FICHE PEDAGOGIQUE

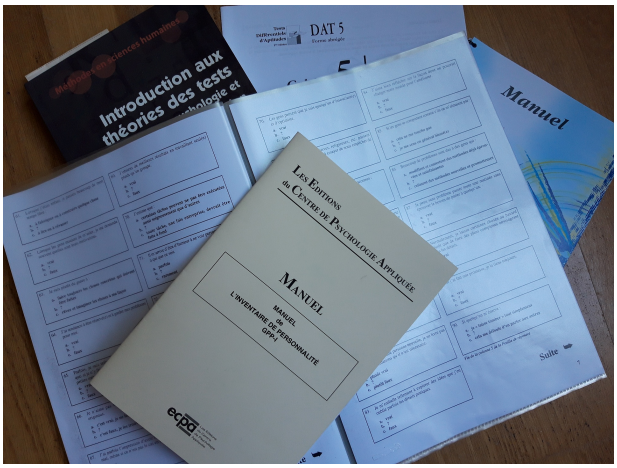
La construction du test psychométrique

V1.00

18/05/2022

Si vous avez téléchargé cette fiche ailleurs que sur le site, assurez-vous d'avoir la dernière version ici :
http://deporientation.free.fr/Ressources/Ressources.html

Fabien BELTRAME, Ph.D.



1 Introduction

Cette fiche pédagogique se donne pour objectif de **vulgariser**, dire simplement, ce qu'est un test psychométrique. La bonne compréhension de la conception, de la fabrication et de la validation d'un test psychométrique permet de mieux comprendre pourquoi, après avoir répondu à un questionnaire validé scientifiquement, **je peux** dire « *je n'ai pas aimé ce test* » mais **je ne peux pas** dire « *ce questionnaire est absurde, il ne vaut rien !* ».

Dire « *je n'ai pas aimé ce test* » est un ressenti tout à fait personnel et tout aussi légitime. Dire « *ce questionnaire est absurde, il ne vaut rien !* », c'est mettre en doute la qualité de la mesure réalisée par l'outil et nous allons voir que pour faire cela, il faut faire la démonstration que les analyses psychométriques de validation sont critiquables. Et ça … c'est relativement compliqué

Cette fiche s'adresse à un public néophyte en psychométrie et prend volontairement des raccourcis pour rendre le plus accessible possible la présentation des différentes phases de fabrication d'un test.

> « *Un test est un outil d'évaluation objective qui présente quatre propriétés :*
> 1. *il est standardisé : toutes les personnes répondent aux mêmes questions ;*
> 2. *il permet de situer les réponses de chaque personne par rapport aux personnes qui lui ressemblent ;*
> 3. *le degré de précision des mesures qu'il permet est évalué ;*
> 4. *la signification théorique de la mesure est expliquée (validité).* »
>
> Inspiré de Michel Huteau et Jacques Lautrey (1997) légèrement vulgarisé

2 De la conception à la validation du test

Etape 1	Etape 2	Etape 3	Etape 4	Etape 5	Etape 6
Choix du modèle scientifique qui sous-tend la/les dimension(s) à mesurer	Rédaction des items conformément aux préconisations du modèle	Création et passation de la Version « 0 » du test	Analyses psychométriques de la V0 du test	Choix des items et création de la V1 du test	Passation et analyse de la V1 du test
Exemple Dimension : personnalité Modèles : - C.G. Jung - Big Five - Eysenck - etc ...	Exemple Dimension : personnalité Définition des caractéristiques de l'introversion par le modèle Prévoir 2 fois plus d'items qu'il n'en restera à la fin	Exemple Dimension : personnalité Un minimum de 300 personnes répond à l'ensemble des items de la V0 du test	Exemple Dimension : personnalité Vérification de la conformité au modèle scientifique retenu des 300 passations Analyse des 300 réponses apportées à chaque item	Elimination des items qui ne satisfont pas aux critères des analyses psychométriques Choix du nombre d'item, assemblage de la V1 du test	Récolte d'au moins 300 passations de la V1 du test. Analyses psychométriques de la V1 Si les analyses sont concluantes, le test est validé. Si non, on reprend à l'étape 2 pour les items qui posent problème

2.1 Etape 1 : le choix du modèle scientifique

Cela va sans dire, mais cela va mieux en le disant : un test en psychologie mesure une dimension « psychologique ». Par exemple, l'introversion vs extraversion, l'estime de soi, la perception temporelle, le sentiment d'efficacité personnelle, etc … sont des dimensions « psychologiques ». Alors que la taille, le poids, le nombre de frères et sœurs, la couleur de la voiture ou encore le nombre de personnes qui composent le foyer ne sont pas des dimensions psychologiques

De plus, ces dimensions psychologiques doivent avoir fait l'objet d'une validation scientifique. C'est à dire que leur existence a été prouvée par des démonstrations reconnues par l'ensemble de la communauté des scientifiques. Valider une dimension psychologique c'est expliquer, présenter, décrire les caractéristiques de cette dimension. Cette description est très importante car elle présente les caractéristiques qui feront l'objet de l'évaluation. Dit autrement, les caractéristiques de la dimension psychologique serviront à rédiger les questions du test. C'est le critère de **validité** du test.

Il faut également que ces dimensions psychologiques montrent une variabilité entre les personnes. Par exemple, certaines personnes sont **très très** extraverties et d'autres sont **légèrement** extraverties. Cette différence de niveau d'extraversion va se traduire dans le comportement des personnes. Dit autrement, s'il n'y avait pas de variabilité alors toutes les personnes auraient le même niveau et ce ne serait pas utile de construire un test puisque l'on connaîtrait le résultat sans avoir besoin de l'évaluer. Ici, le test doit être en capacité de restituer cette variabilité. C'est le critère de **sensibilité** du test.

Enfin, comme tout instrument de mesure, le test doit produire le même résultat lorsqu'il est utilisé plusieurs fois avec la même personne. C'est le critère de **fidélité** du test.

2.2 Etape 2 : la rédaction des items

Les items ce sont les questions du test. Les psychométriciens préfèrent le terme « item » car il existe des tests qui ne sont pas composés de questions rédigées sous forme de phrases. C'est le cas, par exemple dans un test de logique visuo-spatiale où les items sont souvent des images.

Le modèle scientifique retenu à l'étape 1 va définir les caractéristiques de la, ou des dimensions évaluées. Par exemple, le modèle de personnalité en 5 grandes dimensions dit « big five » défini les 5 dimensions suivantes : l'extraversion, l'agréabilité, la conscience, le névrosisme et l'ouverture. Pour chaque dimension, le modèle décrit les caractéristiques qui la composent. Par exemple, la dimension qui va de l'extraversion à l'introversion explique que : « *Les personnes introverties sont réservées, calmes et moins dépendantes de la vie sociale* ». A partir de ces caractéristiques, les concepteurs du test vont rédiger les items.

Ils vont commencer par décider de la forme de l'item : est-ce une question avec 2 réponses possibles (oui ou non), une question à choix multiples (oui, non, je ne sais pas, non concerné), une question avec réponse sur une échelle (tout à fait comme moi, légèrement comme moi, neutre, pas vraiment comme moi, pas du tout comme moi), etc … il existe un grand nombre de formes de réponse.

La rédaction de l'item va donc dépendre de la caractéristique à évaluer et de la forme de réponse retenue. Pour l'introversion, par exemple, ils pourraient rédiger cet item : « l*e vendredi soir, après une semaine de travail, préférez-vous ? Réponse 1 : sortir en discothèque – Réponse 2 : rester à la maison et regarder un DVD* ». Cet item aurait pour objectif d'évaluer la caractéristique « *moins dépendantes de la vie sociale* ». Ils vont procéder de même pour toutes les caractéristiques de la dimension à évaluer. Donc pour la seule dimension « Extraversion vs introversion », ils vont rédiger un grand nombre d'items.

Le nombre total d'items dépend de l'objectif de nombre de questions pour la forme finale du test. Par expérience, on estime que 7 items est un nombre minimum pour évaluer correctement une dimension. Les concepteurs du test vont donc en rédiger 14 pour ne conserver que les 7 « meilleurs » pour la forme finale du test.

Zoom sur un ressenti ...
Souvent, les personnes qui répondent à un test psychométrique, ont l'impression d'avoir déjà lu une question. Elles se disent alors que le test essaie de les « piéger » en leur reposant une même question pour vérifier la cohérence. En réalité, le test ne cherche pas à « piéger ». C'est juste parce qu'il a besoin de plusieurs items qui évaluent la même caractéristique, pour faire une bonne mesure. C'est ce principe qui explique cette potentielle redondance.

2.3 Etape 3 : la version « 0 » du test

A ce stade, les concepteurs du test vont décider de l'assemblage des items. En effet, on a vu qu'il y avait plusieurs dimensions, comme dans le test de personnalité en 5 grandes dimensions, et plusieurs items par dimension. Il convient donc de décider si les items sont présentés :

1. dans l'ordre des dimensions
2. complètement aléatoirement
3. avec une certaine logique comme par exemple le 1er item de chaque dimension, puis le second de chaque dimension, etc …
4. ou tout autre forme de présentation...

Le choix pour cet assemblage, c'est à dire l'ordre dans lequel les items vont être présentés, est important car il peut avoir un impact sur la mesure réalisée. En effet, comme nous l'avons vu précédemment, plusieurs items vont évaluer la même dimension. Ils vont donc « se ressembler ». Et cette ressemblance peut apparaître à l'esprit du répondant lorsque les items sont lus à la suite. Le fait que cet assemblage puisse mettre « la puce à l'oreille » du répondant introduit un potentiel biais dans la mesure car, en fonction du répondant, ce dernier peut également ne pas percevoir cette ressemblance. Deux répondants différents pourraient alors se trouver dans deux situations différentes face à la même mesure. Cela viendrait perturber la standardisation de la passation.

Il est donc fréquent pour les tests avec items sous forme de question rédigée, de ne pas présenter les items d'une même dimension les uns à la suite des autres.

Les concepteurs vont ensuite rédiger la consigne. C'est à dire le texte de présentation du test et les indications pour répondre aux items. Ces indications sont dépendantes de la forme choisie pour les réponses aux items. Bien souvent ici, un item d'exemple est introduit. Cet item est propre à l'exemple, il ne s'agit pas d'un des items qui se trouvent dans le test.

Cette V0 est alors proposée à des personnes qui correspondent à la population cible du futur test finalisé. En effet, ici encore cela va sans dire mais cela va mieux en le disant, des items rédigés en français n'ont pas pour population cible les Allemands du Baden-Württemberg. Sans être aussi caricatural, si le test final est destiné à une population de jeunes de moins de 26 ans par exemple, alors cette V0 ne doit pas être présenté à des adultes de plus de 26 ans.

L'objectif est de récolter un minimum de 300 répondants. Et ces 300 feuilles de réponses serviront à réaliser les analyses psychométriques. Les 300 personnes qui répondent doivent donc être similaires à la population cible de la forme finale du test.

2.4 Etape 4 : les analyses psychométriques

Les concepteurs détiennent au moins 300 passations de la V0 de leur questionnaire. Ils vont étudier et décrire les caractéristiques de la population qui a répondu : classe d'âge, genre, niveau de qualification et toute caractéristique permettant de décrire les répondants. Ils vont ensuite calculer les scores à chaque dimension. A ce stade, on parle de **score brut**. Ce calcul dépend de la forme du test, mais bien souvent il consiste à faire la somme des réponses aux items qui composent chaque dimension.

Illustration ...
Une dimension **extraversion vs introversion** composée de 14 items et d'une échelle de réponse en 5 points (de « pas du tout comme moi » à « tout à fait comme moi ») :
- chaque point de l'échelle de réponse est traduite en chiffre : « pas du tout comme moi » = 0 à « tout à fait comme moi » = 4 et les pas intermédiaires en 1, 2 et 3
- les réponses aux 14 items sont additionnés : le score minimum théorique est 14 x 0 = 0 et le score maximum théorique est 14 x 4 = 56

Parler avec les autres aide à explorer les croyances personnelles :

Pas du tout comme moi	Pas vraiment comme moi	Neutre	Légèrement comme moi	Tout à fait comme moi
0	1	2	3	4

On le comprend, un test psychométrique est avant tout une démarche **quantitative**. Le texte de la question est une sorte de prétexte pour déclencher un comportement de positionnement (de 0 à 4 par exemple).

Les analyses psychométriques sont des procédures statistiques. Elles commencent par des procédures de statistiques descriptives sur la distribution des scores : mode, moyenne, écart-type, étendue de distribution, quartiles, etc … Les concepteurs du test vont s'assurer que toute l'étendue des scores théoriques possibles est utilisée par les 300 répondants. Cela renvoie à la notion de **variabilité** présentée à l'étape 1.

Puis les concepteurs vont réaliser des statistiques de type analyses factorielles exploratoires sur les items pour analyser leur répartition dans chacun des facteurs. Il s'agit de s'assurer que tous les items d'une même dimension se retrouvent bien dans un même facteur. Mais également que ces mêmes items « saturent » le moins possible sur les autres facteurs. Ce principe illustre la caractéristique d'un item qui mesure la dimension pour laquelle il est conçu et qu'il ne mesure pas une autre dimension.

Ensuite, les concepteurs du test vont utiliser la procédure « Alpha de Cronbach » pour vérifier la cohérence d'ensemble des items d'une même dimension. Si la valeur de l'Alpha n'est pas assez élevée alors il faut retravailler la liste des items. Pour cela, on va généralement supprimer l'item qui contribue à faire baisser la valeur de l'Alpha. En procédant par itération, on ne retient au final que la liste des items dont l'ensemble présente la meilleure caractéristique psychométrique. Ici, deux grands principes sont utilisés :

- la corrélation entre les réponses aux items de la dimension : si 2 items mesurent de l'introversion alors pour une même personne, les deux réponses doivent varier dans le même sens. Si pour les 300 personnes, les 2 items varient toujours dans le même sens alors cela montre que ces 2 items sont fortement corrélés
- la contribution des réponses à l'item au score global de la dimension : ce principe est lié au précédent. Pour les 300 répondants, la valeur qu'ils attribuent à l'item doit contribuer de manière homogène au score global qu'ils obtiennent à la dimension.

On comprend donc que ce n'est pas le texte de la question qui est utilisé pour analyser la qualité de l'item, mais le comportement d'évaluation que ce texte engendre. Ce comportement est comparé aux autres comportements d'évaluation de la même personne sur tous les items qui évaluent la même dimension. Ce principe est appliqué aux 300 répondants. Si tous les comportements d'évaluation sont cohérents alors les items concernés sont qualifiés de « bon ». Si tous les comportements d'évaluation sont anarchiques et incohérents alors les items concernés sont qualifiés de « mauvais »

Zoom sur un ressenti ...
Souvent, les personnes qui répondent à un test psychométrique, estiment qu'un item est mal formulé. Que la rédaction est « mauvaise ». En réalité, comme on le voit dans la description des analyses psychométriques, ce sont les 300 réponses des personnes et leur cohérence qui prouvent que l'item ainsi rédigé mesure bien ce qu'il est censé mesurer. Les items qui seraient « mal compris » par les répondants auraient 300 réponses incohérentes qui feraient baisser la valeur de l'Alpha et seraient éliminés. Même si, d'un point de vue personnel (individuel), la rédaction d'un item ne nous plaît pas, globalement (collectivement) l'item est bien compris et provoque un comportement d'évaluation cohérent chez l'ensemble des répondants

2.5 Etape 5 : la forme finale du test

Tous les « mauvais items » étant supprimés, chaque dimension peut alors être composée d'un nombre inégal d'items. En effet, ce sont les analyses psychométriques qui révèlent la qualité d'un item et donc la façon dont les 300 répondants ont choisi de répondre. Le nombre de « mauvais item » n'est donc pas égal sur chaque dimension. Et les items ne sont donc pas éliminés de manière homogène sur toutes les dimensions

Bien que ce ne soit pas une obligation, les concepteurs du test vont généralement uniformiser le nombre d'items par dimension évaluée. Il reste donc à éliminer les plus mauvais des « bons items » pour avoir un nombre égal d'item pour chaque dimension évaluée.

La consigne du test est reprise en l'état ou éventuellement ajustée en fonction des retours des 300 premiers répondants. La V1 du test est alors prête pour la nouvelle phase de validation.

En effet, on pourrait penser que les items étant validés individuellement, la forme finale ne contient que de « bons » items et elle peut être utilisée en l'état. Mais il n'en est rien car un test psychométrique forme un tout homogène. C'est à dire que le fait de lire un item et réfléchir à son choix de réponse peut agir sur le raisonnement du répondant et déteindre sur ses réponses suivantes. Donc le fait d'avoir éliminé des items agit sur le contenu global de l'outil.

On va donc, à nouveau, proposer cette V1 à un panel d'au moins 300 personnes représentatives de la population à qui est destiné le test.

2.6 Etape 6 : la validation de la forme finale du test

Après la récolte des 300 nouvelles passations, on va renouveler les mêmes analyses psychométriques. En fonction des résultats de ces analyses, différentes situations peuvent se présenter. Le meilleur des cas étant que cette V1 réponde parfaitement aux critères de validation d'un test psychométrique. Dans ce cas la V1 est aussi la forme finale du test.

Dans tous les autres cas, on va procéder aux ajustements nécessaires et refaire la collecte de 300 passations pour, à nouveau vérifier les qualités psychométriques de l'outil. Et l'on va répéter cette itération autant de fois que nécessaire pour atteindre un niveau de qualité psychométrique satisfaisant.

A noter que depuis le milieu des années 2000 environ, de nouvelles méthodes statistiques sont venues renforcer l'exigence de rigueur scientifique des tests psychométriques. Il s'agit des méthodes de modélisation par équation structurelle, aussi appelées SEM dans le monde anglophone pour « Structural Equation Modeling ». Pour le dire simplement, il s'agit de faire une analyse factorielle confirmatoire en lieu et place de l'analyse factorielle exploratoire de l'étape 4.

Une analyse factorielle **exploratoire** consiste à identifier, dans l'ensemble des 300 feuilles de réponses, des regroupements d'items qui appartiennent à une même dimension sans à priori sur ces dimensions. Et l'on vérifie ensuite que les dimensions identifiées correspondent bien à celles du modèle scientifique de référence (de l'étape 1).

Une analyse factorielle **confirmatoire** fonctionne, pour ainsi dire en sens inverse. Il s'agit de poser un modèle à priori ; en l'occurrence le modèle scientifique de départ en indiquant l'appartenance de chaque item à la dimension qui lui correspond. Puis l'analyse factorielle confirmatoire va indiquer dans quelle mesure les 300 passations s'accordent avec le modèle ainsi défini.

Cette analyse factorielle confirmatoire est mise en œuvre sur la forme finale du test après que tous les ajustements aient progressivement amélioré les qualités psychométriques de l'outil

Enfin, à partir des 300 dernières passations, les concepteurs du test vont produire les tables d'étalonnages. Il s'agit des scores bruts transformés en **scores étalonnés** pour permettre de situer la personne qui a répondu au test par rapport à la population qui lui ressemble. Ces scores étalonnés permettent de « *donner du sens* » aux scores bruts obtenus par le répondant. Pour en savoir plus sur l'étalonnage, se reporter à la fiche pédagogique dédiée à ce sujet.

3 Réponses aux questions fréquentes

Foire aux questions fréquentes de mes étudiants, alimentée au fur et à mesure des promotions devant lesquelles j'interviens

3.1 Question N°1 : *« Quand j'étais en stage, j'ai utilisé des tests et je me suis rendu compte qu'il y en avait un qui datait de 1992 ! Est-ce bien sérieux d'utiliser des tests aussi vieux ? »*

Un test évalue différentes dimensions comme on le découvre dans cette fiche. Si on reprend le modèle de la personnalité, les études scientifiques ont énormément fait progresser la connaissance que nous avons du modèle descriptif. Aujourd'hui la communauté scientifique s'accorde autour du modèle en 5 grands facteurs. Il serait donc effectivement peu sérieux d'utiliser un test de personnalité plus ancien que le « big five ».

Mais un test peut aussi mesurer une aptitude cognitive propre à l'homme et intrinsèquement présente depuis très longtemps. Prenons l'aptitude visuo-spatiale par exemple. Le cerveau de l'homme a la capacité de former des images mentales depuis la nuit des temps. On pourrait poser l'hypothèse que l'art pariétal paléolithique qui s'exprime sur les parois de la grotte de Lascaux a été réalisé par des hommes qui étaient capables de se représenter leurs dessins sous forme d'images mentales. Donc on peut tout à fait utiliser un test d'aptitude visuo-spatiale, même s'il est très ancien. Par contre il faut s'assurer d'utiliser des **étalonnages qui ont été actualisés récemment**. En effet, on sait que l'environnement a un impact sur les

aptitudes cognitives. Et notre environnement évolue rapidement. Il ne serait donc pas très logique de comparer un score d'une personne aujourd'hui avec une population de référence d'il y a 30 ou 40 ans.

3.2 Question N°2 : *« Les tests gratuits, disponibles sur internet, sont-ils de bons ou de mauvais tests ? »*

La gratuité n'est pas corrélée à la qualité. Il y a des tests « gratuits » de bien piètre qualité comme des tests « gratuits » de très grande qualité.

Un quizz proposé par une revue non spécialisée en psychométrie et souvent, avec plus de publicités à l'écran que de place laissée au questionnaire a peu de chance d'être un outil de qualité. A contrario, un site dédié, sans publicité outrageante et qui consacre une page à la présentation de son modèle scientifique et les qualités psychométriques de son outil gratuit peut tout à fait être un test de qualité.

En fait, il y a un principe fondamental dans l'usage des tests : le professionnel qui utilise un test psychométrique doit bien connaître et maîtriser l'outil qu'il utilise. Et cette maîtrise passe par 2 étapes incontournables :

- Passer lui-même le test concerné
- Bien comprendre le fonctionnement du test et ses qualités psychométriques en lisant le manuel technique

Donc un test, qu'il soit **gratuit** ou pas, doit toujours être accompagné de son **manuel technique** qui précise à minima : le **modèle scientifique** sur lequel il s'appuie et les **analyses psychométriques** de validation.

Utiliser un test, gratuit ou pas, sans manuel technique c'est comme sauter en parachute … sans parachute ...