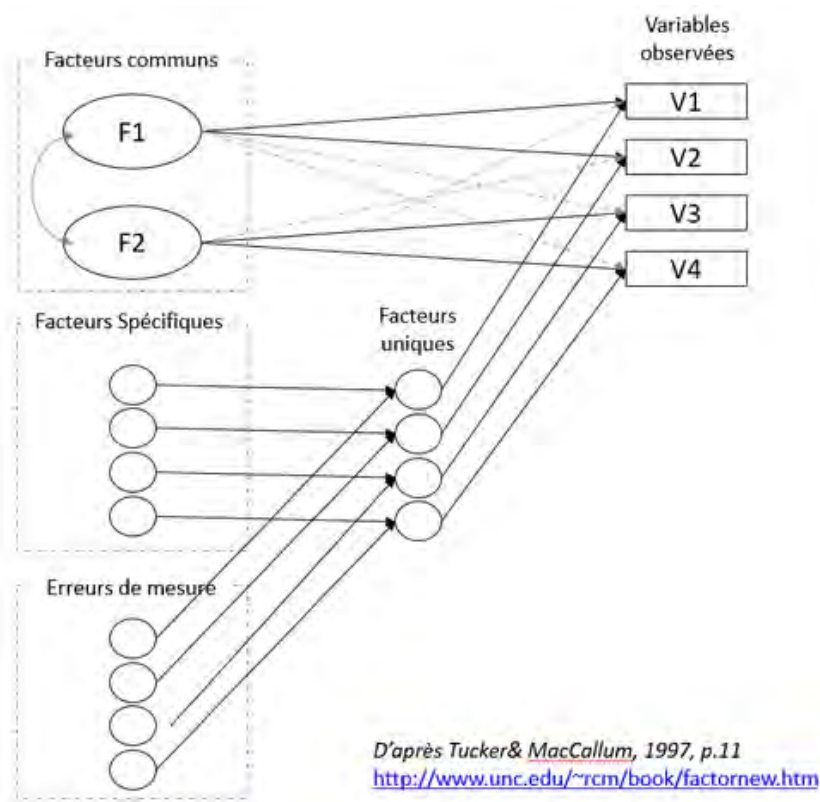
ETHIQUE (usage des tests) : Ensemble de principes moraux qui engagent (responsabilités et obligations) ceux qui construisent les tests mais aussi ceux qui les utilisent.

F

FIDELITE : En psychométrie, la fidélité caractérise la qualité de la mesure et plus particulièrement la capacité à donner des valeurs exemptes d'erreur aléatoire. Assurer la fidélité d'un test c'est assurer que l'on mesure quelque chose (cf. aussi COEFFICIENT DE FIDELITE, INDEX DE FIDELITE)

FACTEUR SPECIFIQUE : Source de variation différente de l'erreur de mesure mais spécifique à une variable observée et source d'une partie des différences interindividuelles.

FACTEUR UNIQUE : terme utilisée en analyse factorielle. Correspond à des variables latentes combinant les facteurs spécifiques et l'erreur de mesure (cf. aussi UNICITE)



FONCTION AFFINE : une fonction affine est une fonction qui à toute valeur x associe le nombre $ax + b$, a et b étant des nombres relatifs qui ne dépendent pas de x . Un cas particulier des fonctions affines est lorsque l'ordonnée à l'origine (b) est nulle, on obtient alors une fonction linéaire.

FONCTION LINEAIRE : cf. FONCTION AFFINE

FONCTION LOGISTIQUE : cf. LOGISTIQUE

FONCTION MONOTONE : cf. MONOTONE

G-H

GENERALISABILITE (théorie) : Alors que TCT décompose la variance de score observée en variance de score vrai et variance d'erreur aléatoire indifférenciée, la théorie de généralisabilité propose une procédure pour estimer les sources de variance d'erreur de mesure à l'aide des méthodes d'analyse

de variance (ANOVA). Elle présente l'avantage de permettre simultanément la quantification de plusieurs sources de variance d'erreur de mesure et leurs interactions (sources supplémentaires de variance d'erreur). Cette théorie est cependant peu présente dans le champs de la psychométrie (en psychologie) mais plus utilisée en Sciences de l'Education (mesure académique ou autres).

HASARD (tirage au hasard) : on parle de tirage au hasard lorsque chaque élément d'un ensemble a la même probabilité d'être sélectionné (ce qui est le cas dans l'échantillonnage probabiliste par exemple). Dans tous les autres cas, le terme de tirage au hasard est incorrect.

HOMOSCEDASTICITE : on parle d'homoscédasticité lorsque la variance des erreurs d'un modèle est identique pour toutes les observations. Par exemple, si la même mesure est effectuée dans 5 sous-groupes différents, on parlera d'homoscédasticité si les variances sont égales et d'hétéroscedasticité si elles sont différentes.

HETEROSCEDASTICITE : s'oppose à homoscédasticité.

INDEX DE FIDELITE : dans la TCT, correspond à la corrélation entre le score vrai et le score observé. On ne doit pas confondre cet index avec le coefficient de fidélité. En général, quand on parle de la fidélité, on fait référence au coefficient de fidélité et non à l'index.

ISOBARYCENTRE : cf. BARYCENTRE.

I

INDICE KMO : cf. KMO

INFLEXION : cf. POINT D'INFLEXION

INTERVALLE DE CONFIANCE : de façon générale, intervalle dans lequel, si le paramètre à estimer ne se trouve pas, il y avait a priori une faible probabilité d'obtenir l'estimation obtenue. Ainsi, un intervalle de confiance à 95 % donnera un encadrement correct 95 fois sur 100 en moyenne et on se tromperait en moyenne 5 fois sur cent.

IPSATIVE (mesure) : de façon générale, se dit d'une méthode de mesure qui utilise comme référence les autres mesures le concernant (cf. aussi TEST IPSATIF)

IRT (ITEM RESPONSE THEORY) : cf. MRI

ITC (International Test Commission). Commission internationale qui fixe des directives générales (guidelines) concernant les tests (construction et usage). Actuellement, la ITC a 21 membres titulaires (associations psychologiques professionnelles nationales), 64 membres affiliés (autres commissions de test, les éditeurs et les organismes de recherche impliqués dans les tests) et plus de 700 membres individuels (personnes qui travaillent ou qui ont un intérêt dans les tests et les essais). Sa composition actuelle couvre la plupart des pays d'Europe occidentale et orientale, d'Amérique du Nord, ainsi que certains pays du Moyen et Extrême-Orient, Amérique du Sud et en Afrique. Pour plus de détails : www.intestcom.org

ITEM : plus petit élément d'un test auquel on assigne, en fonction de la réponse, un score.

J-K-L

KMO (indice KMO) : indice de (Kaiser-Meyer-Olkin qui nous renseigne sur la qualité des corrélations d'une matrice de corrélations. Cet indice prend des valeurs entre 0.0 et 1.0 et sa valeur devrait être égale ou supérieure à .60 et est considéré comme correct à partir de .70.

KURTOSE - KURTOSIS : Le coefficient d'aplatissement de Pearson ou kurtosis nous renseigne sur le degré d'aplatissement d'une distribution (voussure). Il correspond au moment centré d'ordre 4

divisé par le carré de la variance. La kurtosis d'une loi normale (de Gauss) est égale à 3. Dans la plupart des cas, on retranche 3 à la formule ce qui donne le coefficient de Fisher appelé par les anglais excès d'aplatissement ("excess kurtosis"). C'est ce kurtosis normalisé qui est reporté le plus souvent par les logiciels.

LEPTOKURTIQUE - LEPTOCURTIQUE (distribution) : si dit d'une distribution ayant une kurtose (coefficient d'aplatissement) élevée c'est à dire supérieur à celle d'une courbe normale. La distribution est plutôt "pointue" autour de la moyenne, et a des queues de distribution épaisses.

LOGISTIQUE (fonction) : Les courbes représentatives d'une fonction logistique ont la forme d'un S ce qui fait qu'elles sont parfois appelées sigmoïdes. Ces fonctions ont été mises en évidence par P-F Verhulst qui cherchait un modèle d'évolution non exponentielle ou borné d'une population. La fonction logistique à trois paramètres (k, b, a) avec a positif et est de la forme $k/(1+be^{-ax})$. C'est donc une composée de fonction affine, exponentielle et inverse. Le numérateur k étant la limite de de la fonction à l'infini (plafond de la courbe en S) et est symétrique par rapport à son point d'inflexion .

LOGIT : logarithme népérien du rapport de vraisemblance (log odds-unit)

M

MATRICE D'IDENTITE : Une matrice d'identité est une matrice carrée avec des 1 sur la diagonale et des 0 partout ailleurs.

MATRICE SINGULIERE : Une matrice qui est non inversible est singulière. Lorsque l'on réalise une analyse factorielle, une matrice de variances-covariances singulière rendra impossible l'analyse (pas de solution factorielle). On observe des matrices singulières lorsqu'une variable est parfaitement corrélée avec une autre variable ou avec une combinaison de plusieurs variables. Cette condition peut être détectée en calculant le « déterminant » de la matrice.

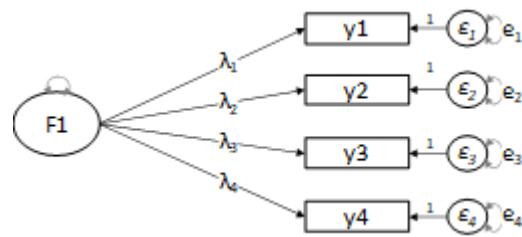
MESOKURTIQUE - MESOCURTIQUE (distribution) : se dit d'une distribution ayant une kurtosis normalisée égale à 0. La distribution normale est mésokurtique (coefficient d'aplatissement égal à 0.

MESURE FORMATIVE. On parle de modèle de mesure formative lorsque les variables mesurées sont la cause du "construit" mesuré. Une variable est dite formative lorsqu'elle est « formée » ou directement modifiée et influencée par les indicateurs ou les items du test. Ce sont donc les indicateurs qui "créent" le construit mesuré. Plus rarement utilisée en psychologie (cf. [chapitre E §2](#)).

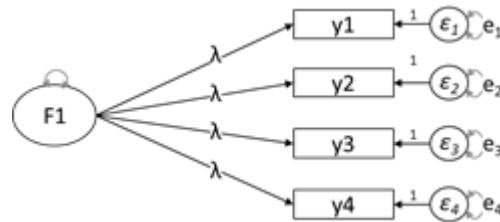
MESURE REFLECTVE. Correspond à la démarche habituelle en psychologie. On suppose qu'il existe une dimension sous-jacente (variable latente) théorique (non observable) et que le résultat au test est causé par cette dimension (la variable latente). La dimension théorique prédit les performances aux items du test qui doivent donc corrélérer entre eux.

MODELE DE MESURE. Les modèles de mesure se réfèrent aux modèles implicites ou explicites qui relient le construit et ses indicateurs. Ces modèles décrivent aussi les conditions d'application, les notions d'erreur, de score vrai, d'erreur-type, d'estimation de la fidélité, les méthodes d'analyse d'items, la dimensionnalité des construits, etc.

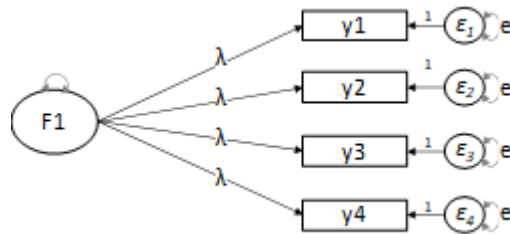
MODELE DE MESURE CONGENERIQUE : Ce terme est un anglicisme. On parle de modèle de mesure congénérique lorsque la variance vraie de chaque variable dépend d'une même variable latente et que les erreurs de mesure ne covarient pas entre elles.



MODELE DE MESURE TAU-EQUIVALENT (τ -équivalent). On parle de tau-équivalence pour un modèle de mesure lorsque celui-ci est congénérique et que de plus la variance vraie mesurée par chaque variable est constante.



MODELE DE MESURE PARRALELE : On parle de parallélisme ou modèle de mesure parallèle lorsque celui-ci est tau-équivalent et que de plus la variance d'erreur (e) est constante.



MULTIDIMENSIONNALITE : on parle de multidimensionnalité lorsque un test à une dimensionnalité supérieure à 1 (*i.e.* n'est pas sous-tendu par un seul trait latent).

MOMENT : Si X est une variable aléatoire, on appelle moment d'ordre k , s'il existe, le nombre $E(X^k)$.

MOMENT CENTRE : Le moment centré d'une variable aléatoire est $E[(X - E(X^k))]^2$. Le moment centré d'ordre 2 est donc la variance de la variable (cf. aussi KURTOSIS et ASSYMETRIE)

MONOTONE (fonction) : fonction uniquement croissante ou décroissante.

MRI (modèle de réponse à l'item) : modèles qui définissent les relations entre les réponses aux items d'un test et le construit par une fonction qui donne la probabilité de fournir une bonne réponse en fonction du niveau de la personne sur un trait latent. Ces modèles (cf. le cours) sont une alternative à la TCT et s'imposent progressivement en psychométrie.

N

NORMALE (DISTRIBUTION) : cf. DISTRIBUTION NORMALE.

NUAGE DE POINTS. Terme souvent utilisé lors du calcul de corrélations ou en analyse factorielle. Le nuage de points est la représentation de l'ensemble des observations dans un espace ayant un nombre de dimensions égal au nombre des variables. Chaque point correspond à un sujet (une observation) et les coordonnées d'un point correspondent aux scores d'un sujet sur chacune des variables mesurées (donc si on a dix variables, le sujet est représenté dans un espace à 10 dimensions).

NUAGE DE POINTS - CENTRE DE GRAVITE : Le centre de gravité d'un nuage de points est le point ayant pour coordonnées les moyennes calculées sur chacune des variables.

NUAGE DE POINTS - VARIANCE (variables quantitatives). La variance du nuage de point est la moyenne des carrés des distances au centre de gravité (inertie). La formule de la variance apprise pour une variable est donc généralisée à un espace à plusieurs variables. **Remarque** : la variance du nuage de points est égale à la somme des variances de chacune des variables (théorème de Huygens).

O-P

OUTLIER : Un outlier (valeur aberrante ou non conforme) est une observation qui est trop peu probable au regard des autres valeurs dans un échantillon aléatoire d'une population. Cette définition "vague" permet de bien comprendre que c'est celui qui analyse les données qui décide de ce qui sera considéré comme anormal ou trop différent (quel que soit la méthode qu'il utilise). Cela suppose aussi par ailleurs que l'on sache caractériser les observations normales.

PARRALLELISME (modèle de mesure parallèle) : cf. MODELE DE MESURE PARRALELE

PARAMETRE (statistique) : Les paramètres statistiques sont des résumés de distributions ou de séries statistiques (tendance centrale, dispersion, asymétrie, kurtose, etc.) qui résume l'information relative à l'observation. A ne pas confondre avec les estimateurs qui ont pour objectifs d'estimer les paramètres à partir d'un sous ensemble d'information.

PERCENTILE (rang) : cf. RANG PERCENTILE

PERCENTILE (score) : cf. SCORE PERCENTILE

PLATIKURTIQUE - PLATICURTIQUE (distribution) : se dit d'une distribution ayant une kurtosis normalisée négative. La distribution est aplatie (excès d'aplatissement).

POINT D'INFLEXION : en mathématique, point où s'opère un changement de concavité d'une courbe. En un tel point, la tangente traverse la courbe.

POPULATION (population mère) : ensemble de tous les individus ou unités d'observation dans lequel on extrait un échantillon.

Q-R

QUANTILE (d'ordre n) : chacune des $n - 1$ valeurs d'un caractère quantitatif qui partagent l'étendue ordonnée des valeurs en n sous-ensembles d'effectifs égaux (ordonnés).

RANG PERCENTILE : pour une valeur donnée p comprise entre 0 et 100, le rang percentile est le score brut pour lequel $p\%$ de l'échantillon ou de la population ont un score inférieur. Par exemple si le rang percentile 80 est la valeur 16, cela signifie que 80% des observations ont des valeurs inférieures à 16.

R (R cran) : À l'origine, logiciel destiné à l'enseignement et à l'apprentissage des statistiques, proche du langage S développé par R. Becker, J. Chambers et A. Wilks (laboratoires Bell). Actuellement c'est surtout un langage et un environnement pour le traitement de données. Il présente de nombreux avantages et est particulièrement bien adapté à l'analyse statistique. Logiciel libre, il permet de disposer d'un outil gratuit, ouvert et en perpétuelle évolution (<https://cran.r-project.org/>).

RAPPORT DE VRAISEMBLANCE (odds en anglais) : rapport entre la probabilité d'occurrence d'un événement sur la probabilité complémentaire de cette occurrence.

REGRESSION (analyse de) : Ensemble de méthodes statistiques pour étudier la relation entre une

variable dépendante et une variable indépendante (régression simple) ou encore plusieurs variables indépendantes (régression multiple). Les variables indépendantes sont aussi appelées prédicteurs et la variable dépendante devient la variable prédite.

REGRESSION VERS LA MOYENNE : Ce phénomène a été décrit par l'anglais F. Galton. Il remarque que les enfants de parents de grande taille étaient souvent plus grands que la moyenne, mais toutefois, en moyenne, plus petits que leurs parents. Inversement, les enfants de parents de petites tailles sont plus petits que la moyenne, mais plus grands que leurs géniteurs. Cela s'explique par le fait que la taille résulte de nombreux facteurs (génétiques et environnementaux) et la probabilité qu'ils agissent consécutivement (deux mesures) dans le même sens est faible. On observe donc en général quand les scores sont élevés pour une mesure, une probabilité plus importante qu'ils soient plus faibles lors de la seconde mesure (et inversement). Un autre exemple : statistiquement, les élèves ayant les meilleurs scores à un contrôle ont en moyenne des scores un peu moins bons lors d'un second contrôle et inversement, les moins bons auront de meilleurs scores. Cet effet traduit que la note dépend de nombreuses facteurs dont certains sont aléatoires (parfois positifs, parfois négatifs).

REGRESSION POLYNOMIALE : Analyse de régression dans laquelle la relation entre la variable dépendante et la variable indépendante est modélisée comme un polynôme du $n^{\text{ième}}$ degré en x . ($y = \sum a_i x^i + \varepsilon$). L'intérêt de la régression polynomiale est donc de pouvoir introduire de la non-linéarité dans la relation entre deux variables.

ROBUSTE- ROBUSTESSE : La robustesse d'un test ou d'un indicateur est sa capacité à ne pas être modifié lorsque les conditions d'application ne sont pas totalement respectées ou, pour un indicateur, d'être peu sensible à la présence d'outliers.

S

SATURATION : En analyse factorielle exploratoire en composantes principales, la saturation correspond à la corrélation entre une variable et un facteur. La saturation varie entre -1 et +1 et le carré de la saturation traduit la proportion de variance de la variable expliquée par le facteur.

SCORE BRUT (RAW SCORE) : somme des scores (pondérés ou non) obtenu à chacun des items d'un test.

SCORE COMPOSITE : score obtenu en additionnant des scores de plusieurs tests ou sous-tests. Souvent présents dans les batterie test, ce sont des scores qui fournissent des mesures générales adaptées et synthétiques (exemple : le QI).

SCORE STANDARD : en statistique correspond à la transformation d'un score brut en un score qui exprime l'écart à la moyenne en fraction d'écart type (cf. z-score). En psychométrie, lorsque l'on parle d'échelle en scores standards on fait parfois aussi référence à des échelles obtenus par simple transformation du score z (exemples : score T, stanine, sten).

SCORE UNIVERS : Terme utilisé dans la théorie de la généralisabilité. Le score univers peut être considéré comme le score vrai dans cette théorie.

SCORE VRAI: Dans le contexte d'une variable aléatoire, le score vrai est défini comme l'espérance de la variable "score observé" pour une personne (moyenne observée si on passait le test une infinité de fois sans effet d'ordre). Une autre façon de dire c'est que c'est le score que l'on obtiendrait s'il n'y avait pas d'erreur de mesure (qui serait donc identique si on répétait la mesure indéfiniment). C'est donc un concept (non observable).

SCORE PERCENTILE : Nombre compris entre 0 et 100 associé à un score brut et correspondant au

pourcentage d'observations dans l'échantillon de standardisation ayant un score inférieur à ce score brut (ne pas confondre avec le rang percentile).

SCORE-Z (ou COTE Z) : Score standard dont la moyenne est 0 et l'écart-type de 1. S'obtient par une simple transformation linéaire $(x-m)/s$ (x_i étant les scores observés, m la moyenne de ces score et s l'écart-type). Exprime ainsi le nombre de fois en écart-type dont le score est éloigné de la moyenne (au dessus si positif, en dessous si négatif).

SINGULARITE : La singularité d'une matrice est un concept fondamental en algèbre linéaire. Une matrice carrée (autant de lignes que de colonnes) est dite singulière lorsque le [déterminant](#) de la matrice est égale à 0.

T

TAU-EQUIVALENCE (modèle τ -équivalent) : cf. MODELE DE MESURE TAU-EQUIVALENT

TAUX DE SONDAGE : Proportion de la population qui e t sélectionnée pour constituer un échantillon. Il est égal à la taille de l'échantillon divisé taille de la population de base (le tout multiplié par 100).

TABLE DE CONTINGENCE : Tableau à double entrée qui croise deux variables (nominales ou ordinales) dont les modalités de la première variable définissent les lignes du tableau et les modalités de la secondes les colonnes de ce même tableau. Les cases de ce tableau contiennent l'effectif des "individus" ou unité d'étude cumulant la conjonction des caractères décrits par la ligne et la colonne considérée. Ces tableaux permettent de détecter d'éventuelles dépendances entre les variables. Ce terme aurait été introduit introduit par K. Pearson en 1904.

TCT (THEORIE CLASSIQUE DES TESTS) : modèle théorique classiquement utilisé en psychométrie qui considère que tous les scores observés sont l'addition de deux composantes : le score vrai (T) et l'erreur de mesure (E) : $X = T + E$

TEST ADAPTATIF : Modalité de passation des tests ou chaque item est sélectionné en fonction du niveau de la personne sur le trait latent. Toutes les personnes ne passent donc pas les mêmes items et le niveau de la personne est réévalué automatiquement après chaque réponse donnée. Ce mode de test peut se développer grâce aux apports des modèles de réponses à l'item (MRI). L'efficacité de ce type d'évaluation permet de réduire significativement le nombre d'items dans les questionnaires de personnalité ou les tests de connaissances sans perte de fidélité.

TEST D'APTITUDE : En psychologie, un test d'aptitude est un test permettant d'évaluer la capacité à acquérir des connaissances ou à traiter des informations dans des domaines particuliers (aptitudes : verbale, spatiale, numérique, etc.).

TESTS DE VITESSE : par opposition aux tests de puissance, les tests de vitesse privilégient l'évaluation d'une aptitude ou de connaissances, le temps d'exécution comme indicateur direct ou indirect de la performances (les items sont le plus souvent des items simples).

TEST DE PUISSANCE : A l'inverse des tests de vitesse, les tests de puissance n'ont pas ou peu de limite de temps et privilégient pour évaluer les connaissances ou les aptitudes des items complexes. Le niveau de complexité des items réussis devient un indicateur de la performance (et non pas le temps d'exécution).

TEST IPSATIF : consiste à comparer les scores d'une personne sur un sous-test à leur propre score sur les autres sous tests. On parle parfois de tests auto-référencés (self-referenced test) par opposition aux tests normatifs (cf. aussi IPSATIVE)

THEOREME CENTRAL LIMITE : théorème de Pierre Simon Laplace qui énonce que toute somme de

variables aléatoires indépendantes et identiquement distribuées tend vers une variable aléatoire gaussienne (loi normale). La portée de ce théorème est essentielle en statistiques et explique l'omniprésence de la loi normale.

TRACE D'UNE MATRICE. La trace d'une matrice carrée est la somme des éléments de sa [diagonale principale](#) c'est à dire la somme des éléments qui ont le même indice a_{ii} . Par exemple, en psychologie on réalise par défaut des ACP normées (sur des variables centrées réduites), la trace de la matrice de variances-covariances soumise à l'analyse est égale au nombre de variables et représente la quantité de variance du nuage de points (chaque coefficient étant la variance de la variable centrée réduite). Pour les AFE la diagonale de la matrice contient la part de variance de chaque variable qui doit être expliquée. La trace représente donc la quantité de variance à expliquer par le système de facteurs extraits.

TRAIT : Terme utilisé dans le domaine de la personnalité. Il définit des caractéristiques psychologiques ou plus exactement des dispositions relativement stables qui permettent de prédire les comportements. En psychologie on distingue le "trait" et "l'état". Par exemple l'anxiété trait est une caractéristique de la personne pouvant affecter plus ou moins l'ensemble des conduites alors que l'anxiété état est un niveau d'anxiété lié à un moment ou une situation et qui n'est pas permanent.

U-V-X-Y-Z

UNICITE : Terme utilisée en analyse factorielle. Il correspond pour une variable manifeste (observée) à la variance non expliquée par le système de facteurs : $u^2 = 1 - h^2$ (h^2 étant la communauté). Cf. aussi "Facteur unique".

UNIDIMENSIONNALITE : En principe on parle d'unidimensionnalité lorsque chaque item d'un test ne dépend que d'une seule dimension (la dimensionnalité est de 1).

UNIDIMENSIONNALITE ESSENTIELLE OU DOMINANTE : L'unidimensionnalité étant rarement respectée, on préfère parler d'unidimensionnalité essentielle ou dominante lorsque qu'une variable latente domine pour expliquer les réponses aux items.

UNIVERS (score) : cf. SCORE UNIVERS.

VALEUR PROPRE ("EIGENVALUE") : Terme associé à une composante ou un facteur en analyse factorielle et qui désigne la somme des carrés des saturations entre ce facteur (ou cette composante) et chacune des variables. La valeur propre représente ainsi la quantité de variance expliquée par le facteur ou la composante. De façon plus formelle la notion de valeur propre s'applique à des applications linéaires d'un espace vectoriel dans lui-même (endomorphisme). Un scalaire λ est une valeur propre d'une matrice carrée $U_{n \times n}$ s'il existe un vecteur x (appelé alors vecteur propre) non nul tel que $u(x) = \lambda x$. Mathématiquement : cf [vecteur propre](#).

VALIDATION : correspond à l'ensemble des procédures mis en place pour évaluer la validité d'un test.

VALIDITE : se dit lorsque le test mesure ce que l'on souhaite mesurer. Ensemble des preuves empiriques et théoriques accumulées pour supporter l'interprétation des résultats d'un test.

VARIABLE ALEATOIRE : application qui associe à tout événement élémentaire d'un ensemble (univers des éventualités) un nombre. Un exemple simple est le résultat d'un jet de dés, pour lequel les valeurs possibles sont 1, 2, 3, 4, 5 ou 6. Cette variable aléatoire est dite discrète (elle est définie sur le sous ensemble des nombres entiers) mais les variables aléatoires peuvent être continues.

VALIDITE APPARENTE : La validité apparente (face validity) est un jugement subjectif sur les items d'un

test ("sont-ils conforme pour mesurer ce que l'on veut mesurer ?"). La validité apparente est parfois considérée comme une forme très faible de validité conceptuelle.

VARIABLE MANIFESTE : Mesure observée ou score(s) observé(s) suite à une opération de mesure (exemple : temps de réaction). Cette variable manifeste (ou observée) peut être composite (somme des scores obtenus à un ensemble de questions par exemple).

VARIABLE OBSERVEE : cf. VARIABLE MANIFESTE

VARIABLE LATENTE : variable non observable (construit statistique) source d'une partie des différences interindividuelles observées sur des variables manifestes. Une variable latente est toujours évaluée indirectement par ces effets sur des variables manifestes. L'analyse factorielle exploratoire (EFA) est une des méthodes permettant d'identifier des variables latentes.

VARIANCE : Écart-type au carré (σ^2), il s'agit donc d'une mesure de la dispersion des valeurs d'une distribution autour de leur moyenne. C'est la moyenne des carrés des écarts à la moyenne : $\sigma^2 = \frac{\sum (x_i - m_x)^2}{n}$

VECTEUR PROPRE : En mathématiques, la notion de vecteur propre s'applique à des applications linéaires d'un espace vectoriel dans lui-même (endomorphisme). Un vecteur propre d'une matrice carrée est un vecteur non nul dont la direction ne change pas lorsqu'on lui applique la matrice, et qui est simplement multiplié par une constante appelée valeur propre. Le nombre maximal de vecteurs propres non colinéaire d'une matrice carré $n \times n$ diagonalisable est égal à n . En analyse factorielle, ces vecteurs propres sont associés à un facteur ou une composante et constituent l'ensemble des saturations entre un facteur (composante) et chacune des variables.

VOUSSURE : cf KURTOSE

Z-SCORE : cf. SCORE-Z

K - Liste des principaux acronymes utilisés

AFE	Analyse Factorielle Exploratoire
ACP	Analyse en Composantes Principales
AF	Analyse Factorielle
AFC	Analyse factorielle des correspondances (en France)
AGFI	Adjusted Goodness of Fit Index (indice de qualité d'ajustement corrigé)
AIC	Akaike Information Criterion (Critère d'information de Akaike)
APA	American Psychological Association
AUC	Area Under Curve (Aire sous la courbe ROC)
CA	Correspondance Analysis (acronyme anglais de l'AFC).
CCI	Courbe Caractéristique d'un Item.
CFI	Comparative Fit Index de Bentler.
CSEM/C-SEM	Acronyme anglais de ESM-C
EAWOP	European Association of Work and Organizational Psychology
EFPA	European Federation of Psychologists' Associations
ESM	Erreur Standard de Mesure
ESM-C	Erreur Standard de Mesure Conditionnelle
FN	Faux Négatif
FP	Faux Positif
GFI	Goodness of Fit Index (indice de qualité d'ajustement)
ITC	International Test Commission
MRI	Modèle de Réponse à l'Item
NCE	Normal Curve Equivalent (scores NCE)
NR	Non Réponse
QCM	Questionnaire à Choix Multiples
QI	Quotient Intellectuel (Qis, standard, Qic Classique)
RCI	Indice de changement fiable (reliable change index)
RMSEA	Root Mean Square Error of Approximation (erreur quadratique moyenne de l'approximation)
ROC	Receiver Operating Charateristic (Caractéristique du Fonctionnement ou Efficacité d'un Récepteur)
SEM	Acronyme anglais de ESM
SRMR	Standardized Root Mean Residual (indice de la racine du carré moyen d'erreur)
TCT/CTT	Théorie Classique des Tests/Classical Test Theory

TRI	Théorie des Réponses à l'Item
VN	Vrai Négatif
VP	Vrai Positif

L - Bibliographie

Ouvrages de référence

Hogan, T. P., Parent, N., & Stephensen, R. (2017). *Introduction à la psychométrie* (2ème édition). Canada: Chenelière-Education.

Laveault, D. & Grégoire J. (2014). *Introduction aux théories des tests en sciences humaines* (3ème édition). Bruxelles, De Boeck.

Compléments

Dickes P., Tournois J., Flieller A., et Kop J.-L. (1994) *La psychométrie : théories et méthodes de la mesure en psychologie*. Paris : Presses Universitaires de France.

Huteau, M. (1995). *Manuel de psychologie différentielle*. Paris : Dunod

Pour aller plus loin

Anastasi, A., Urbina, S. (1997). *Psychological Testing* (7th edition). Prentice Hall : New-York.

Urbina, S. (2014). *Essentials of Psychological Testing* (2nd edition). John Wiley & Sons.

Wasserman, J. D., & Bracken, B. A. (2013). Fundamental Psychometric Considerations in Assessment. In I. B. Weiner, J. R. Graham, & J. A. Naglieri (Eds.), *Handbook of Psychology. Assesment psychology* (2nd ed., Vol. 10, pp. 50–81). Hoboken, NJ, USA: John Wiley & Sons.

Bibliographie (hors ouvrages de références)

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association (Web site: <http://www.aera.net>. Retrieved from <https://eric.ed.gov/?id=ED565876>)

Beavers, Amy S.; Lounsbury, John W.; Richards, Jennifer K.; Huck, Schuyler W.; Skolits, Gary J.; and Esquivel, Shelley L. (2013) Practical Considerations for Using Exploratory Factor Analysis in Educational Research, *Practical Assessment, Research, and Evaluation*, 18, 1-13. doi: <https://doi.org/10.7275/qv2q-rk76>

Bernaud, J.L. (2007). *Introduction à la psychométrie*. Paris: Dunod, collection « Topos ».

Bernier, J.-J., Pietrulewicz, B. (1998). *La psychométrie. Traité de mesure appliquée*. Gaëtan Morin.

Benzecri, J.-P. (1973), *L'analyse des données tome 2 : l'analyse des correspondances*, Paris : Bordas

Benzecri, J.-P., (1982). *Histoire et préhistoire de l'analyse des données*, Dunod.

Binet A., Simon T. (1908, réédition, 1964) *La mesure du développement de l'intelligence chez le jeune enfant*. Paris : Armand Colin

Boake C. (2002). From the Binet–Simon to the Wechsler–Bellevue: Tracing the History of Intelligence Testing, *Journal of Clinical and Experimental Neuropsychology*, 24(3), 383–405. doi: 10.1076/jcen.24.3.383.981

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, 110(2), 203–219. doi : 10.1037/0033-295X.110.2.203

- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, 111(4), 1061–1071. doi : 10.1037/0033-295X.111.4.1061
- Bowman, M. L. (2002). The perfidy of percentiles. *Archives of Clinical Neuropsychology : The Official Journal of the National Academy of Neuropsychologists*, 17(3), 295–303. doi : 10.1016/S0887-6177(01)00116-0
- Brennan, R. L. (1972). A Generalized Upper-Lower Item Discrimination Index. *Educational and Psychological Measurement*, 32(2), 289–303. doi : 10.1177/001316447203200206
- Bratsberg B., Rogeberg O. (2018). Flynn effect and its reversal are both environmentally caused. *Proceedings of the National Academy of Sciences*, 201718793. <https://doi.org/10.1073/pnas.1718793115>
- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2), 245–276.
- Chapanis, A. (1951). Theory and methods for analyzing errors in man-machine systems. *Annals of the New York Academy of Sciences*, 51(7), 1179–1203. doi : 10.1111/j.1749-6632.1951.tb27345.x
- Cho, E. (2016). Making Reliability Reliable: A Systematic Approach to Reliability Coefficients. *Organizational Research Methods*, 19(4), 651–682. doi: 10.1177/1094428116656239
- Cohen, R.J., Swerdlik, M.E., Phillips, S.M. (1996). *Psychological Testing : An Introduction to Tests & Measurement* (3rd edition). Mayfield Publishing : New-York
- Crawford, J. R., & Garthwaite, P. H. (2005). Testing for suspected impairments and dissociations in single-case studies in neuropsychology: evaluation of alternatives using Monte Carlo simulations and revised tests for dissociations. *Neuropsychology*, 19, 318-331.
- Cronbach, L.J. (1984). *Essentials of Psychological Testing* (5th edition, 1990). Harpercollins College Div. : New-York
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302. doi : 10.1037/h0040957
- Dickes P., Flieller A., Tournois A., & Kopp, J.-L. (1994). *La psychométrie*. Paris : Presses Universitaires de France
- Dutton E., & Lynn R. (2013). A negative Flynn effect in Finland, 1997–2009. *Intelligence*, 41(6), 817–820. doi: 10.1016/j.intell.2013.05.008
- Ebel, R. L. (1977), Comments on the measurement theorist's dilemma, *Journal of Educational Measurement*, 14(2), 107–112.
- Eyde, L.D., Robertson, G.J., Krug, S.E. et al (1993). *Responsible Test Use : Case Studies For Assessing Human Behaviour*. Washington DC : American Psychological Association.
- Flora, D. B. (2020). Your Coefficient Alpha Is Probably Wrong, but Which Coefficient Omega Is Right? A Tutorial on Using R to Obtain Better Reliability Estimates. *Advances in Methods and Practices in Psychological Science*, 3(4), 484–501. <https://doi.org/10.1177/2515245920951747>
- Frank, L. K., & Macy, J. (1939). Projective Methods for the Study of Personality. *Transactions of the New York Academy of Sciences*, 1(8 Series II), 129–132. doi:10.1111/j.2164-0947.1939.tb00021.x
- Frank, L. K., & Macy, J. (1948). *Projective Methods*. Springfield III: Charles C Thomas
- Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist*, 18, 510–522. doi:10.1037/h0049294

- Glutting J. J., McDermott P. A., Stanley J.C. (1987). Resolving Differences among Methods of Establishing Confidence Limits for Test Scores, *Educational and Psychological Measurement*, 47(3), 607-614. doi : 10.1177/001316448704700307
- Gorsuch R.L. (1974). *Factor analysis*. Hillsdale, N.J., Lawrence Erlbaum Associates
- Gregory, R.J. (2010). *Psychological Testing : History, Principles, and Applications*(2nd edition). Allyn & Bacon : New-York.
- Groth-Marnat, G. (2009). *Handbook of Psychological Assessment* (5^{ème} édition). J.Wiley & Sons: New-York.
- Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sorbom (Eds.), *Structural equation modeling: Present and future* (pp. 195–216). Lincolnwood, IL: Scientific Software International
- Hattie, J. (1985). Methodology Review: Assessing Unidimensionality of Tests and Items. *Applied Psychological Measurement*, 9(2), 139–164. doi: 10.1177/014662168500900204
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179–185. doi : 10.1007/BF02289447
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 417-441, 498-520.
- Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: conceptual, methodological, and statistical issues. *Psychological Assessment*, 15(4), 446–455. doi:10.1037/1040-3590.15.4.446
- Iliescu, D. (2017). Norming. In D. Iliescu (Ed.), *Adapting Tests in Linguistic and Cultural Situations* (pp. 415–442). Cambridge: Cambridge University Press. doi: 10.1017/9781316273203.011
- Kahneman D. (2012). *Système 1 / Système 2 : Les deux vitesses de la pensée*, Flammarion
- Kieftenbeld, V., & Nandakumar, R. (2015). Alternative Hypothesis Testing Procedures for DIMTEST. *Applied Psychological Measurement*, 39(6), 480–493. doi: 10.1177/0146621615577618
- Massey, F. J. Jr. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46(253), 68–78. DOI: 10.2307/2280095
- Muneaux, M. (2018). *Petit guide de psychométrie clinique à l'usage des praticiens* De Boeck.
- Mcneish, D., & Mcneish, D. (2017). Psychological Methods Thanks Coefficient Alpha , We ' ll Take It From Here. *Psychological Methods*, 23(3), 412–433.
- Lacot, E., Barbeau, E. J., Thomas-Anterion, C., Basaglia-Pappas, S., Pariente, J., Puel, M., & Vautier, S. (2011). Le TOP 12: comment s'en servir pour repérer une pathologie du vieillissement cognitif? *Revue Neuropsychologique*, 3(4), 273–283. doi: 10.1684/nrp.2011.0188
- Laveault, D. (2012). Soixante ans de bons et mauvais usages du alpha de Cronbach. *Mesure et Évaluation En Éducation*, 35(2), 1-7. doi: 10.7202/1024716ar
- Lilliefors, H. W. (1967). On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318), 399–402.
- McArdle, J. J. (2007). John L. Horn (1928-2006). *American Psychologist*. 62 (6): 596–7. doi:10.1037/0003-066X.62.6.596
- Messick,, S. (1989). Validity In R. L. Linn (Ed.),. *Educational Measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan

- Messick, S. (1995). *Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry*. *American Psychologist*, 50(9), 741–749.
- Morin, V., Morin, J., Mercier, M., Moineau, M., & Codet, J. (1998). Les courbes ROC en biologie médicale. *Immuno-Analyse & Biologie Spécialisée*, 13(5), 279–286. doi: 10.1016/S0923-2532(98)80016-1
- Nandakumar, R., & Stout, W. (1993). Refinements of Stout's Procedure for Assessing Latent Trait Unidimensionality. *Journal of Educational Statistics*, 18(1), 41–68. doi: 10.3102/10769986018001041
- Nunnally, J. C., (1978). *Psychometric Theory* (2^e éd.). New York : McGraw-Hill.
- Osterlind, S. J. (2002). *Constructing Test Items* (2nd Ed, Vol. 47). Dordrecht: Kluwer Academic Publishers. doi: 10.1007/0-306-47535-9
- Pearson, K.(1904). Mathematical contribution to the theory of evolution XIII: On the theory of contingency and its relation to association and normal correlation. in *Drapers Company Research Memoirs, Biometric Series*, 1, 1-34.
- Pichot, P. (1997). *Les tests mentaux*. (15^{ème} ed.). Paris: Presses Universitaires de France.
- Piéron H. (1951). *Vocabulaire de la psychologie* (2003, dernière édition). Presses Universitaires de France.
- Pintea, S., & Moldovan, R. (2009). The receiver-operating characteristic (ROC) analysis: Fundamentals and applications in clinical psychology. *Journal of Cognitive and Behavioral Psychotherapies*, 9(1), 49–66.
- R Core Team. (2025). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>
- Reuchlin M. (1976). *Précis de statistique*. Paris, Presses Universitaires de France.
- Revelle, W. (2016). *An introduction to psychometric theory with applications in R*. Retrieved from <http://personality-project.org/r/book/>
- Revelle, W., Rocklin, T. (1979). Very simple structure alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research*, 14(4), 403-414.
- Rindermann, H., Becker, D., & Coyle, T. R. (2017). Survey of expert opinion on intelligence: The Flynn effect and the future of intelligence. *Personality and Individual Differences*, 106, 242–247. doi: 10.1016/j.paid.2016.10.061
- Rouxel, G. (1999). Modèles de Réponse à l'Item pour items polytomiques : exemple d'utilisation du logiciel MULTILOG. *Psychologie et Psychométrie*, 20(3), 113-130
- Rücker, G., & Schumacher, M. (2008). Simpson's paradox visualized: The example of the Rosiglitazone meta-analysis. *BMC Medical Research Methodology*, 8(1), 34. doi: 10.1186/1471-2288-8-34
- Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of a known factorial structure. *Psychological Assessment*, 24(2), 282-292
- Roy, A., Fournet N., Roulin, J.L., Le Gall, D. (2021). *La batterie FEE (évaluation des fonctions Exécutives chez l'enfant)*. Hogrefe France Editions, Paris.
- Sechrest, L. (1963). Incremental validity: A recommendation. *Educational and Psychological Measurement*, 23, 153–158
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428. doi: 10.1037/0033-2909.86.2.420

- Soloman, S. R., & Sawilowsky, S. S. (2009). Impact of Rank-Based Normalizing Transformations on the Accuracy of Test Scores. *Journal of Modern Applied Statistical Methods*, 8(2), 448–462. doi: 10.22237/jmasm/1257034080
- Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52(4), 589–617. doi: 10.1007/BF02294821
- Simpson, E. H. (1951), The Interpretation of Interaction in Contingency Tables, *Journal of the Royal Statistical Society, Ser. B(13)*, 238-241
- Spearman, C. (1904a). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1), 72. doi: 10.2307/1412159
- Spearman C. (1904b). General Intelligence Objectively Determined and Measured, *American Journal of Psychology*, 15, 201-292. Disponible à <http://psychclassics.yorku.ca/Spearman/>
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103, 677-680.
- Svetina, D., & Levy, R. (2014). A Framework for Dimensionality Assessment for Multidimensional Item Response Models. *Educational Assessment*, 19(1), 35–57. doi: 10.1080/10627197.2014.869450
- Tabachnick, B. G., and Fidell, L. S. (2013). *Using Multivariate Statistics* (6th ed). Boston : Pearson.
- Tallent N. (1993). *Psychological report writing* (4th ed.). Upper Saddle River, NJ: Prentice Hall.
- Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. *BMC Medical Education*, 9(1), 1–8. doi: 10.1186/1472-6920-9-40
- Tate, R. (2003). A Comparison of Selected Empirical Methods for Assessing the Structure of Responses to Test Items. *Applied Psychological Measurement*, 27(3), 159–203. doi: 10.1177/0146621603027003001
- Thomas, H. (2009). Discovering Lightner Witmer : A Forgotten Hero of Psychology. *Journal of Scientific Psychology*, (April), 3–13. retrieved from : http://www.psyencelab.com/uploads/5/4/6/5/54658091/discovering_lightner_witmer.pdf
- Thurstone, L. L. (1935), *The Vectors of Mind*, University of Chicago Press
- Toksöz, S., & Ertunç, A. (2017). Item Analysis of a Multiple-Choice Exam. *Advances in Language and Literary Studies*, 8(6), 141. doi: 10.7575/aiac.all.v.8n.6p.141
- Tong, Y., & Kolen, M. J. (2005). Conditional Standard Errors of Measurement. In *Encyclopedia of Statistics in Behavioral Science* (Vol. 29, pp. 285–307). Chichester, UK: John Wiley & Sons, Ltd. doi: 10.1002/0470013192.bsa196
- Thurstone, L.L (1931). Multiple factor analysis. *Psychological Review*, 38, 406–427
- Van Der Linden, W. J. (2010). Item response theory. *International Encyclopedia of Education*, 4, 81–88. doi: 10.1016/B978-0-08-044894-7.00250-5
- Vautier (2014) 8a. Le kappa de Cohen : une solution à un faux problème, in *Épistémologie de la psychologie*, 21/04/2014, <https://epistemo.hypotheses.org/715>.
- Vautier, S. (2015). La psychotechnique des aptitudes : pour différencier une sociotechnique de l'évaluation sans mesurage et une psychologie balbutiante de la compréhension de la performance. *Pratiques Psychologiques*, 21, 1-18.
- Vautier (2017, January 23). Carnet d'enseignement et de recherche de Stéphane Vautier. in *hypotheses.org* [online]. Page consultée le 23 février 2017. <http://epistemo.hypotheses.org/cours-video>

- Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 31, 321–327
- Velleman, Paul F., & Leland Wilkinson (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician*, 47(1), 65-721. doi : 10.1080/00031305.1993.10475938
- Velicer, W., Eaton, C., & Fava, J. (2000). Construct Explication through Factor or Component Analysis: A Review and Evaluation of Alternative Procedures for Determining the Number of Factors or Components. In R. Goffin & E. Helmes (Eds.), *Problems and Solutions in Human Assessment SE - 3* (pp. 41–71). Springer US. doi:10.1007/978-1-4615-4397-8_3
- Vrignaud, P. (2006). La mesure de la littéracie dans PISA : la méthodologie est la réponse, mais quelle était la question ? *Revue Française de Pédagogie*, 157, 27–41. doi: 10.4000/rfp.409
- Warner, W. L. (1960). *Social class in America: A manual of procedure for the measurement of social status*. Science Research Associates.
- Watkins, M. W. (2017). The reliability of multidimensional neuropsychological measures : From alpha to omega. *The Clinical Neuropsychologist*, 31, 1113-1126. <https://doi.org/10.1080/13854046.2017.1317364>
- Watkins, M. W. (2018). Exploratory Factor Analysis: A Guide to Best Practice *Journal of Black Psychology*, 44(3), 219–246. <https://doi.org/10.1177/0095798418771807>
- Wechsler, D. (2005). *Manuel d'administration et de cotation du WISC-IV*. Paris: Editions du Centre de Psychologie Appliquée.
- Wechsler, D. (2016). *Manuel d'administration et de cotation du WISC-V*. Paris : Edition du Centre de Psychologie Appliquée.
- Wolber, G.J., Carne W.F. (2002). *Writing psychological reports. A guide for clinicians* (2nd édition). Petersburg, VA : Professionnal Ressources Press.
- Wang, X. (2004). Le baccalauréat, pivot du système éducatif chinois. *Revue Internationale D'éducation de Sèvres*, (37), 61–69. doi: 10.4000/ries.1400
- Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α , Revelle's β , and McDonald's ω_H : their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, 70(1), 123–133. doi: 10.1007/s11336-003-0974-7

Documents sur Internet

- Ricco Rakotomalala (2012). *Analyse de corrélation*. Cours en économétrie, Université de Lyon. Page consultée en décembre 2016. http://eric.univ-lyon2.fr/~ricco/cours/cours/Analyse_de_Correlation.pdf

Quelques adresses électroniques de sites Internet

Guidelines

- ITC Guidelines on Quality Control in Scoring, Test Analysis, and Reporting of Test Scores. (2014). *International Journal of Testing*, 14(3), 195–217. <http://doi.org/10.1080/15305058.2014.918040> (lien : https://www.intestcom.org/files/ijt_qc_guidelines.pdf)
- International Test Commission (2001) International Guidelines for Test Use, *International Journal of Testing*, 1:2, 93-114, doi: 10.1207/S15327574IJT0102_1. (lien : http://dx.doi.org/10.1207/S15327574IJT0102_1)

Autres

Riandey, B., & Widmer, I. (2009). *Introduction aux sondages à l'usage du plus grand nombre* [Electronic Version], 13. Retrieved 2010. http://statistix.fr/IMG/pdf/Riandey-Widmer_sondages_.pdf

[International test commission](http://www.intestcom.org) : (<http://www.intestcom.org>) Recommandations internationales sur l'usage des tests.

[Buros Institute of Mental Measurements](http://www.unl.edu/buros/) (<http://www.unl.edu/buros/>) : Site officiel de l'Institut Buros qui est spécialisé dans la publication d'analyses critiques de pratiquement tous les tests sur le marché. On y trouve également des informations sur la meilleure façon d'utiliser les publications «Mental Measurements Yearbook» (MMY) et «Tests in Print».

[American Psychological Association. Finding information about psychological tests](http://www.apa.org/science/testing.html) [<http://www.apa.org/science/testing.html>]. Cette page W3 préparée par l'American Psychological Association présente une foule d'informations que tout étudiant en psychologie devrait savoir au sujet des tests psychologiques. Vous pouvez regarder par exemple la rubrique *The Rights and Responsibilities of Test Takers: Guidelines and Expectations* (<http://www.apa.org/science/ttrr.html>) ou encore les FAQ (*cf.ci-dessous*)

[FAQ/Finding Information About Psychological Tests](http://www.apa.org/science/programs/testing/find-tests.aspx) (www.apa.org/science/programs/testing/find-tests.aspx) : Traite d'abord des tests qui ont été publiés et, ensuite, des tests qui n'ont pas fait l'objet d'une publication. Les principaux répertoires de tests sont présentés. Ce site est particulièrement utile pour connaître les outils permettant de trouver un test sur tel ou tel sujet ou de trouver plus d'information sur un test déjà identifié.

[Les tests d'intelligence et la mesure de l'esprit](http://www.canal-u.education.fr/canalu) (sur <http://www.canal-u.education.fr/canalu>). Conférence de Jacques Lautrey qui est un excellent résumé du cours de psychologie différentielle de 2^{ème} année. A ne pas manquer. Il est conseillé aussi de s'intéresser aux questions typiques du public (fin de conférence) et à la façon dont un psychologue doit répondre quant on parle des tests d'intelligence et de QI.

<http://www.indiana.edu/~intell/index.html> (Plucker, J. A. (Ed.). (2003). *Human intelligence: Historical influences, current controversies, teaching resources*. Retrieved [insert month day, year], from <http://www.indiana.edu/~intell>) : Excellent site concernant les théories de l'intelligence et leur histoire (Christopher D. Green, York University, Toronto, Canada). Une brève biographie des principaux chercheurs est présentée ainsi qu'une carte générale décrivant les influences entre les principaux chercheurs. On peut aussi accéder sur ce site à des articles originaux ayant marqués l'histoire de la psychologie, consulter ou participer à un forum sur des questions de la psychologie, etc. Des questions d'actualité ou des controverses importantes sont bien traitées. Par exemple : l'effet Flynn, la controverse de Wissler, The belle curve, La famille Kallikak (Goddard), la théorie multiple de l'intelligence, les enfants doués et les théories de l'intelligence, The task force report (APA).

Associations Nationales en langue française :

- [La SFP](http://www.sfpsy.org/) (<http://www.sfpsy.org/>). Site de la Société Française de Psychologie.
- [La SSP](http://www.psyweb.ch/) (<http://www.psyweb.ch/>). Site de la Société Suisse de Psychologie.
- [La FBP](https://www.bfp-fbp.be/fr/psychologie-en-belgique) (<https://www.bfp-fbp.be/fr/psychologie-en-belgique>). Site de la Société Belge de Psychologie
- [La SCP](https://www.cpa.ca/fr/) (<https://www.cpa.ca/fr/>). Site de la Société Canadienne de Psychologie.

→ [La SLB](https://www.slp.lu/fr/) (https://www.slp.lu/fr/). Site de la société Luxembourgeoise de Psychologie

Fédérations et commissions Internationales :

→ [L'EFPA](http://www.efpa.eu/) (http://www.efpa.eu/) European Federation of Psychologists' Associations

→ [L'ITC](https://www.intestcom.org/) (https://www.intestcom.org/). L'Internationale Test Commission.