CHAPITRE III : LES ÉPREUVES STANDARDISÉES D'ÉVALUATION

✪ Présentation générale.

Le **test** est un **outil** psychologique et il désigne *toutes techniques permettant une description quantitative, contrôlable du comportement d'un individu placé dans une situation définie par référence au comportement des individus d'un groupe défini placé dans la même situation.*

○ Caractéristiques des épreuves standardisées.

Standardisation.

Deux éléments ne sont jamais standardisés : le **sujet** et l'**évaluateur**. Cependant au-delà de ces deux éléments tout doit être identique d'une passation à un autre :

- **Les conditions d'application** (ou consignes) : elles doivent être lues et les mêmes pour tous les sujets.
- **Le matériel** : même objet, mêmes matériaux... quitte à avoir les plans de construction de l'objet.
- **Les principes de notation** : barème précis et clair afin d'avoir une grande fidélité inter-évaluateur. Parfois il est utile d'utiliser l'informatique.

Étalonnage et norme.

Afin de définir la norme on fait passer le test à un échantillon de personnes. Il en ressort une **moyenne** et un **écart-type**.

Puis on passe à la standardisation des notes : c'est-à-dire que les notes brutes deviennent des **notes Z** ou des notes standardisées (ce sont des notes **étalonnées**). Cela permet de situer le sujet dans un groupe de référence.

○ L'échantillon.

Il doit être **représentatif** (*une image réduite mais fidèle d'une population générale*). C'est-à-dire qu'il doit avoir les mêmes caractéristiques que la population d'où il provient.

○ Erreur de mesure et nature probabiliste des scores observés.

Le test ne donne qu'une **estimation**. Mais pourquoi?

- Car on test qu'un **petit échantillon** de comportement, on ne peut pas tous les prendre en compte.
- Et car il existe une **fluctuation intra-individuelle** (dans les émotions, la réussite).

➔ Cependant l'erreur dépend du domaine de mesure : ainsi nous observons une marge d'erreur plus importante dans les tests de personnalité que dans un test cognitif.

✪ Démarche générale de la construction des tests.

➲ **Définir les objectifs** : est-ce un outil de recherche ou d'application? A pour but un pronostic ou un diagnostic?

➲ **Définir le domaine à mesurer** : on parle d'échantillonnage des items (l'échantillon d'items doit être représentatif du domaine à mesurer). Cela est plus ou moins facile. Ainsi il est plus aisé de définir le domaine dans un test de connaissance, mais il est plus compliqué de le faire avec un test d'intelligence (varie selon les cultures).

➲ **L'échantillonnage des sujets**.

➲ **La construction des items** : choisir la forme et le contenu (et par là même vérifier la validité de contenu).

➲ **La standardisation**.

➲ **Le pré-test** : on recueille des premières réponses et réactions (par feedback). Par la suite il arrive que l'on reconstruise le test (pour qu'il devienne plus clair, plus précis...etc).

➲ **La construction d'échelle de validité interne** : cette étape n'est pas obligatoire, mais elle permet (entre autre dans les tests de personnalité) de savoir si les personnes sont sincères ou non.

➲ **Étude des qualités psychométriques** : on analyse tout d'abord la sensibilité des items et du test. Puis la fidélité (savoir si le test est stable et répétable). Et enfin la validité (connaître l'efficacité des mesures, et savoir si les données sont conformes à une théorie).

➲ **Le travail de reconstruction et d'amélioration** : suite aux différentes analyses il arrive que le test doive être modifié (afin que les mesures soient plus convenables).

➲ **La construction de normes et d'étalonnages** : pour ce faire on recueille de nouvelles données sur un nouvel échantillon.

➲ **La contrevalidation** : on doit refaire l'étude au bout de quelques années afin de déterminer si le test est toujours valide.

➲ **La publication éventuelle** : il faut savoir qu'aujourd'hui les tests sont essentiellement informatisés pour éviter tout «vol».

➲ **Les révisions** : on doit éviter au maximum les effets du vieillissement en faisant évoluer le test et en le modifiant (changements culturels).

✪ Élaboration psychométrique.

○ La sensibilité.

Le but du test est de faire ressortir les **différences inter-individuelles.** On parle du **pouvoir classant** du test ou de la **finesse discriminative** de l'épreuve. On utilise alors les indices de dispersion.

Les indicateurs et les échelles de mesure.

- **Échelles nominales** : elles peuvent être dichotomiques/binaires (oui/non par exemple), dans ce cas la sensibilité parfaite serait d'avoir 50%/50%. Mais elles peuvent être également multitomiques (plusieurs réponses, 3, 4, 5...etc). Dans ce dernier cas la sensibilité parfaite serait d'avoir la même répartition de sujets par réponse.

- **Échelles ordinales** : on parle de score ipsatif. Il s'agit pour le sujet d'ordonner ses réponses ainsi A > B > C ou B > C > A par exemple.

- **Échelles d'intervalles** : ce sont des échelles continues.

◦ **La fidélité.**

Pour qu'un test soit fidèle il doit nécessairement être sensible. La fidélité consiste à **évaluer l'erreur de mesure**. L'idéal serait d'obtenir un **score vrai** *: score moyen que l'on pourrait obtenir si on pouvait retester la personne plusieurs fois de suite.*

Test-retest.

Un test a une bonne fidélité dans le temps si appliqué deux fois sur le même échantillon dans les mêmes conditions standards on obtient des résultats similaires. On calcule alors la **corrélation** entre les deux séries de données obtenues (il faut cependant un intervalle de temps court).

Les formes parallèles.

On crée deux fois plus de questions qu'on le souhaitait à la base (par exemple 40 au lieu de 20) et ainsi on fait passer à l'échantillon les 20 premières questions, puis plus tard les 20 autres. Cela évite **l'effet d'apprentissage.**

L'homogénéité ou consistance interne du test.

- **Bi-partition** : on découpe les scores de l'échantillon en deux. Et on calcule la corrélation entre les deux moitiés. Cela permet de dire si le test se réfère bien à une unique dimension (domaine). Une bi-partition possible : **la méthode pair/impair** (prendre les questions pairs d'un côté et les questions impairs de l'autre).

- **Le coefficient d'homogénéité** : on utilise très souvent **l'alpha de Cronbach** = moyenne de toutes les corrélation que l'on pourrait obtenir avec toutes les bi-partition possibles.

La fidélité inter-évaluateurs.

On utilise encore une fois la **corrélation** entre les évaluateurs (voire l'alpha de Cronbach).

→ *Par exemple : si nous obtenons une corrélation de .90 alors le score vrai est égal à 90%, tandis que l'erreur vaut 10%. La corrélation varie également avec le domaine de mesure (ainsi pour un test de personnalité la corrélation sera plus faible que pour un test cognitif).*

- **La validité.**

Un test non fidèle n'est pas valide, mais un test peut-être fidèle sans être valide.

On dira qu'un test est valide s'il mesure bien ce qu'il est censé mesurer.

Validité de contenu : représentativité des items.

Une fois que le domaine est bien délimité, on fait appel à un **expert** pour qu'il confirme la représentativité des items. Cette validité est utilisée dans les domaines **objectifs**.

Validité empirique ou de critère.

Le test est considéré comme un **prédicteur**. On utilise alors un **critère** et on mesure le degré de liaison

(avec un coefficient de corrélation) entre ce critère et le test/ le prédicteur.

→ Par exemple : *on crée un test qui permet de savoir en décembre quels jeunes d'une promotion auront leur bac en juin. On fait passer le test, on recueille les données. Puis en juin on récupère les notes qu'on obtenu ces mêmes jeunes au bac. Et enfin on mesure la corrélation entre les scores au test et les notes du bac. Les notes du bac sont les critères.*

➔ On distingue : la **validité concomitante/ simultanée** (on mesure la validité en même temps que la passation du test) et la **validité prédictive** (comme l'exemple précédent).

Validité théorique ou de construction.

Le test est alors valide si les résultats sont conformes à la théorie. Cette validité est utilisé dans les domaines **subjectifs**. On distingue :

- **La validité (ou structure) interne** : on étudie la corrélation des items deux à deux. On distingue alors des groupes d'items qui ont une forte corrélation entre eux (on parle de clusters). Normalement si on mesure une dimension, on doit observer un unique cluster.

- **La validité conceptuelle** : on étudie la nature psychologique des données.

On utilise différentes méthodes (adaptées au test et aux items) pour mesurer cette validité :

- **Méthodes des groupes contrastés** : par exemple on crée un test qui permet de mesurer «l'introversion». On prend un groupe de sujets très extravertis et un groupe de sujets très introvertis. On leur fait passer le test et chaque groupe doit se trouver d'un côté différent de la médiane.

- **Analyse corrélationnelles/critères convergents** (ou divergents) : en reprenant l'exemple précédent, cette

fois-ci on compare les scores obtenus à ce test avec des scores obtenus à un test qui mesure «l'extraversion» (en sachant que ce test de l'extraversion a déjà été validé).

- **Analyses factorielles**.

Toutes ces validités sont complémentaires.

➪ <u>À savoir dans ce chapitre</u> : *ce chapitre est sûrement le plus important de tous. Il faut vraiment connaître les différents types de validité et de fidélité. Les différents étapes de construction d'un test sont à lire, à comprendre mais pas à savoir.*