COURS DE CONSTRUCTION ET ADAPTATION DES TESTS PSYCHOLOGIQUE EN MILIEU PROFESSIONNEL

Caractéristiques Psychométriques

des Échelles Psychologiques

Validité — Fidélité — Normes

Spécialité : Psychologie du travail et organisation

Niveau : Licence 3

Année universitaire 2025-2026

Par: Dr. YOUCEF KHODJA Adil

Introduction

L'utilisation d'une échelle psychologique ne saurait se réduire à la simple administration d'un questionnaire et au calcul d'un score. Tout instrument de mesure psychologique doit satisfaire à des critères scientifiques stricts qui garantissent la qualité et la légitimité des informations qu'il produit. Ces critères, rassemblés sous l'expression de « caractéristiques psychométriques », constituent le fondement de toute évaluation psychologique rigoureuse et éthiquement responsable (Laveault & Grégoire, 2002).

Les trois piliers de la qualité psychométrique d'une échelle sont la validité, la fidélité et la normalisation. La validité répond à la question fondamentale : « L'échelle mesure-t-elle bien ce qu'elle prétend mesurer ? ». La fidélité s'interroge sur la stabilité et la cohérence des mesures produites : « Obtient-on des résultats similaires si l'on mesure plusieurs fois le même phénomène dans des conditions équivalentes ? ». Les normes, enfin, permettent de donner un sens comparatif aux scores : « Que signifie un score de 45 sur cette échelle par rapport à la population de référence ? » (Bertrand & Blais, 2004 ; Fortin, 2010).

Ce cours présente de manière approfondie et structure ces trois caractéristiques, en s'appuyant exclusivement sur des références francophones ou traduites en français. Chaque notion est définie conceptuellement, illustrée par des exemples concrets et située dans le contexte de la pratique psychométrique contemporaine. La maitrise de ces concepts est indispensable pour tout psychologue souhaitant utiliser, adapter ou construire des outils d'évaluation dans les différents champs de la psychologie appliquée.

Chapitre I : La validité des échelles psychologiques

1.1. Définition générale et fondements conceptuels

La validité d'une échelle psychologique est la propriété la plus fondamentale et la plus exigeante de tout instrument de mesure. Elle désigne le degré auquel l'échelle mesure effectivement le construit théorique qu'elle est censée mesurer, et non une autre variable, voisine ou parasite (Vallerand & Hess, 2000). Un instrument peut être parfaitement fidèle — c'est-à-dire produire des résultats stables et reproductibles — sans pour autant être valide, si ces résultats stables reflètent une variable autre que celle visée.

Anastasi (trad. Juhel, 1994, p. 149) définit la validité comme « le degré auquel le test mesure ce qu'il est censé mesurer ». Cette définition, bien que simple en apparence, recouvre une réalité complexe et multidimensionnelle : la validité n'est pas une propriété absolue et binaire (un test est valide ou ne l'est pas), mais un degré qui s'évalue à travers plusieurs sources de preuves convergentes, dans un contexte d'utilisation spécifique et pour une population définie (Laveault & Grégoire, 2002).

Il est important de souligner que la validité ne se démontre pas une fois pour toutes : elle s'accumule progressivement au fil des études et des usages. Un instrument validé pour une population adulte francophone n'est pas nécessairement valide pour des adolescents, pour une population arabophone ou pour un contexte culturel différent, comme c'est souvent le cas en Algérie (Mokrani, 2015). La validité est donc toujours contextuelle, populationnelle et finaliste.

1.2. Les grandes catégories de validité

1.2.1. La validité de contenu

Définition et enjeux

La validité de contenu évalue dans quelle mesure les items de l'échelle représentent de manière adéquate et exhaustive l'ensemble du domaine conceptuel du construit mesuré (Fortin, 2010). Elle est qualitative par nature et repose sur le jugement d'experts plutôt que sur des analyses statistiques. Une échelle mesurant l'anxiété aura une bonne validité de contenu si ses items couvrent les différentes composantes du construit : la composante

cognitive (pensées intrusives, inquiétudes), la composante somatique (tensions musculaires, palpitations) et la composante comportementale (évitement, inhibition).

L'indice de validité de contenu (IVC)

La procédure la plus rigoureuse pour évaluer la validité de contenu est le recours à un panel de juges experts qui évaluent chaque item sur une échelle en quatre points (1 = pas du tout pertinent ; 4 = très pertinent). L'Indice de Validité de Contenu (IVC) calcule la proportion de juges ayant attribué une note de 3 ou 4 à chaque item (IVC par item) et à l'ensemble de l'instrument (IVC global). Lynn (1986, cité dans Fortin, 2010) recommande un IVC par item supérieur à 0,78 avec un panel de 6 experts ou plus, et un IVC global supérieur ou égal à 0,90.

Selon Laveault et Grégoire (2002), la validité de contenu doit également éviter deux écueils opposés : (1) la sous-représentation du construit, où l'échelle ne couvre qu'une partie de ses dimensions (ex. : une échelle de dépression qui ne mesure que les aspects cognitifs en omettant les aspects somatiques) ; et (2) la contamination du construit, où l'échelle inclut des items qui mesurent en réalité un construit différent ou adjacent (ex. : des items mesurant l'anxiété dans une échelle censée être spécifique à la dépression).

1.2.2. La validité de construit

Définition et place centrale

La validité de construit est aujourd'hui considérée comme la forme suprême de validité, car elle englobe toutes les autres formes dans un cadre théorique intégrateur (Messick, 1995, cité dans Laveault & Grégoire, 2002). Elle répond à la question : « Dans quelle mesure les scores de l'échelle reflètent-ils fidèlement le construit théorique postulé, tel qu'il est défini dans la littérature scientifique ? ». Pour Vallerand et Hess (2000), la validité de construit se démontre par un ensemble de preuves empiriques convergentes plutôt que par une seule analyse.

La validité factorielle

La validité factorielle évalue si la structure dimensionnelle de l'échelle, mise en évidence par l'Analyse Factorielle Exploratoire (AFE) ou Confirmatoire (AFC), correspond à la structure théorique du construit. Une échelle conçue pour mesurer trois dimensions distinctes de l'estime de soi (dimensions scolaire, sociale et physique) devrait produire, lors

de l'analyse factorielle, trois facteurs clairement distincts regroupant respectivement les items de chaque dimension (Bertrand & Blais, 2004).

Dans le cadre de l'AFC, l'adéquation du modèle théorique aux données est évaluée par plusieurs indices d'ajustement (Laveault & Grégoire, 2002) :

- Le chi-deux rapporté au degré de liberté ($\chi 2$/dl) : une valeur inférieure à 3 indique un ajustement satisfaisant.
- Le RMSEA (Root Mean Square Error of Approximation) : une valeur inférieure à 0,06 indique un bon ajustement ; inférieure à 0,08, un ajustement acceptable.
- Le CFI (Comparative Fit Index) et le TLI (Tucker-Lewis Index) : des valeurs supérieures à 0,95 indiquent un bon ajustement.
- Le SRMR (Standardized Root Mean Square Residual) : une valeur inférieure à 0,08 est généralement jugée acceptable.

La validité convergente

La validité convergente établit que les scores de l'échelle corrèlent positivement et significativement avec les scores d'autres instruments mesurant des construits théoriquement proches ou liés. Par exemple, une nouvelle échelle d'anxiété généralisée devrait corréler fortement avec l'Inventaire d'Anxiété de Beck (BAI, traduit et validé en français par Bourque & Beaudette, 1982) et modérément avec des échelles de neuroticisme ou de stress perçu (Vallerand & Hess, 2000).

Selon les critères de Campbell et Fiske (1959, cités dans Vallerand & Hess, 2000), des corrélations convergentes de l'ordre de $r \geq 0{,}50$ à 0,70 entre instruments mesurant des construits voisins sont généralement jugées satisfaisantes, à condition que les instruments utilisent des formats de réponse ou des méthodes de passation différents (afin d'exclure les artefacts de méthode communs).

La validité discriminante (ou divergente)

La validité discriminante est l'image en miroir de la validité convergente : elle établit que les scores de l'échelle ne corrèlent pas, ou très faiblement, avec les scores d'instruments mesurant des construits théoriquement indépendants ou orthogonaux (Vallerand & Hess, 2000). Ainsi, une échelle d'intelligence fluide ne devrait pas corréler significativement avec une échelle d'extraversion, sauf si la théorie prévoit un lien entre ces deux construits.

Dans le cadre de la modélisation par équations structurales (MES), la validité discriminante peut être évaluée quantitativement par la Variance Moyenne Extraite (VME ou AVE) : si la VME d'un facteur est supérieure au carré de sa corrélation avec tout autre facteur, la validité discriminante est établie (Fornell & Larcker, 1981, cités dans Bertrand & Blais, 2004). Une VME supérieure à 0,50 indique par ailleurs que le facteur capte plus de variance des items qu'il n'en laisse aux erreurs de mesure.

La validité de trait et la validité de méthode : la matrice MTMM

La procédure la plus complète pour évaluer simultanément la validité convergente et discriminante est la Matrice Multi-Traits Multi-Méthodes (MTMM) proposée par Campbell et Fiske (1959, cités dans Vallerand & Hess, 2000). Cette procédure consiste à mesurer plusieurs construits (traits) avec plusieurs méthodes différentes (auto-évaluation, évaluation par autrui, observation comportementale), puis à analyser la matrice de corrélations obtenue selon les principes suivants :

- Les corrélations convergentes (même trait, méthodes différentes) doivent être élevées.
- Les corrélations discriminantes (traits différents, même méthode ou méthodes différentes) doivent être faibles.
- Les corrélations convergentes doivent dépasser les corrélations discriminantes correspondantes.

1.2.3. La validité de critère

Définition générale

La validité de critère évalue la capacité des scores de l'échelle à être associés ou à prédire un critère externe pertinent, mesuré indépendamment de l'instrument évalué (Anastasi, trad. Juhel, 1994). Le critère peut être un autre test reconnu comme étalon d'or, une évaluation clinique expérte, un comportement observable ou une issue mesurable dans le temps.

La validité concomitante

La validité concomitante est établie lorsque les scores de l'échelle corrèlent significativement avec un critère mesuré au même moment ou dans un intervalle de temps très court. Elle est particulièrement importante lors de la validation d'un nouvel instrument comme alternative ou substitut à un instrument déjà reconnu (Chartier & Morin, 2009). Par exemple, pour valider une nouvelle échelle de mesure de la dépression en contexte algérien,

on peut établir sa validité concomitante en la corrélant avec le diagnostic psychiatrique clinique établi par un entretien structuré.

La validité prédictive

La validité prédictive est établie lorsque les scores de l'échelle permettent de prédire significativement un critère mesuré dans le futur. Elle est cruciale dans les contextes d'orientation professionnelle et scolaire, de sélection et de pronostic clinique (Bernaud, 2007). Un test d'aptitude professionnelle doit prédire les performances au travail évaluées six mois plus tard ; une échelle de dépistage du burnout doit prédire l'arrêt de travail ou la démission dans les mois qui suivent. Le coefficient de validité prédictive est généralement un coefficient de corrélation de Pearson entre le score de l'échelle et le critère futur.

La validité incrémentielle

La validité incrémentielle, plus exigeante encore, évalue si un nouvel instrument apporte une contribution prédictive supplémentaire par rapport à celle déjà fournie par les instruments existants. Elle est évaluée par des analyses de régression hiérarchique dans lesquelles on introduit d'abord les instruments de référence (bloc 1), puis l'instrument en cours de validation (bloc 2), et l'on teste si l'ajout du nouvel instrument augmente significativement le R^2 prédictif (Vallerand & Hess, 2000). Si tel est le cas, le nouvel instrument apporte une information unique et non redô vis-à-vis des outils existants.

1.2.4. La validité écologique et la validité culturelle

La validité écologique

La validité écologique concerne la mesure dans laquelle les résultats obtenus en situation standardisée (dans un cabinet, un bureau ou une salle de classe) reflètent les comportements réels des individus dans leur environnement naturel (Fortin, 2010). Une échelle d'évaluation du fonctionnement social chez des patients psychiatriques aura une bonne validité écologique si ses scores corrèlent avec les observations directes du comportement de ces patients dans leur milieu de vie quotidienne.

La validité culturelle et l'invariance de mesure

La validité culturelle questionne l'applicabilité d'un instrument à des populations de cultures différentes. Lorsqu'on traduit et adapte une échelle d'une culture à une autre, il ne suffit pas de vérifier la qualité de la traduction : il faut également s'assurer que le construit mesuré a la même signification dans les deux cultures (invariance de mesure) (Bertrand &

Blais, 2004). L'invariance de mesure est évaluée par une série d'AFC multi-groupes testées selon une progression hiérarchique d'hypothèses :

- L'invariance configurationnelle : la même structure factorielle s'applique dans les deux groupes.
- L'invariance métrique : les saturations factorielles sont égales entre les groupes (permettant de comparer les corrélations).
- L'invariance scalaire : les interceptes des items sont égaux entre les groupes (permettant de comparer les moyennes latentes).
- L'invariance stricte : les variances des erreurs de mesure sont égales entre les groupes.

En Algérie, cette problématique est d'une importance cruciale : de nombreux instruments traduits de l'anglais ou du français standard sont utilisés sans vérification de leur invariance culturelle, ce qui fragilise la comparabilité des résultats et peut produire des interprétations erronées (Mokrani, 2015).

1.3. Tableau synthétique des types de validité

Type de validité	Question clé	Méthodes d'évaluation
Contenu	L'échelle couvre-t-elle bien tout le domaine du construit ?	Panel de juges, IVC ≥ 0,78 par item, ≥ 0,90 global
Factorielle	La structure dimensionnelle correspond-elle à la théorie ?	AFE, AFC (RMSEA < 0,06 ; CFI > 0,95)
Convergente	Les scores corrèlent-ils avec des instruments proches ?	Corrélations de Pearson r ≥ 0,50–0,70 ; VME > 0,50
Discriminante	Les scores se distinguent-ils des construits non liés ?	Matrice MTMM, VME > r^2 inter-facteurs
Concomitante	Les scores sont-ils associés à un critère actuel ?	Corrélation avec étalon d'or, diagnostic clinique
Prédictive	Les scores prédisent-ils un critère futur ?	Corrélation avec critère longitudinal, régression
Incrémentielle	L'instrument apporte-t-il une information unique ?	Régression hiérarchique, ΔR^2 significatif
Culturelle	L'instrument est-il applicable à d'autres cultures ?	AFC multi-groupes, tests d'invariance de mesure

Tableau 1. Synthèse des types de validité d'une échelle psychologique (d'après Vallerand & Hess, 2000 ; Laveault & Grégoire, 2002 ; Bertrand & Blais, 2004).

1.4. Les menaces à la validité

Plusieurs facteurs peuvent menacer la validité d'une échelle et biaiser les scores obtenus. Laveault et Grégoire (2002) et Vallerand et Hess (2000) distinguent les principales menaces suivantes :

- Le biais de désirabilité sociale : la tendance des répondants à donner des réponses socialement valorisées plutôt que des réponses authentiques. Ce biais est particulièrement prégnant dans les contextes évaluatifs (sélection, expertise) et dans les cultures où la préservation de l'image de soi est une valeur forte.
- Le biais d'acquiescement : la tendance à approuver systématiquement les items, quelle que soit leur formulation. Pour contrôler ce biais, il est recommandé d'inclure des items formulés en sens inverse (items inversés).
- Le biais de réponse extrême ou de tendance centrale : la tendance à utiliser systématiquement les extrémités ou le milieu de l'échelle de réponse, indépendamment du contenu des items.
- L'effet de halo : dans les échelles d'observation ou d'évaluation par autrui, la tendance de l'évaluateur à laisser son impression générale sur la personne colorée l'ensemble de ses jugements spécifiques.
- La contamination de la mesure par des variables de nuisance : factigue, anxiété situationnelle, bruit, incompréhension des consignes, etc.

Chapitre II : La fidélité des échelles psychologiques

2.1. Définition générale et modèle théorique

La fidélité (ou fiabilité) d'une échelle psychologique désigne le degré auquel les mesures produites par l'instrument sont stables, cohérentes et reproductibles dans le temps et entre les observateurs (Bertrand & Blais, 2004). Un instrument fidèle produit des résultats similaires lorsqu'il est appliqué à la même personne dans des conditions équivalentes, à des moments différents ou par des évaluateurs différents.

Le cadre théorique de référence pour comprendre la fidélité est la Théorie Classique des Tests (TCT), formalisée notamment par Gulliksen (1950) et Cronbach (1951), cités dans Laveault et Grégoire (2002). Selon la TCT, tout score observé (X) se décompose en deux composantes : un score vrai (T, reflet de la vraie valeur du trait mesuré) et une erreur de mesure (E), selon la formule fondamentale :

$$\mathbf{X = T + E}$$

La fidélité est alors définie comme la proportion de la variance totale des scores observés qui est due à la variance des scores vrais, plutôt qu'aux erreurs de mesure. Plus cette proportion est élevée, plus l'instrument est fidèle. Formellement, le coefficient de fidélité ρ est défini par :

$$\mathbf{\rho = \sigma^2T / \sigma^2X = \sigma^2T / (\sigma^2T + \sigma^2E)}$$

où σ^2T est la variance des scores vrais et σ^2X est la variance totale des scores observés (Laveault & Grégoire, 2002). Ce coefficient varie entre 0 (mesure entièrement dominée par l'erreur) et 1 (mesure parfaitement exempte d'erreur). Dans la pratique, les instruments psychologiques se situent généralement entre 0,70 et 0,95.

2.2. Les différentes formes de fidélité

2.2.1. La cohérence interne

Le coefficient alpha de Cronbach

La cohérence interne est la forme de fidélité la plus fréquemment rapportée dans les études de validation d'échelles. Elle évalue le degré d'intercorrélation des items d'une même

dimension, c'est-à-dire dans quelle mesure ils mesurent tous le même construit latent (Chartier & Morin, 2009). Le coefficient alpha de Cronbach (α) est l'estimateur standard de la cohérence interne :

$$\alpha = (k / k{-}1) \times (1 - \Sigma\sigma^2 i / \sigma^2 X)$$

où k est le nombre d'items, $\Sigma\sigma^2 i$ est la somme des variances individuelles des items, et $\sigma^2 X$ est la variance totale du score composite (Bertrand & Blais, 2004). Les seuils d'interprétation de l'alpha sont les suivants (Chartier & Morin, 2009 ; Laveault & Grégoire, 2002) :

Valeur de l'alpha (α)	Interprétation
$\alpha \geq 0{,}90$	Excellent — recommandé pour la prise de décision individuelle clinique
$0{,}80 \leq \alpha < 0{,}90$	Très bon — acceptable pour l'évaluation individuelle
$0{,}70 \leq \alpha < 0{,}80$	Acceptable — suffisant pour la recherche de groupe
$0{,}60 \leq \alpha < 0{,}70$	Discutable — à justifier selon le contexte
$\alpha < 0{,}60$	Insuffisant — révision de l'instrument nécessaire

Tableau 2. Seuils d'interprétation du coefficient alpha de Cronbach (d'après Chartier & Morin, 2009 ; Laveault & Grégoire, 2002).

Limites de l'alpha et alternatives

Malgré sa popularité, l'alpha de Cronbach présente plusieurs limites importantes (Bertrand & Blais, 2004). Premièrement, il est sensible au nombre d'items : un instrument comportant de nombreux items tend mécaniquement à produire des alphas élevés, même si les items sont peu corrélés entre eux. Deuxièmement, il suppose que tous les items contribuent de manière égale au construit (modèle tau-équivalent), hypothèse rarement vérifiée en pratique.

L'oméga de McDonald (ω) est aujourd'hui recommandé comme alternative plus robuste, notamment lorsque les saturations factorielles des items sont hétérogènes (Chartier & Morin, 2009). Contrairement à l'alpha, l'oméga utilise les saturations factorielles de chaque item estimées par analyse factorielle confirmatoire, ce qui lui confère une estimation plus précise de la cohérence interne réelle de l'instrument.

2.2.2. La fidélité test-retest

Principe et procédure

La fidélité test-retest évalue la stabilité temporelle des scores produits par l'échelle, c'est-à-dire le degré auquel les mêmes individus obtiennent des scores similaires lors de deux passations successives, séparées par un intervalle de temps (Vallerand & Hess, 2000). Elle est évaluée par le coefficient de corrélation de Pearson (ou le Coefficient de Corrélation Intraclasse, CCI) entre les scores obtenus aux deux passations.

L'intervalle de temps entre les deux passations doit être soigneusement ajusté en fonction du construit mesuré (Fortin, 2010) :

- Trop court (moins d'une semaine) : risque de contamination par les effets de mémoire et d'apprentissage — les répondants se souviennent de leurs réponses précédentes.
- Trop long (plus de trois mois) : risque que des changements réels survenus dans le trait mesuré soient interprétés comme des erreurs de mesure.
- Intervalle recommandé : 2 à 6 semaines pour les traits stables (personnalité, aptitudes) ; 1 à 2 semaines pour les états transitoires (humeur, anxiété situation-nelle).

Critères d'interprétation

Un coefficient de fidélité test-retest supérieur à 0,80 est généralement considéré comme satisfaisant pour les échelles mesurant des traits relativement stables (Laveault & Grégoire, 2002). Des valeurs entre 0,70 et 0,80 sont jugees acceptables, tandis que des valeurs inférieures à 0,70 suggèrent soit une instabilité excessive de l'échelle, soit que le construit mesuré est lui-même de nature fluctuante. Il convient de noter que pour les échelles mesurant des états transitoires (moods, symptomatologie aiguë), des corrélations test-retest plus faibles peuvent être attendues et tolérées.

2.2.3. La fidélité par formes parallèles

La méthode des formes parallèles consiste à élaborer deux versions équivalentes de l'instrument (Form A et Form B), comportant le même nombre d'items, couvrant le même domaine de contenu et présentant des propriétés statistiques identiques (même moyenne, même écart-type, mêmes intercorrélations entre items). On administre ensuite les deux formes au même groupe de sujets, soit simultanément, soit à un court intervalle, et on calcule la corrélation entre les scores des deux formes (Bertrand & Blais, 2004).

Si les deux formes sont véritablement parallèles, cette corrélation fournit une estimation épurée de la fidélité, évitant les biais de mémoire du test-retest. En pratique, cette méthode est peu utilisée en raison de la difficulté à construire deux formes véritablement équivalentes et du coût élevé que cela représente (Chartier & Morin, 2009). Elle trouve cependant des applications importantes dans les contextes d'évaluations certifiées où plusieurs versions d'un test sont nécessaires pour prévenir la triche ou le partage des réponses.

2.2.4. La fidélité inter-juges

Définition et contextes d'application

La fidélité inter-juges (ou accord inter-évaluateurs) est pertinente pour les instruments qui requièrent le jugement ou l'observation d'un ou plusieurs évaluateurs humains, plutôt qu'une réponse directe du sujet évalué (Fortin, 2010). Elle s'applique notamment aux échelles d'observation comportementale, aux grilles d'évaluation clinique, aux tests projectifs (Rorschach, TAT) et aux échelles d'évaluation des travaux scolaires ou professionnels.

Le Kappa de Cohen

Pour les données catgorielles (nominales ou ordinales), l'accord inter-juges est évalué par le coefficient Kappa de Cohen (κ), qui corrige l'accord observé pour l'accord attendu par le seul fait du hasard (Fortin, 2010). Il est défini par la formule :

$$\kappa = (Po - Pe) / (1 - Pe)$$

où Po est la proportion d'accord observé et Pe est la proportion d'accord attendue par hasard. Les seuils d'interprétation proposés par Landis et Koch (1977, cités dans Fortin, 2010) sont : $\kappa < 0{,}20$ (accord médiocre) ; 0,21–0,40 (accord léger) ; 0,41–0,60 (accord modéré) ; 0,61–0,80 (accord substantiel) ; 0,81–1,00 (accord presque parfait).

Le Coefficient de Corrélation Intraclasse (CCI)

Pour les données continues (intervalles ou rapports), la fidélité inter-juges est évaluée par le Coefficient de Corrélation Intraclasse (CCI), qui prend en compte non seulement la corrélation entre les évaluateurs, mais aussi les différences de niveau systmatiques entre eux (biais) (Bertrand & Blais, 2004). Un CCI supérieur à 0,75 est généralement considéré comme excellent pour une application clinique ; entre 0,60 et 0,74, il est jugé bon ; entre 0,40 et 0,59, modéré (Fortin, 2010).

2.3. L'erreur standard de mesure (ESM)

L'Erreur Standard de Mesure (ESM) est un indice complémentaire à la fidélité qui exprime l'erreur de mesure dans les mêmes unités que le score observé, permettant une interprétation directement applicable à l'évaluation individuelle (Laveault & Grégoire, 2002). Elle est calculée par la formule :

$$\mathbf{ESM = \sigma X \times \sqrt{(1 - \alpha)}}$$

où σX est l'écart-type des scores observés et α est le coefficient de fidélité. L'ESM permet de construire des intervalles de confiance autour des scores individuels : le score vrai d'un individu a approximativement 95 % de chances de se situer dans l'intervalle [score observé ± 1,96 × ESM] (Laveault & Grégoire, 2002). Cette notion est indispensable pour une interprétation clinique rigoureuse des résultats individuels, qui évite de traiter un score ponctuel comme une valeur absolue.

Par exemple, si un étudiant obtient un score de 75 à un test d'aptitudes dont l'ESM est de 3 points, son score vrai probable se situe dans l'intervalle [75 ± 1,96 × 3], soit entre 69,1 et 80,9, avec un niveau de confiance de 95 %. Cela signifie qu'il serait erronné d'interpréter la différence entre deux scores individuels comme significative si elle est inférieure à deux fois l'ESM (Bernaud, 2007).

2.4. Relations entre fidélité et validité

La fidélité et la validité entretiennent une relation logique asymmétrique fondamentale : la fidélité est une condition nécessaire mais non suffisante de la validité (Vallerand & Hess, 2000). En d'autres termes :

- Un instrument peut être fidèle sans être valide : il produit des scores stables et cohérents, mais qui ne reflètent pas le construit visé. Par exemple, un chronometre mesure le temps de manière très fidèle, mais il ne mesure pas l'intelligence, même si on l'utilisait à cet effet.
- Un instrument valide est nécessairement fidèle dans une certaine mesure : des scores entièrement dominés par l'erreur de mesure ne peuvent pas reflettervalablement un construit théorique.
- Il existe un plafond mathématique : le coefficient de validité ne peut dépasser la racine carrée du coefficient de fidélité. Ainsi, si α = 0,64, la corrélation maximale possible avec un critère externe ne peut excéder 0,80 (Laveault & Grégoire, 2002).

Chapitre III : Les normes des échelles psychologiques

3.1. Définition et nécessité des normes

Un score brut obtenu à une échelle psychologique n'a, en lui-même, aucune signification intrinsèque. Savoir qu'un individu a obtenu un score de 42 à une échelle d'anxiété ne permet pas de déterminer si ce score est faible, moyen ou élevé, ni de prendre des décisions cliniques ou orientationnelles fondées. C'est précisément à cette limite que répondent les normes : elles permettent de situer un score individuel dans une distribution de référence représentative d'une population définie (Anastasi, trad. Juhel, 1994).

Laveault et Grégoire (2002, p. 213) définissent les normes comme « des données statistiques permettant de situer la performance d'un sujet par rapport à la performance d'un groupe représentatif défini comme groupe de référence ou groupe normatif ». Bernaud (2007) souligne que les normes transforment un score brut (absolument arbitraire) en un score relatif (porteur d'une signification comparative), condition indispensable à toute décision psychologique fondée sur un test.

La normalisation (ou étalonnage) désigne l'ensemble de la procédure par laquelle sont collectées les données normatives, calculées les statistiques de référence et élaborées les tables de conversion permettant de passer des scores bruts aux scores normés. C'est une étape finale et indispensable de la construction de tout instrument destindé à une utilisation clinique ou évaluative (Chartier & Morin, 2009).

3.2. Les types de normes

3.2.1. Les normes de développement

Les normes de développement situent les performances d'un individu par rapport à des résultats typiques d'individus à différents stades de développement, généralement exprimés en âges ou en niveaux scolaires (Laveault & Grégoire, 2002). Elles étaient particulièrement utilisées dans les premiers tests d'intelligence pour enfants, comme l'échelle de Binet-Simon, qui avait recours à la notion d'âge mental. Bien qu'encore utilisées dans certains tests pédiatriques, elles sont aujourd'hui souvent critiquées pour leur manque de précision et les malentendus qu'elles peuvent générer (Huteau & Lautrey, 1999).

3.2.2. Les normes de groupe (normes in-tra-groupe)

Les normes intra-groupe sont les plus utilisées en psychométrie contemporaine. Elles situent le score d'un individu par rapport à la distribution des scores d'un échantillon normatif représentatif de la population à laquelle appartient cet individu. Elles prennent principalement deux formes : les rangs centiles et les scores standardisés (Anastasi, trad. Juhel, 1994).

3.2.3. Les normes critériées

Contrairement aux normes intra-groupe, les normes critériées (ou références au critère) ne situent pas l'individu par rapport à un groupe, mais par rapport à un critère de performance absolue défini indépendamment de la distribution des scores dans la population (Laveault & Grégoire, 2002). Ce type de norme est utilisé lorsque l'objectif est de déterminer si un individu atteint ou non un seuil de compétence ou de critère clinique prédéfini.

Par exemple, un test de sécuritié routière peut définir un seuil de réussite à 70 % de bonnes réponses, indépendamment de la performance des autres candidats. De même, en psychologie clinique, un seuil de score à partir duquel un individu est considéré comme présentant un niveau cliniquement significatif d'anxiété (point de coupure ou cut-off) peut être établi sur la base de données cliniques ou d'analyses ROC (Receiver Operating Characteristic) (Bernaud, 2007).

3.3. Les scores normés les plus fréquemment utilisés

3.3.1. Les rangs centiles

Le rang centile (ou percentile) d'un score indique le pourcentage d'individus de la population normative dont le score est égal ou inférieur au score considéré. Un rang centile de 75 signifie que l'individu dépasse 75 % des individus du groupe de référence (Laveault & Grégoire, 2002). Les rangs centiles présentent l'avantage d'être facilement comprehénsibles par des non-spécialistes (professeurs, parents, patients) et de ne faire aucune hypothèse sur la distribution des scores.

Leurs principales limites sont les suivantes (Anastasi, trad. Juhel, 1994) :

- Ils ne constituent pas une échelle à intervalles égaux : une différence de 10 centiles au centre de la distribution ne représente pas la même différence réelle qu'une différence de 10 centiles aux extrêmes.

- Ils exagèrent les différences au centre de la distribution et les minimisent aux extrêmes, en raison de la forme en cloche de la distribution normale.
- Ils ne peuvent pas être moyennés arithmétiquement.

3.3.2. Les scores standardisés (scores Z et scores dérivés)

Le score Z

Le score Z (ou score réduit) est la forme la plus élémentaire de score standardisé. Il exprime le score brut d'un individu en nombre d'écarts-types au-dessus ou en dessous de la moyenne du groupe normatif. Il est calculé par la formule :

$$\mathbf{Z = (X - M) / \sigma}$$

où X est le score brut individuel, M est la moyenne du groupe normatif et σ est l'écart-type du groupe normatif (Bertrand & Blais, 2004). Le score Z a par définition une moyenne de 0 et un écart-type de 1 dans la population normative. Un score Z de +1,5 signifie que l'individu se situe 1,5 écart-type au-dessus de la moyenne, ce qui correspond approximativement au 93e centile sous une distribution normale.

Les scores dérivés

Parce que les scores Z peuvent être négatifs et décimaux, ce qui les rend difficiles à communiquer, plusieurs scores standardisés dérivés ont été développés par transformation linéaire (Laveault & Grégoire, 2002). La formule générale est :

$$\mathbf{Score\ dérivé = M' + \sigma' \times Z}$$

où M' et σ' sont la moyenne et l'écart-type conventionnels choisis pour le score dérivé. Les principales formes de scores dérivés utilisées en psychologie sont présentées dans le tableau suivant :

Nom du score	Moyenne	Écart-type	Domaine d'application principal
Score Z	0	1	Recherche, comparaisons inter-tests
Score T	50	10	Psychologie clinique (MMPI, SCL-90-R)
Quotient (échelle QI)	100	15	Tests d'intelligence (WAIS, WISC, Stanford-Binet)
Stanines	5	2	Orientation scolaire, évaluation éducative
Notes standard (NCE)	50	21,06	Mesures de progression scolaire
Sten (Standard Ten)	5,5	2	Tests de personnalité (16PF, IPIP)

Tableau 3. Principales échelles de scores normés et leurs applications (d'après Anastasi, trad. Juhel, 1994 ; Laveault & Grégoire, 2002 ; Bernaud, 2007).

3.3.3. Les normes spécifiques : normes séparées et normes combinées

Une question cruciale lors de l'étalonnage est de déterminer si les normes doivent être établies séparément pour différents sous-groupes ou de manière globale pour l'ensemble de la population (Anastasi, trad. Juhel, 1994). Des normes séparées par sexe, par tranche d'âge, par niveau d'éducation ou par région géographique sont justifiées lorsque ces variables sont associées à des différences significatives et théoriquement pertinentes dans le construit mesuré.

Par exemple, si les scores moyens à un test de vitesse de traitement de l'information diffèrent significativement entre des sujets de 20 ans et des sujets de 60 ans — ce qui est bien établi empiriquement — des normes par tranches d'âge permettront de situer chaque individu par rapport à ses pairs, plutôt que par rapport à une norme globale qui mélangerait des groupes très différents (Huteau & Lautrey, 1999). En revanche, si les différences entre sous-groupes sont négligeables ou théoriquement non pertinentes, des normes combinées sont préférables, car elles permettent une comparaison plus large.

3.4. Constitution et qualité de l'échantillon normatif

3.4.1. Critères de représentativité

La valeur des normes d'une échelle est directement liée à la qualité de l'échantillon normatif. Un échantillon normatif idéal doit satisfaire plusieurs critères (Bertrand & Blais, 2004 ; Laveault & Grégoire, 2002) :

1. Représentativité : l'échantillon doit refléter fidèlement la distribution des variables clés dans la population cible (sexe, âge, niveau d'éducation, milieu de vie, région géographique).
2. Taille suffisante : un échantillon normatif de moins de 200 sujets est généralement jugé insuffisant ; 500 à 1 000 sujets ou plus sont recommandés pour des normes de qualité. Pour des normes stratès (par âge, sexe, etc.), chaque sous-groupe devrait compter au minimum 100 sujets (Anastasi, trad. Juhel, 1994).
3. Actualité : les normes vieillissent. L'effet Flynn, qui décrit l'augmentation séculaire des scores de QI dans de nombreux pays, illustre bien la nécessité de réactualiser périodiquement les normes des tests d'intelligence (Huteau & Lautrey, 1999). En règle générale, les normes devraient être révisées tous les dix à quinze ans.

4. Adéquation culturelle et linguistique : les normes doivent être établies sur une population qui partage la même culture et la même langue que la population à qui s'adresse l'instrument. Utiliser des normes établies en France pour des sujets algériens sans vérification de leur applicabilité est une pratique méthodologiquement problématique (Mokrani, 2015).

3.4.2. Les méthodes d'échantillonnage

Plusieurs méthodes d'échantillonnage peuvent être utilisées pour constituer un échantillon normatif (Fortin, 2010) :

- L'échantillonnage aléatoire simple : chaque individu de la population cible a une probabilité égale d'être sélectionné. C'est la méthode idéale mais la plus difficile à mettre en oeuvre.
- L'échantillonnage stratéfié : la population est divisée en sous-groupes (strates) homogènes et un sous-échantillon aléatoire est prélevé dans chaque strate, proportionnellement à son poids dans la population. C'est la méthode de référence pour les grands étalon-nages nationaux.
- L'échantillonnage par grappes : on sélectionne aléatoirement des groupes naturels (classes, écoles, services hospitaliers) plutôt que des individus. Très utilisé pour les normes scolaires.
- L'échantillonnage de commodité : utilisation de sujets disponibles et volontaires. Peu représentatif mais fréquemment utilisé dans la recherche, notamment en Algérie, faute de moyens (Khodja, 2016).

3.5. La normalité de la distribution et les transformations de scores

Lorsque la distribution des scores bruts d'un test dans l'échantillon normatif s'écarte significativement de la normalité (distribution en cloche symétrique), il peut être nécessaire d'effectuer des transformations préalables avant l'établissement des normes standardisées (Laveault & Grégoire, 2002). Les transformations les plus courantes sont la transformation logarithmique (pour les distributions présentant une asymétrie positive), la transformation racine carrée ou la normalisation par classement en rangs.

Les scores normalisés (ou normalisés par transformation) sont des scores standardisés dans lesquels la transformation a été choisie de manière à ce que la distribution résultante soit approximativement normale, même si la distribution originale ne l'était pas (Anastasi, trad. Juhel, 1994). Les Stanines sont un exemple courant de scores normalisés : ils divisent la

distribution normale en neuf intervalles de largeur égale (0,5 écart-type chacun), sauf pour les deux extrêmes, et attribuent un entier de 1 à 9 à chaque intervalle.

3.6. L'interprétation des scores normés : précautions et éthique

L'interprétation des scores normés exige un ensemble de précautions éthiques et méthodologiques que Bernaud (2007) et Anastasi (trad. Juhel, 1994) détaillent comme suit :

5. S'assurer que les normes sont adaptées à l'individu évalué : l'individu appartient-il à la population pour laquelle les normes ont été établies ? L'utilisation de normes inadaptées peut conduire à des erreurs d'interprétation graves, notamment en situation de sélection ou d'expertise.
6. Intégrer l'erreur standard de mesure : un score normé n'est jamais une valeur exacte mais une estimation entachée d'erreur. L'interprétation doit toujours s'appuyer sur un intervalle de confiance plutôt que sur un score ponctuel.
7. Vérifier l'actualité des normes : des normes vieilles de plus de quinze ans doivent être utilisées avec précaution, notamment pour les tests d'intelligence.
8. Contextualiser les résultats : un score normé n'est qu'une donnée parmi d'autres. Son interprétation doit être intégrée dans une évaluation globale qui prend en compte l'histoire de la personne, ses conditions de passation et les autres informations disponibles.
9. Respecter la confidentialité et informer la personne : les résultats doivent être communiqués à la personne évaluée de manière compréhensible, en respectant son droit à l'information et en garantissant la confidentialité des données.

Conclusion

Ce cours a proposé une exploration approfondie des trois caractéristiques psychométriques fondamentales des échelles psychologiques : la validité, la fidélité et les normes. Ces trois propriétés forment un triptyque indissociable : un instrument qui manque de validité ne mesure pas ce qu'il prétend mesurer et ses résultats sont interprétables de manière erronée ; un instrument qui manque de fidélité produit des scores instables dont la décision clinique ou éducative ne peut se fonder ; un instrument dépourvu de normes adaptées ne permet pas de donner une signification relative aux scores individuels.

La maitrise de ces concepts est une compétence professionnelle essentielle pour tout psychologue praticien. Elle permet d'évaluer de manière critique les instruments disponibles sur le marché, de choisir les outils les plus adaptés au contexte et à la population, et d'interpréter les résultats avec la rigueur et la prudence qu'ils exigent. Elle est également indispensable pour concevoir, adapter et valider de nouveaux instruments, ce qui représente un besoin crucial dans le contexte algérien.

En Algérie spécifiquement, le développement d'échelles psyché psychométriquement rigoureuses, ancrées dans les réalités culturelles et linguistiques locales, étalonées sur des échantillons représentatifs de la population algérienne, constitue un défi majeur mais aussi une opportunité de développement scientifique qui mérite l'investissement des équipes de recherche en psychologie. C'est à cette condition que la mesure psychologique pourra pleinement remplir sa mission au service du bien-être et du développement des individus.

Références bibliographiques

Présentées selon les normes APA (6e édition), en sources francophones ou traduites en français.

Anastasi, A. (1994). Les tests psychologiques (J. Juhel, trad.). Presses Universitaires de France. (Ouvrage original publié en 1988)

Bernaud, J.-L. (2007). Tests et contextes d'évaluation psychologique de l'adulte. Dunod.

Bertrand, R., & Blais, J.-G. (2004). Modèles de mesure : l'apport de la théorie de réponse aux items. Presses de l'Université du Québec.

Bourque, P., & Beaudette, D. (1982). Étude psychométrique du questionnaire de dépression de Beck auprès d'un échantillon d'étudiants universitaires francophones. Revue canadienne des sciences du comportement, 14(3), 211–218.

Chartier, S., & Morin, C. (2009). Guide pratique de psychométrie. Presses de l'Université du Québec.

Fortin, M.-F. (2010). Fondements et étapes du processus de recherche : méthodes quantitatives et qualitatives (2e éd.). Chenière Éducation.

Huteau, M., & Lautrey, J. (1999). Les tests d'intelligence. La Découverte.

Khodja, M. (2016). L'enseignement de la psychologie en Algérie : état des lieux et perspectives. Revue des sciences humaines de l'Université de Constantine, 27(2), 45–62.

Laveault, D., & Grégoire, J. (2002). Introduction aux théories des tests en psychologie et en sciences de l'éducation (2e éd.). De Boeck Université.

Mokrani, D. (2015). Pratiques d'évaluation psychologique en Algérie : enjeux et défis. Psychologie et Société, 8(2), 33–51.

Reuchlin, M. (1992). Psychologie (11e éd.). Presses Universitaires de France.

Vallerand, R. J., & Hess, U. (2000). Méthodes de recherche en psychologie. Gaëtan Morin Éditeur.