Validité et fiabilité des questionnaires d'évaluation de la qualité de vie : une étude appliquée aux accidents vasculaires cérébraux

Fabienne Midy

LATEC

LABORATOIRE D'ANALYSE ET DE TECHNIQUES ÉCONOMIQUES

U.R.A. 342 C.N.R.S.

DOCUMENT de TRAVAIL

UNIVERSITE DE BOURGOGNE

FACULTE DE SCIENCE ECONOMIQUE ET DE GESTION

4, boulevard Gabriel - 21000 DIJON - Tél. 80 39 54 30 - Fax 80 39 56 48

ISSN : 0292-2002

n° 9612

Validité et fiabilité des questionnaires d'évaluation de la qualité de vie. Une étude appliquée aux accidents vasculaires cérébraux

Fabienne MIDY*

septembre 1996

*LATEC (UMR 5601 - CNRS), Université de Bourgogne

Abstract.

This article deals with various means, which could be used by designers in order to estimate error terms in questionnaire tests results, thanks to psychometry. The study must be conceived so as to minimize systematic errors (validity) and random errors (reliability).

Methods settled to test the validity of a study are very subjective, even if some of them seem to be more "scientific", being based on the principle of correlation. However the last-mentined don't stand up to some theoretic weakness and other explicit subjective methods should be chosen instead : for instance experts consensus.

Methods of reliability evaluation have been developed under strong hypothesis to estimate a reliability coefficient. A very commonly-used method is Cronbach's alpha, but some criticisms have spread concerning the right interpretation of the coefficient.

The article concludes on the incompleted aspect of the theory of error measurement. However it's necessary to go on justifying each study thanks to the means which are available : that is to say the expert consensus to validate the choose of items and the alpha coefficient to attest their reliability. Nevertheless, we must be careful when interpreting the results. The last part of this work deals with the empirical results which have been achieved thanks to these methods in a survey about the quality of life post stroke.

Key-words : psychometric theory - validity - reliability - quality of life - stroke.

Résumé.

L'article discute des moyens mis à la disposition des concepteurs de questionnaires d'évaluation de la qualité de vie par la psychométrie pour évaluer l'importance des erreurs de mesure dans les résultats. L'enquête doit être construite de manière à minimiser les erreurs systématiques (validité) et les erreurs aléatoires (fiabilité).
Les méthodes développées pour tester la validité d'une étude sont très subjectives, même si certaines ont tenté de revêtir une dimension plus "scientifique" en se fondant sur le principe de corrélation. Ces dernières ne résistent pas à certaines faiblesses théoriques, et il est préférable de leur préférer des méthodes explicitement subjectives pour le choix des items d'un questionnaire, telles que le consensus d'experts.
Des méthodes d'évaluation ont été développées sous certaines hypothèses, pour estimer un coefficient de fiabilité. La technique du coefficient alpha de Cronbach est la plus couramment utilisée. De conception correcte, elle est cependant l'objet de certaines critiques quant à l'interprétation exacte du coefficient qu'elle permet de calculer.
L'article conclut sur le caractère inachevé de la théorie des erreurs de mesure. Pratiquement, il est toutefois nécessaire de continuer à justifier toute étude à l'aide des

moyens mis à notre disposition : le consensus d'expert pour valider le choix des items et le coefficient alpha pour attester de leur fiabilité, tout en accordant une certaine prudence quant à l'interprétation de leurs conclusions. Dans la dernière partie de ce travail, nous présentons les résultats empiriques que nous avons obtenus à partir de ces méthodes, dans le cadre d'une étude d'évaluation de la qualité de vie post-AVC (Accidents Vasculaires Cérébraux).

Mots-clés : théorie psychométrique - validité - fiabilité - qualité de vie - accident vasculaire cérébral.

TABLE DES MATIERES

La méthodologie de l'évaluation d'une pathologie ou d'une thérapeutique a évolué au cours des deux derniers siècles. Les premières évaluations étaient réalisées sur la base d'un critère quantitatif qui s'exprimait le plus souvent en termes d'espérance de vie. Aujourd'hui, la recherche médicale intègre dans ces objectifs non seulement la survie des patients, mais également leur mieux-être, ceci étant particulièrement frappant en ce qui concerne la recherche gériatrique. La prise en compte explicite de la qualité de vie pose le problème de sa définition (Baldwin, 1990 ; Bucquet, 1993 ; Patrick, 1995). La qualité de vie est en premier lieu un concept multidimensionnel. Elle recouvre "tout ce qui fait que la vie est ce qu'elle est", décrivant ainsi ce que nous appellerons un état de vie. La qualité de vie est ensuite un concept subjectif, qui comporte un jugement de valeur sur cet état de vie. **On peut alors définir la qualité de vie comme la propriété d'être bon ou mauvais d'un état de vie**. Son évaluation consiste donc à demander à un individu de porter un jugement de valeur sur un état de vie défini.

Parmi les méthodes d'évaluation de la qualité de vie, celle du questionnaire permet d'intégrer ces deux dimensions descriptive et évaluative de façon simultanée grâce à sa structure détaillée et modulable. De nombreuses études ont été conduites dans le domaine de la qualité de vie à partir de cet instrument, donnant naissance à une multitude de questionnaires de conceptions différentes. Devant la prolifération de ces études, on peut s'interroger quant à leur qualité en évaluant la part d'erreur de mesure, d'ailleurs inévitable, dans les résultats obtenus.

L'évaluation des propriétés psychométriques d'un instrument de mesure doit permettre de s'assurer de la non significativité des erreurs de mesure, en distinguant les erreurs systématiques et les erreurs aléatoires. Les propriétés psychométriques correspondantes sont respectivement la validité et la fiabilité.

Ce document présente, dans une première partie, les différentes méthodes permettant de tester les propriétés psychométriques d'un instrument de mesure. L'exposé de leurs fondements théoriques va nous permettre de mettre en lumières les nombreuses critiques qui réduisent l'applicabilité des techniques d'évaluation de la validité et de la fiabilité. Une revue critique de ces méthodes nous a semblé nécessaire afin d'éviter leur utilisation abusive pour justifier une étude. Nous avons ensuite regroupé dans une deuxième partie, les résultats que nous avons obtenus dans le cadre d'une évaluation de la qualité de vie post-AVC (Accidents Vasculaires Cérébraux), pour laquelle un questionnaire spécifique à été élaboré.

PARTIE 1 : FONDEMENTS THEORIQUES ET METHODES.

Chapitre 1 : Généralités.

Une réflexion antérieure sur la définition du concept de qualité de vie nous a conduit à rechercher un instrument pouvant recueillir **des jugements de valeur** (aspect **subjectif** de la qualité de vie) auprès des **patients** eux-mêmes (aspect **personnel**). Un tel instrument est-il concevable?

Dès la fin du 19ème siècle, des **expériences psychophysiques**, c'est-à-dire qui étudient la façon dont les gens perçoivent des phénomènes physiques, ont montré qu'une personne peut faire des estimations numériques correctes et cohérentes à propos de phénomènes tels que l'intensité d'une lumière ou la tonalité d'un son. Plus tard, la **psychométrie** a adapté les méthodes psychophysiques afin de **mesurer des stimulus qui ne se réfèrent pas à une échelle physique**. Cette transposition repose sur des expériences démontrant qu'une "personne pouvait faire des jugements subjectifs de façon remarquablement cohérente, même lorsqu'il lui est demandé de faire des comparaisons abstraites" (Mc Dowell I., Newell C., 1987).

Cependant, nous voulons attirer l'attention sur le domaine de la santé, où nous pouvons douter de l'adéquation parfaite entre le stimulus (l'état de santé) et la réponse du patient :

- d'une part, étant donné l'abstraction du concept de qualité de vie et le flou qui règne quant à sa définition, on ne peut jamais être sûr que le stimulus tel qu'il est élaboré par le chercheur sera le même que celui auquel répondra le patient ;
- d'autre part, le jugement de valeur ne concerne plus un stimulus externe, mais un stimulus interne au sujet. Celui-ci se sentant impliqué dans sa réponse peut alors l'exagérer ou au contraire la minimiser selon les enjeux, que ce soit de façon consciente ou non.

Les réponses recueillies dans une enquête sur la perception de la santé en termes de qualité de vie peuvent ainsi être biaisées pour un certain nombre de raisons.

§1 : Les différents types de biais .

Les sources de déviation des réponses par rapport à l'état ressenti effectif sont nombreuses et variées, et elles font l'objet d'une abondante littérature tant d'un point de vue théorique que pratique (Streiner et Norman, 1989; Dickes, 1994). On peut cependant répertorier trois sources d'erreur de mesure principales : les items, les sujets et les observateurs. Nous nous contenterons dans les paragraphes qui suivent de citer quelques exemples pour chaque cas.

Les items ne sont pas neutres dans le choix de réponse du sujet :

- *Effet de caractéristiques formelles* : la complexité, la longueur, la forme négative ou positive...)
- *L'effet scénario ("framing effect")* décrit par Kahneman et Tversky (Kahneman et Tversky, 1984) : le choix d'une personne entre deux alternatives dépend de la façon dont les alternatives sont présentées (par exemple en termes de réduction des décès ou d'accroissement de vies sauvées).
- *Le dispositif de réponse* : échelle, item à choix multiple.
- *L'effet contexte* : concerne l'instrument de mesure dans son ensemble et non plus les items pris isolément (place des questions les unes par rapport aux autres, regroupement des items concernant un même trait ou présentation dispersée).

Influence du sujet ou "styles de réponse":

Les sujets répondent non pas en fonction du contenu de l'item, mais en fonction de tendances plus ou moins marquées.

- *La désirabilité sociale* : l'individu ne cherche pas à mentir de façon délibérée, mais il oriente ses réponses de façon à correspondre aux normes collectives de la société dans laquelle il vit.
- *Le trucage: positif* : la personne crée une impression positive de façon délibérée (par exemple pour pouvoir sortir de l'hôpital plus vite).
- *Le trucage négatif* : la personne va exagérer les symptômes, par exemple pour éviter le service militaire ou avoir un arrêt de travail.
- *La déviation* : tendance naturelle à donner une réponse légèrement déviée.
- *L'acquiescement* : tendance naturelle à répondre de façon positive.
- *L'aversion des extrêmes* : le sujet a des difficultés à faire des jugements extrêmes et n'utilise pas les bornes de l'échelle.
- *La distribution oblique* : la distribution des réponses sur les différentes modalités n'est pas régulière et entraîne un déplacement non représentatif de la moyenne.
- *Le halo* : le jugement demandé sur un aspect seulement de l'individu est influencé par l'aspect général de l'individu.

On peut également citer les biais attribuables à des facteurs personnels transitoires (humeur, fatigue...).

L'influence de l'observateur et de la situation.

- La situation dans laquelle l'enquête est menée : domicile, présence d'un tiers.
- La confiance qui s'instaure avec le sujet (souvent fonction de l'âge, du sexe..)
- L'observateur peut diriger les réponses du sujet, même inconsciemment.
- Les facteurs mécaniques (mauvais report des réponses, scorage...).

Il serait donc illusoir de croire que l'on peut avoir adéquation complète entre la valeur observée et la vraie valeur de l'objet étudié. Les différents effets déformants, dont nous venons de voir des exemples, peuvent être catégorisés en deux groupes, en fonction du type d'erreurs qu'ils produisent.

§2 : Les différents types d'erreurs.

Elles sont de deux sortes : **les erreurs aléatoires** et **les erreurs systématiques ou biais**. Pour illustrer la différences entre ces deux catégories d'erreur, nous reprendrons l'exemple d'une mesure de température avec un thermomètre, tel qu'il est décrit par Nunnaly :

"Il y aura erreur systématique dans le cas où un chimiste n'a qu'un thermomètre à sa disposition, qui indique toujours deux degrés de plus que ce qu'il devrait indiquer, et cela bien que le chimiste en fasse une lecture précise. L'erreur aléatoire apparaît lorsque le thermomètre est juste, mais que le chimiste étant atteint de myopie est incapable de lire correctement le résultat au cours des différentes mesures. A cause de sa vision trouble, le chimiste peut enregistrer indifféremment des résultats en dessus ou en dessous de ce qu'ils sont réellement." (Nunnaly J., 1978).

Dans chaque mesure, il y a une part d'erreur constante et une part d'erreur variable. A partir de cet exemple, on peut formaliser le modèle qui servira de base à la théorie des erreurs de mesure :

$$Y=T+E+B$$

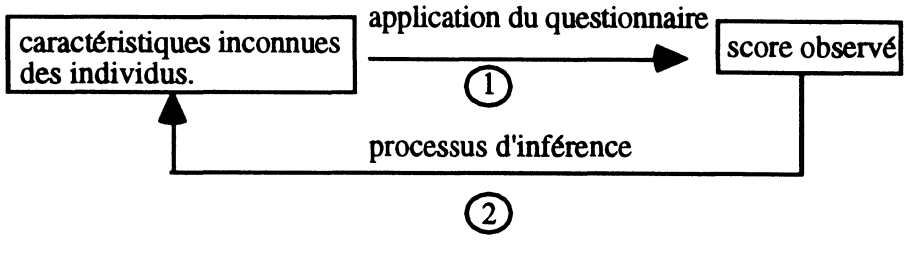
où Y est la valeur observée, la mesure,
T est la valeur réelle,
E est l'erreur aléatoire,
B est l'erreur systématique.

Au travers de ce modèle, il devient évident que l'on doit vérifier les propriétés d'une mesure afin de pouvoir utiliser ses résultats pour faire des conclusions sur l'objet de l'étude. Lorsque l'erreur aléatoire est faible, on dira que la mesure est fiable, lorsque l'erreur systématique est faible, on dira que la mesure est valide. **Un peu à la manière d'un archer, le chercheur doit s'assurer que la cible visée est la bonne (validité) et qu'il peut l'atteindre (fiabilité).**
Pour ce faire, il existe des méthodes parfois critiquables, permettant de tester la validité, et la fiabilité d'un instrument de mesure.

Chapitre 2 : La validation.

Rappelons tout d'abord qu'**une mesure est dite valide lorsque la part du biais dans les résultats obtenus est peu importante**. Cela revient à se demander si la cible visée par l'instrument est effectivement celle que l'on s'était fixé comme objectif de l'étude : "D'une façon générale, un instrument de mesure est dit valide, s'il mesure ce pour quoi il a été construit" (Nunnaly J., 1978).

Au cours de ces vingt dernières années, on est passé de la notion de validité à la notion de validation. Au delà de la terminologie, ce changement marque une évolution théorique très importante. Avant les années soixante-dix, on concluait sur les propriétés psychométriques de l'instrument lui-même. Depuis Cronbach (1971), la validation change d'objet : on met en avant les hypothèses qui permettent de déterminer les caractéristiques inconnues des individus à partir des résultats observés au cours de l'étude : **"On valide, non pas un instrument, mais l'interprétation de données produites par une procédure particulière"** (Cronbach L.J., 1971). Essayons pour simplifier, de schématiser le déroulement d'une évaluation.



On peut séquencer une mesure en deux étapes principales:
- étape 1 : collecte des observations manifestes par un instrument de mesure;
- étape 2 : inférence des caractéristiques latentes à l'aide d'un système d'hypothèses.
Ces deux étapes constituent le processus de mesure qui relie des concepts abstraits à des indicateurs empiriques.

Pour donner un exemple pratique, imaginons que l'on veuille déterminer la sévérité de l'arthrite d'un groupe de patient. Le degré de sévérité de l'arthrite est donc la caractéristique inconnue que l'on veut évaluer. Pour cela on met en place un test (étape 1), qui consistera à faire soulever un poids par le patient. Les résultats seront alors interprétés à l'aide d'une batterie d'hypothèses (étape 2) afin d'en déduire le degré d'arthrite. Dans ce cas, on suppose que plus le poids soulevé est faible, plus l'arthrite est importante.

La première remarque que l'on peut faire aux vus de cette nouvelle définition, c'est que "Strictement parlant, on ne valide pas un instrument de mesure mais plutôt l'utilisation qui va en être faite" (Nunnaly J., 1978). C'est pourquoi **une étude de**

validation doit précéder toute enquête, même si on utilise un instrument qui a déjà été validé dans d'autres circonstances.

On peut ensuite observer deux avantages à une réflexion en termes de validation. D'une part, elle valide le processus de mesure dans son ensemble, c'est-à-dire les deux étapes. En effet, lorsque la validation est interprétée en termes d'instrument, seule l'étape 1 est validée. Mais si on valide le corpus d'hypothèses à partir desquelles on a élaborer l'instrument (étape 2), celui-ci se trouve indirectement validé. En effet, c'est parce que l'on a fait l'hypothèse que la force déclinait avec l'arthrite, que l'on a mis au point un test consistant à soulever un poids.

D'autre part, une interprétation dirigée non vers l'intégrité d'un questionnaire, mais vers les inférences qui peuvent être faites à partir des résultats de l'étude, permet une plus grande applicabilité de l'outil. Il pourra être utilisé à d'autres fins que celles pour lesquelles il a été construit, à partir du moment où l'on a pu démontrer que les déductions qu'il permet de faire, même sorti de son contexte d'origine, sont correctes.

Valider une étude ne doit plus se limiter à la question de savoir si l'instrument mesure effectivement ce pour quoi il a été construit. Le test des hypothèses sous-jacentes à l'instrument a pour objectif de justifier l'évaluation de caractéristiques latentes à partir d'observations manifestes. Pour répondre à cette question, trois types de validité sont couramment définis dans la littérature : la validité de contenu, la validité de critère et la validité de construit.

§1 : La validité de contenu (*content validity*)

L'objectif est de caractériser un individu à l'intérieur d'un domaine particulier. Bien souvent, le domaine auquel on fait référence est trop vaste pour que l'on puisse envisager de tester l'individu sur tous les aspects du domaine. Il faut donc faire un échantillonnage d'items ou d'épreuves, **qui permettrons de faire les bonnes déductions en ce qui concerne les caractéristiques latentes de l'individu**. Si c'est le cas, on parlera de validité de contenu. La validité de contenu s'évalue à partir de la représentativité (pertinence et exhaustivité) des éléments qui composent l'instrument de mesure, par rapport au concept mesuré. Si on oublie un, ou plusieurs, aspects caractéristiques du domaine d'intérêt dans la construction du test, on peut ensuite être amené à faire de fausses déductions. En revanche, si la validité de contenu est importante, cela montre que les bons résultats de l'étude ne sont pas dus aux circonstances particulières dans lesquelles elle a été menée, mais qu'il sont dus à une bonne spécification du domaine. L'étude pourra donc être reproduite dans des situations différentes, sans que cela ne remette en question la portée des hypothèses de déduction qui la fondent.

Malheureusement, il est difficile de s'assurer de la validité de contenu d'une mesure car il n'existe pas de méthode scientifique pour démontrer l'adéquation d'un test

(ensemble d'items ou d'épreuves) avec le but de l'étude, tel qu'il a été spécifié *a-priori*. "La plupart du temps, la validité de contenu n'est pas testée formellement : il peut être impossible de montrer de façon concluante que les items choisis sont représentatifs de tous les items possibles" (McDowell I., Newell C., 1987). En fait, il s'agit plutôt d'**une démarche subjective** basée sur la connaissance que possède le chercheur de l'univers d'étude, il est donc recommandé, dans un premier temps, de mener une recherche bibliographique poussée sur le sujet, afin de pouvoir spécifier le plus précisément possible la structure du domaine étudié. Il s'agit d'identifier de la façon la plus détaillée possible ce qui le compose. Une fois ce premier travail de tri effectué, il est nécessaire de s'adresser à des personnes compétentes qui pourrons confirmer, compléter ou réduire cette première sélection. Ces personnes peuvent être des professionnels de la pathologie ou, plus directement, les patients concernés. L'important est d'arriver à un consensus, "the acceptance of the universe of content as defining the variable to be measured is essential" (Cronbach L.J., 1955).

Cette étape essentielle est le prélude incontournable de toute étude d'évaluation. Cependant, son manque de rigueur a poussé les chercheurs à développer d'autres techniques de validation qui, bien qu'elles portent des noms différents, nous semblent tendre vers les mêmes conclusions que la validité de contenu. On peut classer ces techniques en deux catégories, la validité de critère (*"criterion validity"*) lorsqu'elles se réfèrent à un étalon ou *"gold standard"* et la validité de construit (*"construct validity"*) lorsqu'il n'en existe aucun.

§2 : La validité de critère (*criterium validity*).

"La validité de critère est la propriété d'une méthode (...) qui mesure correctement le phénomène étudié en référence à un critère. Par définition, si le critère doit faire office de norme à fin de vérification, il doit être une mesure supérieure et très fidèle du phénomène étudié : un '*gold standard*'." (Coste, 1995). C'est le cas le plus simple où il existe une référence avec laquelle on va pouvoir comparer les résultats attachés à un instrument. Ces références peuvent être une étude antérieure sur le même sujet, une population (validité discriminante) ou les valeurs effectives futures (validité de prédiction). Ces méthodes sont perçues comme étant plus scientifiques car elle reposent sur les techniques statistiques de la corrélation. Nous verrons cependant que, malgré cela, elles ne résistent pas à des critiques à la fois pratiques et logiques.

- Lorsqu'elle existe, on peut faire référence à une étude qui a été utilisée précédemment dans le même domaine et qui a été validée à cette occasion. Pratiquement, on applique les deux instruments de façon simultanée sur la même population. Les résultats doivent présenter une forte corrélation pour que l'on puisse **déduire de la validité de l'instrument de référence, la validité du nouvel outil**. Bien

qu'elle se fonde sur des techniques statistiques confirmées, cette méthode n'apporte pas toujours de solutions aux lacunes de la validité de contenu.

En effet, concrètement, il n'est pas toujours possible de mener deux enquêtes simultanément. Le coût en temps, et en termes de taux de non-réponse (étant donné la longueur cumulée des deux instruments), peut rendre cette technique de validation totalement impraticable.

Conceptuellement, cette technique peut être discutable dans la mesure où elle repose sur l'interprétation "ancienne" de la validité. Or si on admet que ce n'est pas l'instrument que l'on valide, mais l'utilisation qui en est faite, dès que l'on sort l'instrument de son contexte (lieu géographique, caractéristiques socioprofessionnelles de la population, époque...), il faut de nouveau étudier la pertinence de son utilisation. Paradoxalement, cette méthode propose de valider un instrument à partir d'un autre instrument dont on ne peut prouver la validité. Finalement, cette méthode n'est valable que si la validité de contenu de l'instrument de référence est parfaite, et que l'on peut donc l'utiliser dans des situations différentes, sans remettre en cause sa validité. On retrouve alors un problème de validation de contenu.

- Lorsque l'étude a un objectif de discrimination, c'est-à-dire la classification des individus en fonction d'un critère donné, on se sert d'une population de référence dont on connaît parfaitement les caractéristiques. Cette population doit être composée de deux groupes, le groupe A qui possède le critère discriminatif et le groupe B qui ne le possède pas. On compare les résultats obtenus afin de vérifier que l'instrument possède effectivement un pouvoir discriminatif, c'est-à-dire qu'il est **sensible et spécifique**. Pour évaluer sa sensibilité, on teste son aptitude à révéler le critère dans la population A ; pour évaluer sa spécificité, on teste son aptitude à ne pas révéler le critère dans B.

Cette méthode semble méthodologiquement correcte, cependant elle est d'un intérêt limité, puisque la connaissance parfaite de la population de référence implique qu'il existe déjà un instrument discriminatoire. L'intérêt d'un nouvel instrument ne peut s'expliquer que par un souci de simplification (questionnaire plus court, plus compréhensible...) ou par une baisse des coûts d'utilisation de l'instrument.

- Lorsque l'instrument est développé dans un but prédictif, on fait référence à un critère qui ne sera disponible que dans le futur. La validité de prédiction de notre cadre d'hypothèse est établie si la comparaison de la prédiction et des résultats *ex-post* ne montre pas de différence significative. L'avantage de cette méthode est qu'elle ne réclame pas l'évaluation de la validité du critère puisqu'il est constitué de la valeur réelle future de l'objet. On peut toutefois remarquer que le délai nécessaire pour obtenir la valeur future peut devenir prohibitif lorsqu'il est trop long.

Toutes ces méthodes de validation font référence à un *"gold standard"*. Lorsqu'aucun étalon n'est disponible, on utilise les techniques de validation de construit.

§3 : La validation de construit (*construct validation*).

On utilise ce type de validation si l'étude porte sur un domaine abstrait, lui-même difficile à spécifier. Dans ce cas, puisqu'il n'existe pas de définition unique du domaine d'étude, il ne peut pas exister de *"gold standard"* universel.
La première étape consiste à établir un modèle constitué d'hypothèses qui permettent de définir le domaine abstrait, c'est ce modèle que l'on appelle un construit : **"Un construit est une représentation théorique du domaine que l'on veut mesurer"** (Guyatt G, 1993). L'instrument est étudié sur la base de ce construit. Campbell et Fiske (Campbell et Fiske, 1959) ont défini la validité de construit par deux concepts : la validité de convergence et la validité de divergence, et en ont proposé une méthode de test : la technique multitrait-multiméthode.

3.1 : Validité de convergence et de divergence.

La validité de convergence indique que la mesure d'un construit est indépendante du processus de mesure qui a été suivi, c'est-à-dire que deux mesures par des instruments différents d'un même concept doivent aboutir au même résultat. Le principe technique repose sur un calcul de corrélation entre les résultats de l'étude à valider et les résultats d'une étude valide sur le même domaine. Les deux instruments sont utilisés simultanément sur la population cible. On teste alors l'hypothèse selon laquelle les résultats de l'étude à valider sont corrélés avec les résultats de l'étude de référence.

La validité de divergence vérifie que le construit mesuré ne recouvre pas un autre construit existant. Dans ce cas, soit les deux construits concernent effectivement un même domaine et ils sont donc redondants, soit ils ne devraient pas recouvrir le même domaine, auquel cas l'un des deux construits est mal spécifié. La technique repose sur le même principe de corrélation, mais à partir d'une étude valide sur un domaine différent. On teste alors l'hypothèse selon laquelle il n'y a pas de corrélation entre les deux résultats.

Dans le cas d'un domaine abstrait plus que dans tout autre, il semble qu'il n'existe aucune procédure qui puisse offrir seule la preuve définitive de la validité; elles souffrent toutes de limites pratiques et logiques.

D'un point de vue pratique, le problème est de déterminer le niveau de corrélation acceptable pour conclure à la validité d'une étude. GC Helmstadter propose de comparer la corrélation observée avec la corrélation maximale atteignable, qui est égale au produit des racines carrées des coefficients de fiabilité de chacune des deux mesures (GC Helmstadter, 1966). Le coefficient de fiabilité représente la part de variabilité de la mesure observée que l'on peut attribuer à la variance de la valeur réelle de l'objet. Aucune règle n'a cependant été établie dans ce domaine jusqu'à présent.

D'un point de vue logique, on retrouve la critique relative à la validité de critère, selon laquelle la référence doit elle-même être validée dans le cadre spécifique de l'étude. Dans ce cas précis, on ne peut même plus justifier la technique statistique de la corrélation par l'existence d'une base commune. En effet, les mesures sont appliquées dans un même domaine, mais il est fort peut probable que les construits soient identiques puisqu'ils ont été élaborés par des personnes différentes, dans des contextes différents. En fait, cette technique évalue la corrélation existant entre des mesures inexactes, de **concepts, ou construits similaires mais non identiques**. On peut donc s'interroger sur la signification réelle des corrélations observées.

3.2 : L'analyse Multitraits- Multiméthodes.

Cette technique statistique a été développée par Campbell et Fiske (Campbell et Fiske, 1959) pour évaluer les deux validités de convergence et de divergence de façon systématique et simultanée. Cette technique est fondée sur la matrice multitraits-multiméthodes qui comprend toutes les corrélations entre des concepts différents, mesurés avec des méthodes différentes. Cette méthode permet de révéler si la corrélation entre un même trait (facteur, dimension de qualité de vie...) est plus élevée que la corrélation entre deux traits différents lorsqu'ils sont mesurés avec les mêmes méthodes. Cette méthode n'a pour l'instant pas fait l'objet de beaucoup d'applications (D. Hadorn, 1991). Elle est très lourde a mettre en place car elle nécessite qu'au moins deux dimensions différentes soient mesurées simultanément par au moins deux méthodes différentes. Elle est bien sûr soumise aux critiques développées précédemment.

Face à ces critiques, nous rejoignons l'opinion de Mc Dowell et Newell : "La validation, en termes de validité de construit, d'une mesure de la santé est par certains aspects une science, mais elle est en grande partie une forme d'art" (Mc Dowell et Newell, 1987).

Nous venons de le voir, si le concept de validité est bien défini, les méthodes qui ont été proposées pour le tester ne sont elles-mêmes généralement pas valides, même si elles reposent sur des instruments statistiques éprouvés. En conséquence, nous pensons qu'il est préférable, dans l'état actuel de la recherche, de justifier un questionnaire en termes de validité de contenu. La technique de consensus entre les différents experts a le mérite, en particulier, d'afficher explicitement sa subjectivité. Les personnes qui seront amenées à travailler avec les résultats de l'étude seront donc averties et conscientes des limites de leur interprétation ainsi que de leur utilisation. Les validations de construit et de critère nous semblent dangereuses car elles masquent des faiblesses certaines par des méthodes statistiques reconnues. Nous cautionnons donc tout à fait la prudence de G.

Guyatt lorsqu'il écrit que : " nous ne devrions peut-être jamais conclure qu'un questionnaire a été validé, mais plutôt suggérer qu'une forte présomption de validité a été observée dans un certain nombre d'échantillons et d'études différents" (G. Guyatt, D. Feeny, D. Patrick, 1993). En conclusion, nous reprendrons la définition de la validation donnée par Dickes : "**un processus hypothético-déductif de recherche continu qui s'appuie sur un faisceau convergent d'arguments et de preuves**" (Dickes et al, 1994).

Chapitre 3 : La fiabilité (reliability).

La validité est une condition nécessaire mais non suffisante de la justification d'un questionnaire. Après s'être assuré que l'on vise la bonne cible, on doit pouvoir l'atteindre de façon régulière. On parlera alors de fiabilité. **Lorsque l'erreur de mesure aléatoire est mince, la mesure est dite fiable.** Nous présentons ici le modèle du score vrai selon la théorie classique. Ce modèle, souvent présenté comme un modèle de test, est en fait un modèle de mesure. Il suffit de raisonner non plus sur le terme d'erreur de mesure mais sur celui de vrai score pour s'en convaincre.

§1 : Le coefficient de fiabilité.

1.1 : Le modèle général.

Pour prouver la fiabilité, nous devons calculer la part d'erreur aléatoire dans la mesure afin de démontrer qu'elle est minime. Pour cela, nous reprenons notre modèle, auquel nous avons soustrait l'erreur systématique, la mesure est donc valide. En posant que la variance de la mesure qui n'est pas expliquée par la valeur réelle, n'est due qu'à l'erreur aléatoire, nous pouvons déduire les hypothèses du modèle.

$$\boxed{Y = T + E}$$

avec

$E \quad N(0, \sigma^2 e)$

$T \quad N(\mu, \sigma^2 t)$

$\text{cov}(T, E) = 0.$

Sous ces hypothèses, la variance de la mesure s'écrit :

$$V(Y) = V(T+E)$$
$$= V(T) + V(E) + 2\,COV(T, E)$$
$$= V(T) + V(E)$$

La variabilité des résultats observés dans une série de mesures sur différents sujets peut s'expliquer par deux composantes : la différence réelle existante entre des individus différents et une variance aléatoire. La fiabilité est alors définie comme la proportion de variance observée attribuable aux différences entre les sujet.

On définit le coefficient de fiabilité de Y, mesure de T, comme le ratio de la variance réelle sur la variance observée (Carmines EG, Zeller RA., 1988).

$$\rho_y = \frac{V(T)}{V(Y)}$$

La variance due à la valeur réelle n'est par définition pas observable, puisque c'est elle que l'on cherche à estimer. On peut cependant calculer précisément le coefficient de fiabilité en faisant l'hypothèse de deux mesures parallèles.

1.2 : Le modèle des mesures parallèles.

En faisant l'hypothèses de mesures parallèles il est possible de calculer le coefficient de fiabilité. Deux formes différente d'un même instrument de mesure sont dites parallèles si elles produisent des distributions de scores observés identiques, bien qu'elles contiennent des items différents.

Le modèle.

Ce modèle est développé à partir de variables centrées sur les individus.

$y_1 = t + e_1$

$y_2 = t + e_2$

avec :

y_1 : score centré observé sur le test 1

y_2 : score centré observé sur le test 2

t : score centré réel

e_1 : erreur du test 1

e_2 : erreur du test 2

Hypothèses du modèle :

$\sigma\, y_1 = \sigma\, y_2$

$\bar{y}_1 = \bar{y}_2$

$\rho\,(y_1, t) = \rho\,(y_2, t)$

$\rho (e_1, t) = \rho (e2, t)$

$\rho (e_1, e_2) = 0$

$E(e_1) = E(e_2) = 0$

ρ : coefficient de corrélation.

σ : écart-type.

E : espérance mathématique.

Le calcul du coefficient de fiabilité repose sur le calcul de la corrélation entre les deux mesures parallèles. Cela se justifie par le fait qu'elles ont un élément commun qui est la valeur réelle de l'objet. En calculant la corrélation existant entre les deux mesures on va pouvoir extrapoler la part de l'élément qu'elles ont en commun.

Soit i, le nombre d'observations : i = 1,.., n.

$$\rho(y_1, y_2) = \frac{1/N \sum_{i=1}^{n} y_{i1} y_{i2}}{\sigma y_1 \sigma y_2}$$

Par hypothèse, $\sigma y_1 = \sigma y_2$

En remplaçant par le modèle, on obtient :

$$\rho(y_1, y_2) = \frac{1/N \sum_{i=1}^{n} (t_i + e_{i1})(t_i + e_{i2})}{\sigma^2 y_1}$$

$$\rho(y_1, y_2) = \frac{1/N \sum_{i=1}^{n} (t^2_i - te_{i1} - te_{i2} + e_{i1}e_{i2})}{\sigma^2 y_1}$$

Puisque les variables sont centrées :

$$\rho(y_1, y_2) = \frac{\sigma^2 t + \sigma te_1 + \sigma te_2 + \sigma e_1 e_2}{\sigma^2 y_1}$$

Sur deux mesures parallèles, les erreurs ne sont pas corrélées entre elles et ne sont pas corrélées avec les scores réels. Les trois termes de covariance s'annulent.

$$\rho(y_1, y_2) = \frac{\sigma^2 t}{\sigma^2 y_1} = \rho y_1$$

$$\rho(y_1, y_2) = \frac{\sigma^2 t}{\sigma^2 y_2} = \rho y_2$$

On a donc démontré que **sous l'hypothèse de deux mesures parallèles, le coefficient de fiabilité est égal à la corrélation entre les deux mesures**. On peut par ailleurs montrer que la racine carrée du coefficient de fiabilité est égale à la corrélation entre une mesure y parallèle et la vraie valeur t (Nunnaly J., 1978) :

$$\rho(y_1, t) = \sqrt{\rho(y_1, y_2)} = \sqrt{\rho y_1} = \sqrt{\rho y_2}$$

Les hypothèses qui fondent les mesures parallèles sont très restrictives, mais elles sont surtout invérifiables puisqu'elles sont basées sur des concepts inobservables : l'égalité des corrélations entre les différentes mesures et la vraie valeur (Bravo G., 1991). Les hypothèses de strict parallélisme ont été relâchées pour donner lieu à d'autres définitions.

Deux formes tau-équivalentes admettent des différences entre les variances d'erreurs, donc également entre les variances des scores observés.

Deux formes essentiellement tau-équivalentes admettent des différences entre les variances et les moyennes des scores observés.

Deux formes sont dites congénériques lorsqu'on considère que les scores vrais des deux versions sont parfaitement corrélés.

De ces différentes formes de mesure dépendra l'interprétation des divers coefficients de fiabilité que l'on peut calculer à partir du modèle plus général que l'on va voir maintenant et dont le modèle de mesure parallèles n'est qu'un cas particulier : le modèle de l'échantillonnage du domaine.

1.3 : Le modèle de l'échantillonnage du domaine (*"sampling-domain model"*).

Dans ce modèle, toute mesure particulière est considérée comme se composant d'un échantillon aléatoire d'items à l'intérieur du domaine hypothétique de l'ensemble des items (Nunnaly J., 1978). Le but de la mesure est alors d'estimer la valeur qui aurait été obtenue si tous les items du domaine avaient été employés. Implicitement, lorsque l'on parle de la mesure ou de la valeur dans ce modèle, on parle de la somme des scores des items. Cette remarque est importante pour déterminer le type de données que l'on peut traiter dans ce modèle.

On ne dispose bien entendu pas de l'ensemble des items disponibles et on va travailler avec un groupe d'items aléatoires ou supposés aléatoires.

Supposons dans un premier temps que l'on dispose de l'ensemble des items du domaine. On fait l'hypothèse de scores standardisés, de façon à ce que la corrélation moyenne de chaque item avec les autres items du domaine soit égale à la corrélation moyenne de l'ensemble des items (qui est en fait le coefficient de fiabilité). On peut alors calculer la corrélation de chaque item avec la somme de tout les items du domaine.

$$\rho(y_k, \sum_{\substack{i=1;n \\ i\neq k}} y_i) = \frac{1+n\overline{\rho}}{\sqrt{n+(n^2-n)\overline{\rho}}}$$

$\overline{\rho}$: corrélation moyenne de l' ensemble des items (coefficient de fiabilité).

n : nombre d' items

y_i : mesure de l' item i.

En faisant l'hypothèse que l'univers des items est infini, on a :

$$\rho(y_k, \sum_{\substack{i=1;n \\ i \neq k}} y_i) \xrightarrow[n \to +\infty]{} \sqrt{\overline{\rho}}$$

La corrélation d'un item avec le score réel de l'objet est égale à la racine carrée de la corrélation moyenne de l'ensemble des items, ou coefficient de fiabilité. Il est bien évident que nous ne connaissons pas cette corrélation moyenne, nous devons donc estimer le coefficient de fiabilité, à partir d'un échantillon aléatoire d'items appartenant au domaine. L'hypothèse de construction aléatoire est très importante car elle permet de supposer que les mesures obtenues tendent à être parallèles, on peut alors accepter l'hypothèse selon laquelle les scores sont standardisés. **On estime la corrélation moyenne de l'ensemble des items du domaine par la corrélation moyenne des items de l'échantillon.**

Ce modèle repose sur l'hypothèse d'un échantillonnage aléatoire des items parmi l'ensemble des items du domaine, ce qui permet d'**estimer** le coefficient de fiabilité par la corrélation moyenne des items de l'échantillon. Dans la pratique, cette hypothèse n'est que rarement respectée. En effet, comme nous l'avons expliqué dans notre chapitre sur la validité, les items sont sélectionnés *a-priori* pour leur représentativité de l'objet étudié. Les deux objectifs de validité et de fiabilité sont donc, d'un point de vue empirique, contradictoires. Cette difficulté est contournée en supposant un échantillonnage aléatoire des items, ce qui permet de présenter les méthodes mises au point sur la base des deux modèles de mesures parallèles et d'échantillonnage aléatoire.

§2 : Les différentes méthodes d'estimation du coefficient de fiabilité.

2.1 : La méthode test-retest.

Lorsque l'on met sur pied une étude chronologique, cette méthode détecte les erreurs de mesures consécutives à l'évolution dans le temps. Pratiquement, deux mesures (y_1 et y_2) d'une variable fixe, faites à deux instants différents, sur une même personne, doivent donner les mêmes résultats. La valeur de l'objet étant invariable, on a :

$Y_1 = T + E_1$

$Y_2 = T + E_2$

$Y_1 = Y_2 \Rightarrow E_1 = E_2$

donc : $\sigma^2 e_1 = \sigma^2 e_2$

$\sigma^2 y_1 = \sigma^2 y_2$

alors : COV (E1,T) = COV(E2,T) = 0

On retrouve l'hypothèse fondatrice du modèle des mesures parallèles. On peut alors calculer directement le coefficient de fiabilité, puisque **d'après les conclusions**

du modèle, le coefficient de fiabilité est égal à la corrélation entre les deux mesures parallèles.

Cette méthode est avantageuse puisqu'elle permet de donner le calcul exact du coefficient et non une estimation. Cependant, elle fait l'objet de vives critiques.
D'un point de vue pratique tout d'abord, il n'est pas toujours possible de réaliser deux mesures consécutives du même objet sur une même personne. Mais la critique la plus problématique est celle qui concerne le laps de temps qu'il faut laisser entre les deux mesures, pour être sûr que la valeur T est restée inchangée (condition de deux mesures parallèles). En effet, d'une part la variable peut avoir réellement évolué, mais d'autre part la façon dont le sujet la perçoit peut également avoir changé. Le simple fait de réaliser une étude, de poser certaines questions peut amener les sujets à modifier leur point de vue, ce que l'on appelle phénomène de réaction à l'enquête. Cette modification peut également être la conséquence d'un événement extérieur ignoré par l'observateur. Cela peut avoir pour effet de relativiser les réponses précédentes, ou au contraire de mettre un lumière des problèmes dont le sujet n'avait pas conscience.

La réduction du laps de temps, entre les deux interviews, a été proposée pour limiter ces modifications qui ne relèvent pas de l'erreur, mais sont bien réelles. Cependant, dans quelle mesure faut-il réduire ce laps de temps? Si le délai est trop court entre les deux passages, le sujet peut se rappeler de ses réponses précédentes et biaiser ainsi les résultats.

Nous avons choisi de ne pas appliquer cette méthode d'une part car son application est très lourde, et d'autre part car il nous semble illusoir d'espérer recueillir deux mesures parallèles dans le cadre d'une étude sur la perception d'une pathologie chronique telle que les AVC.

2.2 : La méthode Split-halves.

Cette méthode a été développé afin de pouvoir appliquer le modèle des mesures parallèles, lorsque l'on ne dispose que d'une mesure. L'estimation se fait à partir d'un partitionnement des items en deux parties, chacune des deux moitiés étant alors considérées comme deux formes alternatives d'un même test (mesures parallèles y_1 et y_2). On peut alors calculer la corrélation au carré du score observé y et du score réel t :

$$\rho^2(y,t) = \frac{2\rho(y_1,y_2)}{1+\rho(y_1,y_2)}$$

Cette formule est le cas particulier, où l'on a deux mesures de la formule de Spearman-Brown (Spearman, 1910). La forme générale de la formule est :

$$\rho^2(y,t) = \frac{k\overline{\rho}(y_i, y_j)}{1 + (k-1)\overline{\rho}(y_i, y_j)}$$

$\overline{\rho}$: corrélation moyenne

k : nombre de formes alternatives parallèles comparées.

Cette formule n'est valable que pour des mesures parallèles. Cette hypothèse justifie que la méthode Split-halves repose uniquement sur la corrélation entre deux formes alternatives et non sur l'ensemble des corrélations entre tous les partitionnements possibles. La corrélation calculée est égale à toutes les corrélations possibles de l'univers des items. Cette méthode pose la question de la répartition des items qui fournira deux moitiés parallèles, sur laquelle les avis sont partagés : Nunnaly recommande une répartition aléatoire des items (Nunnaly J., 1978) ; Dickes préfère une assignation des items en deux moitiés qui respectent les contenus et l'ordre de présentation des items dans l'instrument global (Dickes et al, 1994). Par ailleurs, nous avons déjà précisé que le parallélisme de deux mesures n'est pas démontrable, d'ailleurs bien souvent dans la pratique, la procédure Split-halves donne des résultats différents selon la façon dont les items sont regroupés : cela tend à démontrer que les mesures ne sont pas parallèles (Cronbach L.J., 1951). En conséquence, la formule a été travaillée afin d'être applicable à deux moitiés tau-équivalentes (Rulon P.J., 1939).

$$\rho^2(y,t) = 2\left(1 - \frac{\sigma^2 y_1 + \sigma^2 y_2}{\sigma^2 y}\right)$$

Les deux hypothèses de mesures parallèles et tau-équivalentes sont trop restrictives pour que la méthode Split-halves ait une réelle applicabilité. La formule Spearman-Brown a été simplifiée et traduite en termes de variance et non plus de corrélation, afin d'abandonner l'hypothèse de mesures parallèles et de rester dans le cadre plus large du modèle d'échantillonnage du domaine.

2.3 : Le coefficient alpha de Cronbach.

23.1 : Le développement de la formule.

En faisant l'hypothèse que la corrélation moyenne d'une collection aléatoire (on tend vers des mesures parallèles) d'items est une bonne estimation de la corrélation moyenne hypothétique du domaine, on peut reprendre la formule de Spearman-Brown :
L'échantillon est composé de k items indicés i et j.

$$\rho^2(y,t) = \frac{k\overline{\rho}(y_i, y_j)}{1 + (k-1)\overline{\rho}(y_i, y_j)}$$

$$\rho^2(y,t) = \frac{k^2\overline{\rho}(y_i, y_j)}{k + (k^2 - k)\overline{\rho}(y_i, y_j)}$$

Le dénominateur est estimé par la somme de tous les éléments de la matrice des corrélations. Le numérateur est estimé par cette même somme, moins les éléments de la diagonale, ce qui nécessite une pondération particulière.

$$\rho^2(y,t) = \frac{(k / k-1)\sum_{i \neq j} \rho(y_i, y_j)}{\sum_{i=1}^{k}\sum_{j=1}^{k} \rho(y_i, y_j)}$$

y_i et y_j représentent les scores des k items du questionnaires. La démonstration complète est décrite dans l'ouvrage de J. Nunnaly (Nunnaly J., 1978).

Traduite en termes de variances et de covariances afin d'être applicable à des données brutes et non plus standardisées, cette formule simplifiée s'écrit :

$$\rho^2(y,t) = \frac{k}{k-1}\left(\frac{\sum_{i=1}^{k}\sum_{j=1}^{k} \mathrm{cov}(y_i, y_j) - \sum_{i=1}^{k} \sigma^2 y_i}{\sum_{i=1}^{k}\sum_{j=1}^{k} \mathrm{cov}(y_i, y_j)} \right)$$

La variance de la somme des items est égale à la somme des éléments de la matrice des variances/covariances (Nunnaly J., 1978).

Soit $y = \sum_{i=1}^{k} y_i$

$\sigma^2 y = \sum_{i=1}^{k}\sum_{j=1}^{k} \mathrm{cov}(y_i, y_j)$ d'où :

$$\boxed{\rho^2(y,t) = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma^2 y_i}{\sigma^2 y} \right)}$$

Ce résultat est appelé **coefficient alpha de Cronbach** (Cronbach L.J., 1951), il est noté : α. En fait, la formule de Spearman-Brown n'est qu'un cas particulier du coefficient alpha, dans le cas d'items parallèles.

Lorsque le questionnaire est composé d'items dichotomiques, la formule s'écrit en termes de proportion, afin de donner le coefficient KR20 (Kuder G., Richardson M., 1937) :

$$KR20 = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} p_i q_i}{\sigma^2 y}\right)$$

23.2 : Interprétations.

Nous allons voir maintenant que le coefficient alpha peut être interprété de plusieurs façons

Interprétation n°1 : Le coefficient alpha est une estimation du coefficient de fiabilité du questionnaire.

La formule développée plus haut nous donne :

$\sqrt{\alpha} = \rho(y,t)$

En fait, cette égalité n'est vérifiée que dans le cas d'items parallèles ou essentiellement tau-équivalentes. Dans le cas de mesures congénériques (Kristof, 1983), le coefficient alpha donne une sous-estimation du coefficient de fiabilité : "α gives a lower bound to the true reliability" (Cronbach L.J., 1951)[1].

Interprétation n°2 : Le coefficient alpha est égal à la moyenne de tous les coefficients Split-halves possibles pour un questionnaire donné.

Le coefficient alpha est égal à la corrélation estimée d'un test, avec toutes les formes alternatives à ce test, qui comportent le même nombre d'items. La technique du coefficient alpha apparaît ici comme une bonne alternative de la méthode Split-halves lorsqu'on ne dispose pas de mesures parallèles mais tau-équivalentes, puisqu'elle lève l'une des principales critiques, à savoir qu'une corrélation unique ne pouvait pas, dans ce cas, représenter tout un ensemble de corrélations[2].

Interprétation n°3 : Le coefficient alpha est une mesure de la cohérence interne du questionnaire.

Le terme de cohérence interne désigne l'homogénéité des items dans le questionnaire. Cette interprétation se fonde sur l'utilisation dans le développement de la formule alpha de la corrélation entre les différents items du questionnaire. Un fort coefficient peut alors être interprété comme une bonne cohérence globale des items à l'intérieur du questionnaire : les items concernent tous le même trait. Le calcul du coefficient alpha

---

[1]Novick a démontré ce résultat en 1967 (Novick M., Lewis C., 1967, p.5).

[2] La démonstration de ce résultat se trouve page 10 de l'article de Novick et Lewis (Novick M., Lewis C., 1967).

trouve sa justification dans l'hypothèse implicite que la cohérence interne d'une échelle unidimensionnelle psychométrique doit être maximisée (Ray, 1988)
Mais une bonne cohésion générale ne signifie pas que tous les items, pris séparément, sont cohérents avec les autres. Il est alors intéressant de connaître la conséquence individuelle de chacun des items sur la cohérence du groupe. La formule suivante permet de calculer le coefficient alpha lorsque l'on supprime l'un des items (SAS, 1990).

$$\alpha_k = \left(\frac{k-1}{k-2}\right)\left(1 - \frac{\sum_{i \neq k} \sigma^2 y_i}{\sigma^2\left(\sum_{i \neq k} y_i\right)}\right)$$

Si $\alpha_k < \alpha$, cela signifie que si l'on retire l'item k, la cohérence interne du questionnaire diminue. L'item k doit être gardé. Inversement, il pénalise la cohérence de la mesure si $\alpha_k > \alpha$. Il est cependant difficile d'interpréter ces résultats car il n'existe aucun test de significativité de la différence entre α_k et α, il convient donc de rester prudent. De plus, faut-il retirer un item lorsqu'il n'est pas cohérent avec les autres? En effet, lorsque le questionnaire est appliqué à un concept multidimensionnel, il est normal que les items aient une faible cohérence entre eux, puisqu'ils mesurent des dimensions différentes d'un même construit.

23.3 : Les limites.

Les différentes méthodes d'évaluation du coefficient de fiabilité, bien qu'elles aient un visage plus "scientifique" que les méthodes de validation, sont l'objet de différentes critiques, aussi bien pratiques que logiques.

Le coefficient alpha est fonction du nombre d'items. Ce nombre doit donc entrer dans l'interprétation des résultats. Lorsque l'on accroît le nombre des items, on augmente alpha. Un mauvais chiffre peut indiquer indifféremment un questionnaire trop court ou une mauvaise cohérence interne. Symétriquement, "l'alpha peut être élevé, même s'il n'y a pas de facteur commun, puisque (1) il est influencé par le nombre des items et des répétitions parallèles d'items, (2) il s'accroît quand le nombre de facteurs relevant de chaque item s'accroît, et (3) il décroît modérément quand les "communalities" des items s'accroît" (Hattie, 1985).

L'interprétation qui soulève le plus de difficultés concerne la cohérence interne, qui ne serait pas un bon indicateur de fiabilité (Hattie, 1985). Pratiquement, on recommande un alpha supérieur à 0,8 pour conclure en faveur du questionnaire testé. Mais il n'existe en fait aucune règle de décision incontestée. Des contradictions entre les différentes interprétations possibles du coefficient alpha peuvent apparaître. En effet, un fort coefficient (supérieur à 0,9) peut être interprété positivement comme la preuve d'une faible part d'erreur de mesure (interprétation n°1). Mais il peut également être le signe

d'une redondance des items (interprétation n°3), la conclusion est alors négative pour le questionnaire puisqu'il "suggère que le questionnaire est trop étroit et trop spécifique...si on construit des items qui sont virtuellement des paraphrases les uns des autres, les résultats seront d'une cohérence interne élevée et d'une validité très faible" (Kline, 1979). Ce problème d'interprétation en termes de cohérence interne met en opposition les deux concepts de validité et de fiabilité : "une forte cohérence interne peut être (...) contraire à une forte validité (...) l'importance de la fiabilité en termes de cohérence interne a été exagérée par la psychométrie" (Kline, 1986).

Enfin, l'interprétation de cohérence interne, comme nous l'avons déjà évoqué précédemment, révèle l'hypothèse implicite d'un concept à mesurer unidimensionnel, ce qui n'est certainement pas le cas de la qualité de vie. Dans le cas d'un domaine vaste, le concept d'homogénéité des items perd tout son sens : "Dans tous les cas, et spécialement dans les domaines de la motivation, de la personnalité et de l'humeur, une homogénéité des items modérée à faible est réellement préférable si on veut s'assurer d'une couverture large des construits mesurés" (Boyle, 1991).

Enonçons maintenant une limite plus problématique car elle concerne une hypothèse du modèle général : on suppose qu'il est développé sur la base d'erreurs aléatoires. Cela signifie qu'il ne peut s'appliquer que sur une étude validée. En d'autres mots, ce modèle ne s'applique qu'en l'absence d'erreurs systématiques ce qui, nous l'avons souligné, est loin d'être facile à prouver dans la pratique. La théorie de la généralisabilité a été développée pour répondre à cette critique.

§3 : La théorie de la généralisabilité

La théorie classique du score réel que nous venons de développer largement repose sur une hypothèse simple : Tout score observé peut être décomposé en un score réel (*"true score"*) et en un score d'erreur. Le terme d'erreur est supposé aléatoire et indifférencié. Cette hypothèse est à la base du modèle :

$$Y = T + E.$$

A partir de ce modèle, on détermine un coefficient de fiabilité égal au rapport de la variance du 'vrai' score et de la variance du score observé.

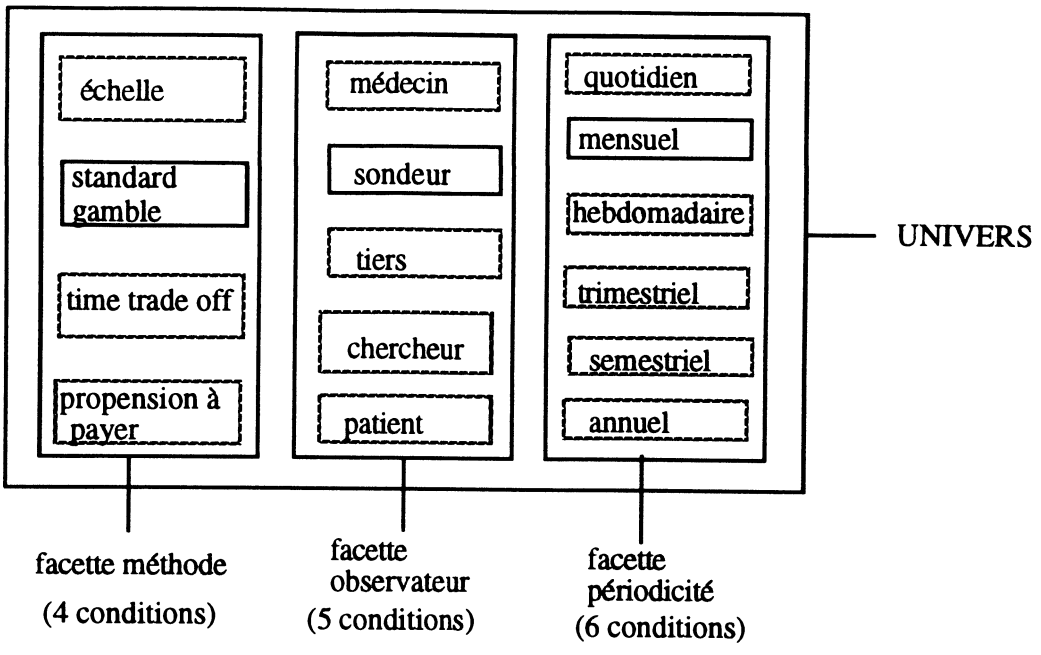
$$\boxed{\rho_y = \frac{V(T)}{V(Y)}}$$

Cette hypothèse est particulièrement simpliste, puisqu'elle ne s'applique que dans le cas d'une mesure valide, or nous avons exposé précédemment toutes les difficultés que l'on rencontre pour démontrer la validité d'une mesure. La première hypothèse du modèle classique du score vrai en fait donc un modèle très limité qui ne tient pas compte des biais systématiques, plus : il les ignore.

La théorie de la généralisabité développée par Cronbach et son équipe dans les années soixante (Cronbach et al, 1963, 1965) admet toutes les sources d'erreur sans distinction. L'essence de la théorie est de reconnaître que dans toutes situations de mesure il existe de multiples, en réalité infinie, sources de variance. Ils sont partis de la réflexion selon laquelle "la plupart des procédures de mesure comportent un ou plusieurs aspects variables qui ne sont pas spécifiés dans la définition opérationnelle de la procédure" (Gleser et Cronbach, 1965). Cela recouvre le jour de passage de l'étude (début ou fin de semaine), le moment dans la journée, la qualité de l'observateur... Or il y a de fortes présomptions pour que ces aspects, qui peuvent sembler anecdotiques, aient une influence sur la réponse du sujet et donc sur les résultats de l'étude. Le but initial de la théorie de la généralité est de s'assurer que les résultats particuliers de l'étude sont assez généraux pour rester valables dans d'autres circonstances.

On suppose que toute mesure est " un échantillon appartenant à l'ensemble de toutes les mesures qui auraient pu être faites" (Evans, 1981). Ce modèle repose sur l'hypothèse que toutes les observations sont aléatoirement et indépendamment échantillonnées. Comme dans le cas du modèle d'échantillonnage du domaine, cette hypothèse est plus admise que démontrée dans la pratique. Le score obtenu lors de cette mesure n'a d'intérêt que dans la mesure où il est représentatif de l'ensemble des observations possibles. Chaque observation est déterminée par un certain nombre de conditions. Il nous semble nécessaire ici de faire une revue du vocabulaire employé dans cette théorie.

La population désigne l'ensemble des objets de l'étude. On définit chaque procédure de mesure en décrivant son mode opérationnel : un observateur, un instrument de mesure... Chacune de ces précisions est appelée facette de la mesure et constitue une source de variance. Ces facettes sont à leur tour définies par les conditions de mesure qui constituent ensemble l'univers des observations possibles. L'univers est défini de façon pratique et théorique par l'investigateur. Pour plus de clarté, illustrons ces définitions d'un exemple. Nous supposerons une mesure réalisée par un sondeur professionnel, tous les mois, à l'aide de la méthode du standard gamble. Nous avons représenté l'univers à partir duquel est échantillonnée cette observation. Les cadres en trait plein indiquent la condition choisie au sein de chaque facette pour la mesure particulière.



"Chaque observation doit être par conséquent entendue comme un échantillon de l'univers des observations possibles" (Gleser et Cronbach, 1965). En théorie, on peut considérer que "le score moyen du sujet sur toutes ces observations potentielles est le score de l'univers" (Gleser et Cronbach, 1965). On peut dès à présent définir le coefficient de généralité comme le ratio de la variance du score de l'univers sur la variance du score observé. Il est l'équivalent d'un coefficient de fiabilité non plus défini sur un univers monofacette mais sur un univers multifacettes. Il est bien sûr impossible de calculer le score de l'univers, qui devra donc être inféré à partir des observations réellement réalisées.

Nous présenterons maintenant cette méthode d'un point de vue pratique à partir d'un exemple développé par Evans (Evans, 1981). L'univers est composé de trois facettes : l'observateur, le positionnement dans le temps des observations et les sujets. La facette observateur comporte 3 conditions, la facette temps comporte 5 conditions et la facette sujet comporte 15 conditions. Chaque sujet sera interrogé par des observateurs différents, à des moments différents.

Dans un premier temps, on réalise une G-Study (*Generabilizability study*). Il s'agit d'identifier les principales sources de variance potentielles puis de collecter des données pour estimer les composantes de variance du score observé correspondant aux facettes variées du processus de mesure. Cette décomposition de la variance du score se fait par une analyse de type ANOVA.

Dans un deuxième temps, on procède à une D-Study (*Decision Study*) qui correspond à la phase d'interprétation des résultats de l'étude précédente en vue d'améliorer le processus de mesure. On cherche à dégager l'impact, en termes de fiabilité, d'une décision tendant à modifier la procédure opérationnelle de la mesure. Dans ce but, on

génère un ensemble de coefficients de généralité qui dépendent des conditions particulières de l'observation. Ces coefficients sont calculés en faisant varier certaines facettes, d'autres restant fixes. Dans notre exemple, il y a 12 possibilités de mesures différentes et donc 12 coefficients de généralité. Chaque coefficient indique dans quelle mesure on peut généraliser à l'ensemble d'une facette des résultats obtenus sur la base d'une condition particulière. Evans considère qu'un coefficient de 0,8 est un minimum acceptable pour conclure à la fiabilité d'une facette.

Etudions par exemple les coefficients obtenus par Evans sur l'ensemble des 15 sujets.

Facette variable	**Facette fixe**	**Coefficient de généralité**
observateur	temps	0,997
temps	observateur	0,820
observateur+temps	-	0,817

Source : Evans, 1981.

Le coefficient de 0,817, obtenu en faisant varier les deux facettes indique que l'on peut mesurer l'état de santé du patient, indépendamment de l'observateur choisi ou de la date du test. Le résultat obtenu pour chaque patient est généralisable à l'ensemble de l'univers.

Si on se limite à nos trois observateurs, on ne modifie que très peu le coefficient de généralité (0,820), donc le choix de l'observateur n'influe pas sur la variance d'erreur.

Par contre, le choix des points dans le temps accroît fortement le coefficient (0,997), cela signifie que le choix de ces dates de passage sont déterminantes pour la fiabilité de l'étude et qu'elles sont dans ce cas particulièrement adéquates.

Pour résumer, l'approche générale de la théorie de la généralité consiste dans un premier temps à isoler les sources variées de variance dans les scores et dans un deuxième temps à calculer l'impact du choix des conditions d'observation sur la fiabilité de la mesure. Cette analyse multifacette de l'erreur comporte des avantages par rapport au modèle classique du score vrai.

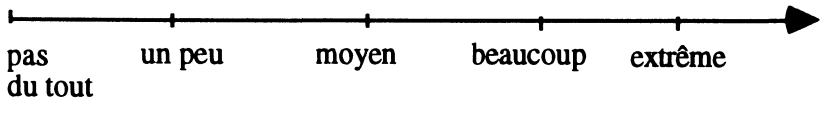
Cette technique prend en considération de façon explicite les différentes facettes d'une opération de mesure et révèle ainsi des ambiguïtés ignorées par le modèle classique. Par exemple elle révélera le fait qu'une mesure puisse être fiable quelque soit l'observateur choisi, mais ne plus l'être si on modifie la date de l'interview.

Cette technique permet de rendre compte des interactions existant entre des facettes différentes, interactions qui sont inaccessibles par le modèle classique. On aura ainsi une meilleure compréhension du processus de mesure.

Lorsque l'on travaille avec le modèle classique, on est obligé de réaliser plusieurs études pour présenter la fiabilité dans le temps (test-retest), la fiabilité inter-observateur etc...Ici, l'analyse de généralisabilité résume en une seule étude de fiabilité l'ensemble des variances d'erreur.
Enfin, elle possède une étape d'interprétation du processus de mesure. Elle apporte à l'investigateur les informations nécessaires à la mise au point de procédures de mesure plus efficaces.

Pour résumer les objectifs de cette théorie, nous reprendrons une phrase de Evans : " **Comprendre où naît l'erreur et la corriger**" (Evans, 1981).

Cependant, une limite du modèle ne peut être résolue par la théorie de la généralité puisque cette limite s'applique aussi bien au modèle classique qu'au modèle général.
L'avantage majeur du modèle développé dans les différentes approches de la fiabilité est sa simplicité, puisque nous avons affaire à un modèle linéaire. Mais comme souvent, cette simplicité se transforme vite en limite. La linéarité du modèle d'échantillonnage du domaine le restreint à l'étude de données cardinales. En conséquence, toutes les méthodes qui en découlent intègrent l'addition des scores des items en un score global afin d'évaluer la corrélation entre le score observé et le score réel.
Lorsque l'on veut appliquer ces techniques à des échelles ordinales (dont nous avons reproduit un exemple ci-dessous), on est obligé de faire deux hypothèses fortes. En effet, il est impossible d'appliquer l'opération mathématique de l'addition à des données qui ne reflètent que la notion d'ordre.



Afin de pouvoir tester la fiabilité de telles échelles à partir des méthodes présentées, il est nécessaire de supposer qu'elles ont une structure cardinale. Les deux hypothèses qui vont permettre de passer de la notion d'ordre à la notion de ratio sont les suivantes :
- tous les individus ont une origine commune,
- les écarts entre les différents paliers sont identiques à tous les individus.

Si l'intérêt pratique de ces hypothèses est évident, leur applicabilité réelle est plus critiquable. Ces deux hypothèses signifient que tous les termes ponctuant l'échelle ont la même signification pour tous. On peut parfaitement imaginer, par exemple, que "pas du tout" puisse être entendu comme une absence totale de symptômes (pour des personnes bien portantes), ou une absence totale de symptômes significatifs (pour des personnes ayant déjà une mauvaise santé). Si cette difficulté peut être amoindrie par une description très pointilleuse du terme "pas du tout" lors du déroulement de l'étude, l'hypothèse d'une même représentation des écarts est plus problématique. En effet, ce que recouvre la

distance séparant deux points sur cette échelle est d'ordre particulièrement subjectif. Supposer que l'écart entre "un peu" et "beaucoup" est invariant, quelle que soit la personne interrogée, ne nous semble pas recouvrir une réalité quelconque. En conséquence, nous pensons qu'il est préférable d'appliquer aux données ordinales d'autres modèles de mesure qui leur sont appropriés en insistant fortement sur l'étude de validation, puisqu'aucune étude de fiabilité n'a été développée pour ces données.

Notre but ici n'était pas de prôner l'abandon du modèle de mesure classique mais d'insister sur ses limites. Les méthodes développées sur la base de ce modèle sont de portée différente. Comme J. Nunnaly (Nunnaly J., 1978), nous pensons qu'il est préférable de ne pas utiliser les méthodes qui reposent sur le modèle des mesures parallèles (méthode test-retest, méthode split-halves), sauf si on peut prouver de façon indiscutable que leurs conditions d'application sont respectées. Il est aujourd'hui impossible de publier une étude sans test préalable de ses propriétés psychométriques, et le coefficient alpha de Cronbach est la technique la plus employée. Nous l'expliquons par le fait que les critiques que nous avons développées à son encontre, concernent moins sa construction ou le modèle sur lequel il repose que les interprétations qui en sont faites. En rejetant le test de Cronbach, Mc Donald (1981) jette le bébé avec l'eau du bain. Ce n'est pas l'alpha qui est à remettre en cause mais son interprétation systématique en termes de cohérence interne des items. Dans l'état actuel des recherches, le coefficient alpha est certainement le plus acceptable.

Dans la prochaine partie, nous vous présentons une application des tests de validation et de fiabilité que nous avons réalisé dans le cadre d'une étude d'évaluation de la qualité de vie auprès de personnes ayant subi un Accident Vasculaire Cérébral (AVC). Suite à notre réflexion théorique, nous avons privilégié pour valider notre questionnaire la méthode du consensus d'experts et pour tester sa fiabilité, la technique de Cronbach, tout en rappelant que l'interprétation qui en est faite est hautement subjective.

PARTIE 2 : APPLICATION - L'EVALUATION DES QUESTIONNAIRES AVC.

L'accident vasculaire cérébral est une complication aiguë liée à la diminution du débit sanguin cérébral. Cette pathologie constitue, d'après la plupart des travaux, la troisième cause de décès et la principale cause d'invalidité dans les pays industrialisés. Afin d'évaluer la qualité de vie des personnes ayant subi un AVC, nous avons élaboré un questionnaire comportant des questions de type dichotomique[3]. Cette partie présente les résultats de l'étude de validité et de fiabilité du questionnaire, réalisée courant 1995.

Chapitre 1 : La validité.

Nous avons expliqué précédemment que les méthodes de validation de construit, qui devraient s'appliquer dans le cas de la qualité de vie, ne nous semblaient pas préférables à une simple validation de contenu, mais au contraire qu'elles nous semblaient plus dangereuses : d'une part, parce qu'elles établissent la validité d'une étude à partir d'autres études dont la validité est contestable ; et d'autre part, parce que l'utilisation de la technique statistique de la corrélation cache une interprétation tout aussi subjective sous des dehors "scientifiques". Nous avons donc pris le parti de justifier le choix de nos items, premièrement par une étude bibliographique poussée, deuxièmement par des avis d'experts (professionnels de la santé et patients).
La recherche bibliographique nous a permis de pointer toutes les conséquences physiques (symptômes), psychiques et sociales consécutives à une attaque vasculaire cérébrale. Au total, nous avions relevé une quarantaine de sujets de plaintes. Nous les avons présentés à Monsieur Giroud, Professeur en neurologie à l'hôpital général de Dijon, avec lequel nous avons finalement sélectionné 33 items. Ces items ont été présentés dans un premier questionnaire à un groupe de 26 patients. Leurs réponses nous ont permis de réduire le nombre d'items à 28.

Au cours d'une deuxième étude, nous avons procédé à des tests de significativité pour chacun des items, en demandant à 44 patients de répondre au questionnaire dichotomique. Nous avons proposé le même questionnaire à leur médecin. Les résultats sont présentés dans le tableau 1.

Test de significativité de la proportion.
On suppose que la taille de l'échantillon est suffisante pour que la proportion d'échantillonnage, notée p_n, soit distribuée selon une loi normale.

[3] Nous ne pouvons pas publier ce questionnaire pour cause de confidentialité.

On se propose de tester l'hypothèse H0, selon laquelle la proportion de réponses positives à un item est significative :

$$\begin{cases} H_0 : p = 0 \\ H_1 : p \neq 0 \end{cases}$$

Pour cela, on calcule pour chaque item une fonction appelée fonction discriminante sous H0 :

$$z = \frac{pn}{\sqrt{\frac{pn(1 - pn)}{n}}}$$

L'hypothèse H0 sera acceptée pour $z \in [-1,96 ; 1,96]$, pour un seuil de significativité de 5%.

Tableau 1 : Test de significativité des différents items.

variables	obs. médecin	obs. patient	Fréquence médecin	Fréquence patient	significativité médecin	significativ patient
Handicap sup.	30	31	69,77	70,45	9,961	10,243
Handicap inf.	25	28	58,14	63,63	7,728	8,775
Tactile	11	14	25,58	31,81	3,844	4,531
Aphasie	23	24	53,49	54,54	7,0320	7,266
Vision	4	5	9,30	11,36	2,10	2,375
Incontinence	13	15	30,23	34,09	4,317	4,770
Vertige	5	6	11,628	13,63	2,379	2,636
Douleur	12	14	27,907	31,82	4,08	4,531
Fatigue	26	31	60,46	70,45	8,11	10,243
Poids	1	1	2,32	2,27	**1,012**	**1,011**
Déprime	27	27	62,79	61,36	8,52	8,359
Nervosité	22	24	51,16	54,54	6,712	7,266
Mémoire	30	31	69,76	70,45	9,9612	10,243
Concentration	31	28	72,09	63,63	10,54	8,775
Sommeil	28	32	65,11	72,73	8,96	10,832
Maux de tête	4	2	9,30	4,54	2,10	**1,447**
Toilette	11	13	25,58	29,54	3,844	4,295
Habillage	17	21	39,53	47,72	5,302	6,338
Nourriture	16	18	37,21	40,91	5,048	5,519
WC	18	24	41,86	54,54	5,564	7,266
Mobilité lit	12	16	27,90	36,36	4,08	5,014
Déplacement	22	22	51,163	50	6,712	6,633
Activités	31	38	72,09	86,363	10,54	16,693
Environnement	21	25	48,83	56,82	6,406	7,608
Logement	8	14	18,60	31,82	3,135	4,531
Conjoint	15	19	34,88	43,18	4,8	5,782
Famille	18	26	41,86	59,09	5,56	7,972
Amis	11	17	25,58	38,63	3,844	5,263

Pour $\alpha = 5\%$, on accepte l'hypothèse selon laquelle l'item est non significatif, si la valeur de la fonction discriminante est comprise dans l'intervalle [-1,96; 1,96].

Les résultats montrent la non significativité de l'item "poids" pour les patients et les médecins. L'item "maux de tête" n'est pas significatif pour les patients, mais il le reste

du point de vue des médecins. Tout les autres items sont significatifs. Il semble que seul l'item "poids" ne doive pas être intégré dans le score final.

Chapitre 2 : La fiabilité.

L'évaluation de la fiabilité de l'étude est plus complexe. En fait, à partir des items sélectionnés au cours de l'étude de validité, nous avons construit deux types de questionnaires. Un premier questionnaire est composé de l'ensemble des 28 items, avec un système de réponses dichotomiques. Ce questionnaire est proposé aux médecins et aux patients. Les mêmes items ont été regroupés par dimensions dans un second questionnaire, uniquement présenté aux patients, et avec un mode de réponse de type échelle. Les deux questionnaires réclament un traitement différent. Pratiquement, les différents tests reposent sur la population des dijonnais ayant subit un AVC entre le 01.01.1995 et le 15.04.1995, ce qui représente 44 personnes.

Le coefficient alpha de Cronbach donne les mêmes résultats pour des items dichotomiques que le coefficient KR20 (Streiner D.L., Norman G.R.,1989). Nous avons donc calculé les coefficients alpha pour le questionnaire médecin et le questionnaire patient à partir du programme SAS. Nous en présentons ci-dessous les résultats .

§1 : Questionnaire médecin.

Le coefficient alpha calculé pour la globalité du questionnaire est égal à :

$\alpha = 0{,}9183304.$

Ce coefficient est supérieur à 0.9, il correspond aux critères d'acceptation d'un questionnaire. Le tableau 2 donne le coefficient alpha lorsque l'on retire un item.

Tableau 2 : Coefficient alpha partiel.

Variables	Corrélation avec le total	Alpha de Cronbach
Handicap supérieur	0,446639	0,916593
Handicap inférieur	0,730448	0,91172
Sensation tactile	0,022621	**0,92284**
Déplacement	0,721298	0,911854
Aphasie	0,308939	**0,91911**
Vision	0,308939	**0,9195**
Incontinence	0,630188	0,91361
Vertiges	0,203373	**0,9193**
Douleur	0,366012	0,917823
Fatigue	0,585067	0,914309
Poids	0,057127	**0,9197**
Déprime	0,680802	0,912654
Nervosité	0,713884	0,911989
Mémoire	0,539756	0,915088
Concentration	0,576434	0,914512
Sommeil	0,597284	0,914113
Maux de tête	0,344652	0,917744
Toilette	0,628055	0,913739
Habillage	0,689542	0,912476
Nourriture	0,756137	0,911334
WC	0,678734	0,912649
Mobilité lit	0,619823	0,913818
Activités	0,378541	0,917628
Environnement	0,5922789	0,914175
Logement	0,207719	**0,91971**
Conjoint	0,580145	0,914404
Famille	0,509264	0,915631
Amis	0,595005	0,914254

Les items dont le retrait augmente le coefficient global sont : les sensations tactiles (0,922842), l'aphasie (0,919112), la vision (0,919499), les vertiges (0,9193), le poids (0,919701) et le logement (0,919708). Cependant, bien qu'il n'existe pas de règle de

rejet d'un item à partir du coefficient partiel, on peut remarquer que les écarts sont de l'ordre du millième. Nous ne modifierons pas le questionnaire sur ces bases.

§2 : Questionnaire patient.

Le coefficient alpha calculé pour la globalité du questionnaire est égal à :

$\alpha = 0{,}922837.$

Ce coefficient est supérieur à 0.9, comme précédemment il correspond aux critères d'acceptation d'un questionnaire. Le tableau 3 donne le coefficient alpha partiel pour chaque item.

Tableau 3 : Coefficient alpha partiel.

VARIABLES	*Corrélation avec le total*	*Alpha de Cronbach*
Handicap supérieur	0,509546	0,920437
Handicap inférieur	0,50852	0,920484
Sensation tactile	0,036559	**0,92747**
Déplacement	0,775855	0,916052
Aphasie	0,374601	0,922691
Vision	0,326145	0,922618
Incontinence	0,584502	0,919283
Vertige	0,154529	**0,92455**
Douleur	0,350347	0,922859
Fatigue	0,52438	0,920214
Poids	0,131001	**0,923882**
Déprime	0,718978	0,91708
Nervosité	0,675914	0,91775
Mémoire	0,613984	0,91886
Concentration	0,745072	0,916685
Sommeil	0,52626	0,920187
Maux de tête	0,142726	**0,92394**
Toilette	0,724009	0,917177
Habillage	0,733517	0,916774
Nourriture	0,751779	0,91651
WC	0,718016	0,917043
Mobilité	0,675201	0,917819
Activités	0,471519	0,921042
Environnement	0,65312	0,918137
Logement	0,258368	**0,92423**
Conjoint	0,47449	0,921066
Famille	0,631861	0,918496
Ami	0,546107	0,919892

Les items dont le retrait accroît la valeur du coefficient global sont : les sensations tactiles (0,927473), les vertiges (0,924545), le poids (0,923882), les maux de tête (0,923938) et le logement (0,924231). Les écarts sont à nouveau très faibles, et nous ne pensons pas qu'ils soient à prendre en considération.

En conclusion, nous pensons avoir démontré la validité et la fiabilité de nos deux questionnaires dichotomiques, auxquels on devra cependant retirer l'item poids qui est non significatif pour les deux populations testées. Il est à noter par ailleurs que son retrait ne pourra qu'accroître la valeur du coefficient alpha.

BIBLIOGRAPHIE.

BALDWIN S., GODFREY C., PROPPER C. (1990)
"Quality of life".
Routledge, London.

BOYLE G. (1991)
"Does item homogeneity indicate internal consistency or item redundancy in psychometric scales?"
Person individ diff, 12(3) : 291-294.

BRAVO G., POTVIN L.(1991)
"Estimating the reliability of continuous measures with Cronbach's alpha or intraclass correlation coefficient : toward the integration of two traditions."
J.Clin.Epidemiol, 44(4-5) : 381-390.

BUCQUET D. (1993)
"Qualité de vie, santé perceptuelle. Définition, concept, évaluation."
in : HERISSON C., SIMON L. "Evaluation de la qualité de vie". Problèmes en médecine de rééducation. Masson, Paris, 1993.

CAMPBELL D., FISKE D. (1959)
"Convergent and discriminant validation by multitrait-multimethod matrix."
Psychological bulletin, 56 : 81-105.

CARMINES E.G., ZELLER R.A. (1988)
"Reliability and validity assessment".
Sage publications, Beverly Hills, London.

COSTE J. (1995)
"Methodological and statistical problems in the construction of composite measurement scales."
Statitics in medicine, 14 : 331-345.

CRONBACH L.J. (1951)
"Coefficient alpha and the internal structure of tests."
Psychometrika, 16(3) : 297-333.

CRONBACH L.J. (1955)
"Construct validity in psychological tests"?
Psychol Bull., 52 : 281-302.

CRONBACH L.J., GLESER G.C., NANDA H. RAJARATNAM N. (1971)
"The dependability of behavioral measurements."
New York, Wiley.
DAVID C, HADORN M.D. (1991).
"Multitrait.Multimethod analysis of health related quality of life measures".
Medical Care, 29(9) : 829-840.
DICKES P., TOURNOIS J., FLIELLER A., KOP J.L. (1994)
"La psychométrie".
Paris, PUF.
EVANS W.J. et al (1981)
"Determining the generalizability of rating scales in clinical settings".
Medical Care, 19(12) : 1211-1220.
GLESER G.C., CRONBACH L.J. (1965).
"Generalizability of scores influenced by multiple sources of variance"
Psychometrika, 30(4) : 395-418.
GUYATT G.H. , DEYO R. (1989).
"Responsiveness and validity in health status measurement : a clarification."
J.Clin.Epidemiol, 42(5) : 403-408.
GUYATT G.H. et al (1992).
"Measuring health status : what are the necessary measurement properties?)
J.Clin.Epidemiol, 45(12) : 1341-1345.
GUYATT G.H. , FEENY D., PATRICK D (1993).
"Measuring health related quality of life."
Annals of internal medicine, 118 : 622-629.
HARDON D. (1991)
"The role of public values in setting health care priorities".
Soc. Sci. Med, 52 : 773.
HATTIE J. (1985)
"Methodology review : assessing unidimensionality of tests and items"
Applied psychological measurement, 9 : 139-164.
HELMSTADTER G.C. (1966)
"Principles of psychological measurement"
London, methnen.
HERISSON C., SIMON L.(1993).
"Evaluation de la qualité de vie"
Masson, Paris.
KAHNEMAN D., TVERSKY A. (1984)
"Choices, values and frames".
Am. Psychologist, 39(4) : 341-350.

KIRSCHNER B., GUYATT G. (1985)
"A methodological framework for assessing health indices".
J.Chron.Dis., 38(1) : 27-36.
KLINE P. (1979)
"Psychometrics and psychology"
Academic press, London.
KLINE P. (1986)
"A handbook of test construction : introduction to psychometric design"
Methuen, New York.
KRISTOFF W. (1991)
"Klassische testtheorie und testkonstruktion."
in Feger H. Bredenkamp J. (eds), Mesen und testen, vol 3,. Göttingen, Hogrefe, 1991, pp 544-603.
KUDER G.F., RICHARDSON M.W. (1937.
"The theory of the estimation of test reliability."
Psychometrika, 2 : 151-160.
LAUNOIS R. et al (1994)
"Construction et validation d'un indicateur spécifique de qualité de vie : le cas des insuffisances veineuses".
Journal.d'économie médicale, 12(2-3) : 109-126.
MARQUIS P. (1995).
"Mise au point d'une échelle qualité de vie : aspects méthodologiques".
Le courrier de l'évaluation en santé, n°7, 18-19.
Mc DONALD R. (1981)
"The dimensionality of tests and items"
British journal of mathematical and statistical psychology, 34 : 110-117.
Mc DOWELL I., NEWELL C. (1987)
"Measuing health"
Oxford University Press, Oxford.
MORET L., CHWALLOW J. et al (1993)
"Evaluer la qualité de vie : construction d'une échelle".
Rev.Epidem. et Santé Publ., 41 : 65-71.
NOVICK M., LEWIS C. (1967)
"Coefficient alpha and the reliability of composite measurement."
Psychometrika, 32(1) : 1-13.
NUNNALY J. (1978)
"Psychometric theory".
2nde édition, New York, Mc Grawhill book company.

PATRICK D. (1995)
"Qualité de vie : définition et évaluation.
Le courrier de l'évaluation en santé, mai 1995, n°7.
RAY J. (1988)
"Semantic ovezrlap between scale items may be a good thing : reply to Smedslund"
Scandinavian journal of psychology, 29 : 145-147.
SAS user's guide : statistiques (1990).
Version 6, 4ème édition.
Cary NC, USA.
SPEARMAN C. (1910)
"Correlation calculed from faulty data".
British journal of psychology, 3 : 271-295.
STREINER D.L., NORMAN G.R. (1989)
"Health measurement scales : a practical guide to their development and use".
Oxford. Oxford University Press.