COURS DE CONSULTATION ET ADAPTATION DES TESTS PSYCHOLOHIQUE EN MILIEU PROFESSIONNEL

Tests et Mesure Psychologique

Psychométrie — Niveaux de mesure — Construction des échelles

Spécialité : Psychologie

Niveau : Licence 3

Année universitaire 2025-2026

Par: Dr.YOUCEF KHODJA Adil

Introduction

La mesure en psychologie constitue l'un des défis épistémologiques les plus complexes des sciences humaines et sociales. Comment quantifier des phénomènes aussi intangibles que l'intelligence, l'anxiété, la motivation ou les attitudes ? C'est à cette question fondamentale que tente de répondre la psychométrie, discipline à la croisée de la psychologie et des mathématiques, dont l'objet est précisément de fournir des outils rigoureux pour mesurer les variables psychologiques (Reuchlin, 1992).

Ce cours se structure autour de trois axes complémentaires et progressifs. Le premier explore l'évolution historique de la psychométrie, depuis ses origines européennes jusqu'à son implantation en Algérie, montrant comment cette discipline s'est construite en réponse à des besoins sociaux concrets. Le second axe examine les niveaux de mesure, cadre conceptuel indispensable pour comprendre la nature des données psychologiques et les traitements statistiques qui leur sont applicables. Le troisième axe, enfin, présente de manière détaillée les étapes de construction et de validation d'un test ou d'une échelle psychologique, depuis la définition du construit jusqu'à l'étalonnage de l'instrument.

L'ensemble de ce cours s'appuie exclusivement sur des références francophones ou traduites en français, conformément aux normes de citation de l'American Psychological Association (APA, 6e édition).

Chapitre I : Les niveaux de mesure en psychologie

1.1. La mesure en sciences humaines : enjeux et fondements

La mesure est une opération qui consiste à attribuer des valeurs numériques à des objets ou à des événements selon des règles explicites, de manière à représenter les quantités de la propriété mesurée (Stevens, 1946, cité dans Reuchlin, 1992). Dans les sciences exactes, la mesure bénéficie d'instruments calibrés et de propriétés physiques directement accessibles. En psychologie, en revanche, les propriétés à mesurer — l'intelligence, l'anxiété, l'estime de soi — sont des construits théoriques (variables latentes) qui ne peuvent être observés qu'indirectement, à travers des indicateurs comportementaux.

Cette spécificité impose une réflexion épistémologique sur la nature de la mesure psychologique et sur les opérations statistiques que l'on peut légitimement appliquer aux données recueillies. C'est à Stanley Smith Stevens que l'on doit la classification la plus influente des niveaux de mesure, publiée en 1946 dans la revue Science et traduite et commentée par de nombreux auteurs francophones (Reuchlin, 1992 ; Vallerand & Hess, 2000 ; Fortin, 2010). Stevens distingue quatre niveaux : nominal, ordinal, intervalle et rapport, qui définissent une hiérarchie croissante de propriétés métriques.

1.2. Les quatre niveaux de mesure

2.2.1. Le niveau nominal (ou catégoriel)

Le niveau nominal est le niveau de mesure le plus élémentaire. Il consiste à assigner des individus ou des événements à des catégories qualitatives mutuellement exclusives et exhaustives, en leur attribuant des chiffres ou des étiquettes qui servent uniquement d'identificateurs, sans aucune valeur quantitative (Reuchlin, 1992). Les nombres ne font qu'indiquer l'appartenance à une catégorie et ne peuvent être ni ordonnés ni additionnés.

Exemples en psychologie : le sexe (1 = homme, 2 = femme), la catégorie diagnostique selon le DSM-5 (dépression, trouble anxieux, schizophrénie), la langue maternelle, le type de stratégie d'apprentissage utilisé. Pour Fortin (2010), la mesure nominale est dite « classificatoire » : elle permet uniquement de distinguer des groupes différents, sans indiquer de relation d'ordre ou de quantité entre eux.

Les opérations statistiques légitimes à ce niveau se limitent au dénombrement des effectifs de chaque catégorie (fréquences et pourcentages), au calcul du mode (valeur la plus fréquente) et aux tests du chi-deux pour l'analyse des associations entre variables catégorielles (Vallerand & Hess, 2000).

1.2.2. Le niveau ordinal

Le niveau ordinal ajoute à la propriété classificatoire du niveau nominal une relation d'ordre entre les catégories : les valeurs peuvent être classées du plus petit au plus grand (ou inversement), mais les intervalles entre les valeurs successives ne sont pas nécessairement égaux (Laveault & Grégoire, 2002). Les chiffres assignés indiquent une position relative, un rang, mais non une différence absolue.

Exemples en psychologie : le rang de classement dans une promotion universitaire (1er, 2e, 3e...) ; les réponses aux échelles de Likert (Tout à fait d'accord > D'accord > Neutre > En désaccord > Tout à fait en désaccord) ; les niveaux socio-économiques (bas, moyen, élevé) ; les stades de développement moral selon Kohlberg. Dans ces exemples, on sait que la catégorie 3 est supérieure à la catégorie 2, mais on ne peut affirmer que l'intervalle entre 2 et 3 est identique à l'intervalle entre 1 et 2.

Les statistiques appropriées à ce niveau incluent la médiane, le rang centile, le coefficient de corrélation de Spearman (rho) et les tests non paramétriques tels que le test de Mann-Whitney ou le test de Kruskal-Wallis (Fortin, 2010). L'utilisation de la moyenne arithmétique est techniquement discutable à ce niveau, bien que très répandue en pratique, notamment pour les scores composites d'échelles de Likert (Laveault & Grégoire, 2002).

1.2.3. Le niveau intervalle (ou d'intervalles égaux)

Le niveau intervalle possède toutes les propriétés du niveau ordinal, et y ajoute l'égalité des intervalles entre les valeurs successives de l'échelle. Cela signifie que la distance entre deux points consécutifs est constante tout au long de l'échelle, ce qui permet d'effectuer des opérations d'addition et de soustraction (Bertrand & Blais, 2004). Cependant, ce niveau ne dispose pas d'un zéro absolu : le zéro est arbitraire et ne signifie pas l'absence totale de la propriété mesurée.

L'exemple canonique est l'échelle de température en degrés Celsius : la différence entre 10 °C et 20 °C est identique à celle entre 30 °C et 40 °C, mais 0 °C ne signifie pas l'absence de température. En psychologie, les scores de QI sont conventionnellement traités comme des

mesures d'intervalle : un QI de 130 n'est pas « deux fois plus intelligent » qu'un QI de 65, mais l'intervalle entre 85 et 100 est considéré équivalent à celui entre 100 et 115 (Huteau & Lautrey, 1999).

Les tests psychologiques standardisés (tests d'intelligence, d'aptitudes, scores composites de personnalité) sont le plus souvent traités comme des mesures d'intervalle en pratique, même si cette hypothèse est parfois discutée. Les opérations statistiques permises incluent la moyenne arithmétique, l'écart-type, la variance, la corrélation de Pearson, et les tests paramétriques comme le test t de Student et l'analyse de variance (ANOVA) (Vallerand & Hess, 2000).

1.2.4. Le niveau de rapport (ou ratio)

Le niveau de rapport (ou niveau de ratio) est le niveau de mesure le plus élevé. Il possède toutes les propriétés du niveau intervalle, mais dispose en outre d'un zéro absolu qui indique l'absence totale de la propriété mesurée (Reuchlin, 1992). L'existence d'un zéro absolu permet d'effectuer toutes les opérations arithmétiques, y compris la multiplication et la division, et de former des rapports entre les valeurs.

Exemples physiques : le poids, la taille, le temps de réaction (en millisecondes). En psychologie, les mesures de ratio sont relativement rares, car peu de variables psychologiques présentent un zéro absolu vérifiable. On peut citer le nombre de comportements observés dans une unité de temps (fréquence comportementale), le nombre d'erreurs commises à une épreuve, ou encore le temps mis pour accomplir une tâche. Des mesures comme l'amplitude de l'activité électrodermale en psychophysiologie peuvent également atteindre ce niveau.

À ce niveau, toutes les opérations statistiques sont permises, incluant le calcul de coefficients de variation et de proportions. Dans la pratique psychologique, les analyses statistiques utilisées pour les niveaux intervalle et ratio sont souvent identiques (Fortin, 2010).

1.3. Tableau récapitulatif des niveaux de mesure

Niveau	Propriétés	Opérations	Statistiques	Exemples psy.
Nominal	Classification, identité	= / ≠	Mode, fréquences, chi-deux	Sexe, diagnostic
Ordinal	+ Ordre, rang	= / ≠ / > / <	Médiane, rang	Likert, niveaux

Niveau	Propriétés	Opérations	Statistiques	Exemples psy.
			centile, rho de Spearman	socio-éco.
Intervalle	+ Intervalles égaux, zéro arbitraire	+ / −	Moyenne, écart-type, Pearson, ANOVA	QI, scores composites
Rapport	+ Zéro absolu	× / ÷	Toutes statistiques	Temps de réaction, fréquences comportementales

Tableau 1. Récapitulatif des niveaux de mesure (d'après Stevens, 1946, cité dans Reuchlin, 1992 ; Vallerand & Hess, 2000).

1.4. Implications pour la pratique psychométrique

La question du niveau de mesure n'est pas purement académique : elle détermine directement la légitimité des analyses statistiques que le chercheur ou le clinicien peut conduire sur ses données. Utiliser une moyenne arithmétique pour des données nominales, ou appliquer une corrélation de Pearson à des données purement ordinales sans vérification préalable de la distribution, constitue une erreur méthodologique susceptible de produire des conclusions erronées (Laveault & Grégoire, 2002).

Dans le contexte algérien, cette problématique est particulièrement importante. De nombreux travaux de recherche utilisent des échelles de Likert et les traitent directement comme des données d'intervalle sans justification théorique. Si cette pratique est répandue et défendue par certains auteurs (notamment lorsque l'échelle comporte au moins cinq points et présente une distribution proche de la normalité), elle mérite d'être mentionnée et discutée explicitement dans tout rapport de recherche (Fortin, 2010).

Chapitre II : Étapes de construction des tests et échelles psychologiques

2.1. Vue d'ensemble du processus de construction

La construction d'un test ou d'une échelle psychologique est un processus long, rigoureux et itératif, qui exige une expertise à la fois théorique et méthodologique. Il ne suffit pas de formuler des items et de les regrouper pour obtenir un instrument valide : chaque étape du processus doit répondre à des critères scientifiques précis, sous peine de produire un outil dont les résultats seront ininterprétables ou, pire, trompeurs (Chartier & Morin, 2009).

Plusieurs auteurs ont proposé des modèles décrivant les étapes de ce processus. Parmi les plus cités en langue française, on trouve Vallerand et Hess (2000), Laveault et Grégoire (2002), Bertrand et Blais (2004), Fortin (2010) et Bernaud (2007). Si leurs découpages varient légèrement, ils s'accordent sur un ensemble d'étapes fondamentales que nous présentons ici de manière détaillée et articulée.

2.2. Étape 1 : Définition du construit et revue de la littérature

2.2.1. La définition conceptuelle

Toute démarche de construction d'un test commence par la définition rigoureuse du construit psychologique que l'on souhaite mesurer. Un construit est une variable théorique non directement observable (comme l'intelligence, l'anxiété, l'estime de soi ou le burn-out) que le chercheur postule pour expliquer des comportements observables (Vallerand & Hess, 2000). La définition du construit doit préciser : (a) sa nature (trait stable, état transitoire, attitude, compétence) ; (b) ses dimensions ou facettes constitutives ; (c) ses limites conceptuelles par rapport aux construits voisins (ce qu'il est et ce qu'il n'est pas) ; et (d) la population cible à laquelle l'instrument est destiné.

Cette étape s'appuie sur une revue exhaustive de la littérature scientifique existante. Le chercheur doit identifier les théories qui sous-tendent le construit, examiner les instruments déjà existants dans le domaine, et analyser leurs forces et leurs limites. Cette analyse permettra de justifier la nécessité de construire un nouvel instrument plutôt que d'adapter un outil existant (Bertrand & Blais, 2004).

2.2.2. La définition opérationnelle

Une fois le construit défini au plan conceptuel, il faut l'opérationnaliser, c'est-à-dire le traduire en indicateurs comportementaux mesurables. La définition opérationnelle précise exactement ce que le test va mesurer et comment, en termes de comportements, de réponses ou de performances attendus. Laveault et Grégoire (2002) soulignent que cette étape est décisive pour la validité de contenu de l'instrument : si la définition opérationnelle est trop étroite, le test risque de sous-représenter le construit ; si elle est trop large, il risque de mesurer des éléments hétérogènes ou parasites.

2.3. Étape 2 : Spécification du plan de l'instrument

Sur la base de la définition opérationnelle, le chercheur établit un plan de l'instrument qui précise : (a) le nombre de dimensions ou sous-échelles à mesurer ; (b) le nombre d'items par dimension ; (c) le format des items (affirmations, questions, situations à évaluer) ; (d) le format de réponse (choix forcé, Likert, différentielle sémantique, vrai/faux) ; et (e) les modalités de passation (individuelle ou collective, papier-crayon ou informatisée, durée estimée) (Chartier & Morin, 2009).

Cette étape implique également une réflexion sur le sens de la cotation : certains items peuvent être formulés dans le sens du construit (items directs) et d'autres en sens inverse (items inversés ou renversés), afin de réduire les effets d'acquiescement — la tendance des répondants à accepter les affirmations quelle que soit leur contenu (Vallerand & Hess, 2000). Un bon plan d'instrument équilibre systématiquement les items directs et inversés.

2.4. Étape 3 : Rédaction des items

2.4.1. Principes généraux de rédaction

La rédaction des items est une étape qui exige à la fois une compétence linguistique, une maîtrise du construit et une sensibilité aux biais potentiels. Selon Vallerand et Hess (2000) et Fortin (2010), les items d'une bonne échelle psychologique doivent respecter les principes suivants :

- Clarté et précision : chaque item doit exprimer une seule idée, formulée de manière non ambiguë. Les termes techniques, les doubles négations et les formulations vagues sont à éviter.
- Pertinence théorique : chaque item doit être un indicateur valide du construit ou de la dimension qu'il est censé représenter.

- Adaptation à la population cible : le vocabulaire, la syntaxe et les références culturelles doivent être accessibles et pertinents pour la population visée.
- Neutralité : les items ne doivent pas suggérer de réponse « socialement désirable » ni révéler le sens de la cotation aux répondants.
- Équilibre : prévoir environ 30 à 50 % d'items inversés pour contrôler les biais d'acquiescement et de réponse extrême.

2.4.2. La génération d'un pool d'items

Il est recommandé de générer initialement un nombre d'items nettement supérieur au nombre final souhaité — généralement deux à trois fois plus (Chartier & Morin, 2009). Si l'on souhaite obtenir une échelle de 20 items, on commencera par en rédiger 50 à 60. Cette surproduction est nécessaire car de nombreux items seront éliminés lors des étapes de validation. Le chercheur peut s'appuyer sur plusieurs sources pour générer les items : la littérature théorique et empirique, des entretiens qualitatifs avec des membres de la population cible, des groupes de discussion (focus groups), et le jugement d'experts du domaine (Bertrand & Blais, 2004).

2.5. Étape 4 : Validation du contenu par les experts

Avant toute collecte de données empiriques, les items rédigés sont soumis à un panel de juges experts — généralement des chercheurs ou des praticiens reconnus dans le domaine du construit. Ces experts évaluent chaque item selon des critères de pertinence (l'item est-il un bon indicateur du construit ?), de clarté (l'item est-il compréhensible ?) et d'exhaustivité (l'ensemble des items couvre-t-il bien toutes les dimensions du construit ?) (Fortin, 2010).

Le degré d'accord entre les experts peut être quantifié à l'aide d'indices statistiques tels que l'Indice de Validité de Contenu (IVC) proposé par Lynn (1986, cité dans Fortin, 2010), qui calcule la proportion de juges ayant évalué chaque item comme « pertinent » ou « très pertinent ». Un IVC supérieur à 0,80 pour chaque item et supérieur à 0,90 pour l'ensemble de l'échelle est généralement requis pour conclure à une validité de contenu satisfaisante. Les items ne satisfaisant pas ces critères sont révisés ou éliminés.

2.6. Étape 5 : Étude pilote et analyse des items

2.6.1. L'étude pilote

L'étude pilote consiste à administrer la version préliminaire de l'instrument à un échantillon de taille modeste (généralement entre 50 et 200 participants) représentatif de la population

cible. Cette étape permet de vérifier la compréhensibilité des consignes et des items, d'évaluer la durée de passation, d'identifier les items problématiques (trop faciles, trop difficiles, ambigus ou redondants) et de recueillir les commentaires des participants (Chartier & Morin, 2009). À l'issue de cette phase, les items peuvent être reformulés, réorganisés ou supprimés avant de procéder à la collecte de données principale.

2.6.2. L'analyse statistique des items

Les données de l'étude pilote font l'objet d'analyses statistiques détaillées visant à évaluer la qualité de chaque item individuellement. Ces analyses comprennent :

1. L'indice de difficulté (ou d'endossement) : pour les items à réponse dichotomique, il correspond à la proportion de sujets ayant donné la bonne réponse (ou la réponse dans le sens du construit). Un item trop facile (endossé par > 90 % des sujets) ou trop difficile (< 10 %) apporte peu d'information discriminante (Laveault & Grégoire, 2002).
2. L'indice de discrimination (corrélation item-total corrigée) : mesure le degré de corrélation entre la réponse à un item donné et le score total de l'échelle (ou de la sous-échelle), après soustraction de l'item en question. Une corrélation supérieure à 0,30 est généralement requise pour qu'un item soit conservé (Bertrand & Blais, 2004).
3. L'alpha de Cronbach si l'item est supprimé : indique si la suppression de l'item améliorerait la cohérence interne globale de l'échelle. Si la valeur de l'alpha augmente significativement lorsque l'item est retiré, ce dernier est candidat à l'élimination (Chartier & Morin, 2009).
4. La distribution des réponses : l'examen des distributions marginales permet de détecter les items présentant des effets plancher ou plafond, ou une distribution fortement asymétrique, qui réduisent la sensibilité de l'instrument (Vallerand & Hess, 2000).

2.7. Étape 6 : Collecte des données principale et analyse de la structure factorielle

2.7.1. Constitution de l'échantillon de validation

Une fois la version révisée de l'instrument établie, une collecte de données principale est organisée sur un échantillon de taille suffisante pour les analyses statistiques multivariées. La règle empirique la plus répandue suggère un minimum de 5 à 10 participants par item, avec un plancher absolu de 200 participants pour une Analyse Factorielle Exploratoire

(AFE) et de 300 à 500 pour une Analyse Factorielle Confirmatoire (AFC) (Bertrand & Blais, 2004). L'échantillon doit être représentatif de la population cible en termes d'âge, de sexe, de niveau d'éducation et, dans le contexte algérien, de région et de langue.

2.7.2. L'Analyse Factorielle Exploratoire (AFE)

L'AFE est utilisée lorsque la structure factorielle de l'instrument n'est pas encore connue ou doit être vérifiée empiriquement. Elle vise à identifier les dimensions latentes communes aux items, en regroupant les items qui partagent une variance commune. Les méthodes d'extraction les plus utilisées sont l'analyse en composantes principales (ACP) et la factorisation par axes principaux. La méthode de rotation la plus courante est la rotation Varimax (orthogonale) pour des facteurs non corrélés, ou Oblimin/Promax (oblique) lorsque les facteurs sont attendus comme corrélés (Bertrand & Blais, 2004).

La détermination du nombre de facteurs à retenir repose sur plusieurs critères : la règle de Kaiser (valeurs propres > 1), le test du coude de Cattell (graphique des valeurs propres), la variance expliquée (en visant généralement 50 à 60 % pour les sciences humaines) et l'interprétabilité théorique des facteurs. Un item est retenu dans un facteur si sa saturation factorielle est supérieure à 0,40, et il est éliminé s'il présente des saturations comparables sur plusieurs facteurs (items complexes) (Chartier & Morin, 2009).

2.7.3. L'Analyse Factorielle Confirmatoire (AFC)

L'AFC est utilisée lorsque le chercheur dispose d'un modèle théorique précis de la structure factorielle qu'il souhaite tester. Contrairement à l'AFE, l'AFC ne découvre pas la structure : elle évalue dans quelle mesure les données observées s'ajustent au modèle postulé a priori. Elle s'inscrit dans le cadre de la modélisation par équations structurales (MES), dont les logiciels les plus utilisés dans le monde francophone sont LISREL, AMOS, Mplus et R (package lavaan) (Bertrand & Blais, 2004).

L'évaluation de l'ajustement du modèle repose sur plusieurs indices : le chi-deux (et son rapport au degré de liberté, idéalement < 3), le RMSEA (Root Mean Square Error of Approximation, valeur < 0,06 jugée bonne), le CFI et TLI (Comparative Fit Index et Tucker-Lewis Index, valeurs > 0,95 jugées bonnes), et le SRMR (Standardized Root Mean Square Residual, valeur < 0,08 jugée acceptable) (Fortin, 2010 ; Bertrand & Blais, 2004).

2.8. Étape 7 : Évaluation de la fidélité de l'instrument

La fidélité (ou fiabilité) désigne la consistance et la stabilité des mesures produites par l'instrument. Un instrument fidèle produit des résultats similaires lorsqu'il est administré dans des conditions identiques à un même individu ou groupe (Laveault & Grégoire, 2002). On distingue plusieurs formes de fidélité :

- La cohérence interne : mesurée par l'alpha de Cronbach, elle évalue l'intercorrélation des items d'une même dimension. Un alpha supérieur à 0,70 est généralement considéré comme acceptable pour la recherche, et supérieur à 0,80 pour un usage clinique individuel (Chartier & Morin, 2009). L'oméga de McDonald constitue une alternative plus robuste, notamment en présence de facteurs corrélés.
- La fidélité test-retest : mesurée par la corrélation entre les scores obtenus à deux passations successives du même instrument, séparées par un intervalle de temps suffisant (généralement 2 à 4 semaines). Un coefficient de corrélation supérieur à 0,70 est généralement requis (Vallerand & Hess, 2000).
- La fidélité inter-juges : pertinente pour les instruments faisant appel à l'observation ou au codage, elle mesure le degré d'accord entre différents évaluateurs. Elle est quantifiée par le coefficient Kappa de Cohen pour les données nominales ou le coefficient de corrélation intraclasse (CCI) pour les données continues (Fortin, 2010).
- La fidélité par formes parallèles : consiste à élaborer deux versions équivalentes de l'instrument et à mesurer leur corrélation. Cette méthode est peu utilisée en pratique en raison de la difficulté à construire des formes véritablement équivalentes (Bertrand & Blais, 2004).

2.9. Étape 8 : Évaluation de la validité de l'instrument

2.9.1. La validité de contenu

La validité de contenu, déjà examinée lors de la consultation des experts (étape 4), renvoie à la représentativité des items vis-à-vis du construit mesuré. Elle est évaluée de manière qualitative et logique, en vérifiant que l'ensemble des items couvre bien l'étendue conceptuelle du construit sans lacune ni chevauchement avec d'autres construits (Fortin, 2010).

2.9.2. La validité de construit

La validité de construit est la forme de validité la plus importante en psychométrie contemporaine. Elle évalue dans quelle mesure le score de l'instrument est un reflet valide du construit théorique qu'il est censé mesurer. Elle est évaluée par plusieurs approches

complémentaires : la validité convergente (les scores corrèlent-ils positivement avec ceux d'instruments mesurant des construits proches ?), la validité divergente ou discriminante (les scores sont-ils suffisamment différents de ceux mesurant des construits non liés ?), et la validité factorielle (la structure factorielle obtenue correspond-elle à la structure théorique postulée ?) (Vallerand & Hess, 2000).

Dans le cadre de l'analyse multitrait-multiméthode (MTMM) proposée par Campbell et Fiske (1959, cités dans Vallerand & Hess, 2000), on attend à la fois des corrélations convergentes élevées entre des instruments différents mesurant le même construit, et des corrélations discriminantes faibles entre des instruments mesurant des construits différents.

2.9.3. La validité de critère

La validité de critère évalue la capacité du score de l'instrument à prédire ou à être associé à un critère externe pertinent. On distingue la validité concomitante (le score est corrélé avec un critère mesuré simultanément) et la validité prédictive (le score prédit un critère mesuré dans le futur, ce qui est particulièrement utile dans les contextes de sélection et d'orientation) (Bernaud, 2007). Par exemple, un test d'aptitude professionnelle doit prédire significativement les performances au travail mesurées plusieurs mois après la passation.

2.10. Étape 9 : Étalonnage et normalisation

L'étalonnage consiste à établir des normes qui permettent d'interpréter le score brut d'un individu en le situant par rapport à un groupe de référence (Anastasi, trad. Juhel, 1994). Un score brut de 45 à un test d'anxiété ne signifie rien en lui-même ; il prend sens uniquement lorsqu'on sait qu'il correspond, par exemple, au 75e percentile chez des adultes de la population générale — ce qui indique un niveau d'anxiété supérieur à celui de 75 % des personnes du groupe normatif.

Les normes les plus couramment utilisées sont les rangs percentiles (position relative dans la distribution), les notes standards (transformation des scores bruts en scores dont la moyenne et l'écart-type sont fixés conventionnellement : scores Z, scores T avec M = 50 et ET = 10, notes de quotient avec M = 100 et ET = 15) et les notes d'âge ou de grade (utilisées pour les tests développementaux) (Laveault & Grégoire, 2002).

L'échantillon normatif doit être suffisamment large (plusieurs centaines à plusieurs milliers de sujets) et représentatif de la population cible. Des normes distinctes sont établies pour les sous-groupes pertinents (hommes/femmes, tranches d'âge, niveaux d'éducation, régions).

En Algérie, l'établissement de normes représentatives reste un défi majeur, en raison des coûts logistiques élevés et de la diversité socio-culturelle et linguistique du pays (Khodja, 2016).

2.11. Étape 10 : Rédaction du manuel et diffusion

La dernière étape du processus de construction est la rédaction d'un manuel technique de l'instrument. Ce manuel doit contenir : (a) une description du construit mesuré et de ses fondements théoriques ; (b) la description complète des items et des formats de réponse ; (c) les procédures standardisées de passation, de cotation et d'interprétation ; (d) les données psychométriques (fidélité et validité) obtenues lors de la validation ; (e) les normes et les tables d'étalonnage ; et (f) les précautions éthiques d'utilisation (Bernaud, 2007).

La diffusion de l'instrument doit respecter les principes éthiques encadrant l'utilisation des tests psychologiques. En France et dans les pays francophones, les directives de la Commission Internationale des Tests (CIT/ITC) et de la Société Française de Psychologie précisent que l'accès aux tests doit être réservé aux professionnels qualifiés, que les résultats doivent être communiqués de manière compréhensible aux personnes évaluées, et que les tests ne doivent pas être utilisés à des fins discriminatoires (Bernaud, 2007).

Conclusion générale

Ce cours a présenté un panorama complet et structuré des fondements de la psychométrie et des tests psychologiques, articulé autour de trois axes complémentaires. L'étude de l'évolution historique de la discipline a montré comment la psychométrie s'est progressivement construite, des premiers travaux de Galton et Binet jusqu'aux développements contemporains de la Théorie de la Réponse à l'Item, en réponse à des besoins sociaux précis. Elle a également mis en lumière le développement particulier de cette discipline en Algérie, de la période coloniale à la psychométrie contemporaine, avec ses richesses et ses défis spécifiques.

L'examen des niveaux de mesure a fourni le cadre conceptuel indispensable pour comprendre la nature des données psychologiques et choisir les analyses statistiques appropriées. La distinction entre les niveaux nominal, ordinal, intervalle et rapport conditionne directement la rigueur des analyses et la légitimité des conclusions que le chercheur peut tirer de ses données.

Enfin, la présentation détaillée des dix étapes de construction d'un test ou d'une échelle psychologique a illustré la complexité et l'exigence de ce processus : de la définition conceptuelle du construit à l'étalonnage final de l'instrument, en passant par la rédaction des items, l'analyse factorielle et l'évaluation de la fidélité et de la validité. Cette rigueur est la condition sine qua non pour que la mesure psychologique puisse remplir ses fonctions essentielles : décrire, expliquer, prédire et orienter les comportements humains dans une perspective à la fois scientifique et éthique.

Pour la psychologie algérienne, ces enseignements prennent une résonance particulière : le développement d'instruments authentiquement ancrés dans les contextes culturels et linguistiques locaux, la constitution d'équipes de recherche interdisciplinaires et le renforcement des infrastructures de recherche constituent les conditions d'une psychométrie réellement au service des populations algériennes.

Références bibliographiques

Les références ci-dessous sont présentées selon les normes APA (6e édition).

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

Benbouzid, B. (2002). La psychologie en Algérie : histoire, formation et pratiques. Revue algérienne de psychologie, 1(1), 5-24.

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.

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.). Chenelière 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.

Zazzo, R. (1969). Manuel pour l'examen psychologique de l'enfant (3e éd.). Delachaux et Niestlé.