METHODOLOGIE

Fabienne **FORTIN, PH-D. Professeur** Faculté des Sciences **infirmières - Université de** Montréal

PROPRIÉTÉS MÉTROLOGIQUES DES INSTRUMENTS DE MESURE (FIDÉLITÉ - VALIDITÉ)

La qualité de la recherche ne dépend pas uniquement du devis de recherche (schéma opératoire), mais aussi **de la qualité métrologique des outils de mesure. Qu'ils soient développés à des fins de recherche ou adaptés et traduits dans une autre langue, les outils de mesure doivent être fidèles et valides. Toute mesure prend forme à partir de la question de recherche et des** définitions conceptuelle et opérationnelle des concepts à l'étude. Comment peut-on assurer la qualité des outils **de mesure si ce n'est au moyen des processus de fidélité et de validité ? Bien que les concepts de fidélité et de validité soient des sujets complexes et souvent controversés, ils demeurent particulièrement importants en ce qui touche la mesure des** comportements, **des attitudes ou des habiletés. Quand il s'agit de mesu**rer des propriétés physiques ou physiologiques telles la **pulsation, la tension artérielle, c'est beaucoup plus simple que de pondérer ou mesurer, par exemple le degré d'autoritarisme, le degré d'anxiété ou encore** l'état de santé. Dans ces cas, il n'existe pas de règles à utiliser, pas plus que d'échelles de la mesure directe de **ces phénomènes. Il devient donc nécessaire de créer ou d'élaborer des moyens indirects pour mesurer les indicateurs de ces concepts. Ces moyens peuvent être parfois trop indirects, rendant ainsi la validité difficile à réaliser. En outre, il faut se rappeler qu'il n'existe pas de validité absolue et qu'il appartient aux chercheurs de déterminer les types de validité et de fidélité correspondant au contexte d'application des instruments de mesure et au but qu'ils se proposent d'atteindre. Par exemple, un instrument de mesure peut être valide à certaines fins et non à d'autres. C'est pourquoi, on ne demande pas « tel outil de mesure est-il valide ? », mais plutôt « est-il valide dans le contexte de son application ? ». Selon BRINBERG et** McGRATH (1985), **la validité c'est l'intégrité, la qualité de la mesure à évaluer** dans des circonstances particulières. Un outil peut être **valide pour mesurer le degré d'autonomie dans les activités de la vie quotidienne** (AVQ) **auprès d'une population âgée, mais non valide si utilisé auprès des** jeunes.

La fidélité et la validité sont des caractéristiques **essentielles de tout instrument de mesure. La fidélité et la** validité des instruments s'évaluent en degré, et non par leur présence ou leur absence. La fidélité est un préala**ble à la validité, c'est-à-dire que si un instrument de** mesure n'attribue pas les scores de façon consistante, **il ne peut être utile pour atteindre le but proposé. Toutefois, la fidélité n'est pas une condition suffisante pour établir la validité. Un instrument peut mesurer un phénomène de façon constante, mais cela ne signifie pas qu'il mesure bien le phénomène que l'on veut mesurer** (WALTZ, **STRICKLAND et LENZ, 1991). Une des distinctions qui caractérisent les deux concepts a trait à l'objet d'évaluation. L'étude de la fidélité réfère à l'évaluation du degré de corrélation d'un instrument de mesure avec lui-même ; l'étude de la validité réfère à l'évaluation du degré de corrélation d'un instrument** avec autre chose que l'instrument de mesure lui-même.

Les concepts de fidélité et de validité ne s'appliquent pas uniquement aux nouveaux instruments de mesure, mais également aux traductions d'instruments dans une autre langue.

LA FIDÉLITÉ DES MESURES

La fidélité réfère au degré de constance ou d'exactitude **avec lequel un(e) instrument ou technique mesure le concept ou le phénomène** qu'il(elle) **est supposé(e) mesurer.** Par **exemple, si un individu se pèse plusieurs fois consécutives** sur **un pèse-personne, on s'attend à ce que son poids soit le même d'une fois à l'autre. Quand on considère la fidélité d'un instrument de mesure, on s'intéresse aux erreurs aléatoires qui peuvent provenir de conditions temporaires chez** l'individu ou de l'administration de l'instrumentation qui varie **dans le temps. Sous-jacente au concept de fidélité, se trouve l'estimation de la proportion de la variation** totale des scores qui serait due au hasard ou aux erreurs **découlant du hasard.**

La fidélité, étant une propriété des instruments de mesure, peut être estimée de différentes façons, au moyen **de procédés statistiques, en utilisant les données d'un**

PROPRIÉTÉS MÉTROLOGIQUES DES INSTRUMENTS DE MESURE (FIDELITE - VALIDITÉ)

échantillon de sujets, Puisque tous les types de fidélité ont trait au degré d'accord ou de concordance entre deux ensembles de scores, la fidélité s'exprime sous la forme d'un coefficient de corrélation qui s'établit de ,00 pour l'absence de fidélité à 1,00 pour une fidélité parfaite. Les coefficients de fidélité sont des coefficients de détermination ; mis au carré, ils représentent la proportion de la variance d'une variable mesurée qui est la vraie variance. Un coefficient de fidélité de 0,8 indique que la mesure reflète 64 % de la *vraie* variance du concept mesuré. Qu'en est-il du coefficient-critère qui permet de conclure qu'une mesure est fidèle à partir d'un critère ? Les travaux recensés suggèrent qu'un coefficient devrait excéder 0,7 s'il s'agit de nouvelles échelles et de 0,8 dans le cas d'échelles bien rodées. Un coefficient de corrélation de 0,7 indique que 50 % de la variance dans les scores obtenus avec l'échelle est la variance d'erreur.

Les sources de variation dans la mesure des concepts

Quand des évaluations sont effectuées auprès d'un échantillon de sujets ou de cas, on peut s'attendre à trouver un certain degré de variation parmi ces mesures. Une partie de cette variation est due aux différences actuelles qui existent entre les sujets ou les cas. L'autre partie de la variation observée est due au processus de mesure lui-même. Il sera utile d'avoir une idée du déroulement du processus de la mesure de manière à distinguer la variance attribuée à l'erreur de mesure et celle de la vraie différence qui existe entre les sujets ou cas.

L'erreur de mesure peut être attribuable aux différences entre les individus qui appliquent la mesure, entre les mesures successives appliquées au même sujet ou cas par le même observateur, ou entre les formes parallèles d'un même instrument de mesure.

Si nous avons à utiliser un instrument de mesure, soit une échelle, un inventaire d'items, qui a été développé par d'autres chercheurs, on s'attend à ce que ceux-ci aient déjà publié des données pour démontrer quelque évidence de la fidélité de leur instrument.

Chacune des façons d'estimer la fidélité reflète un type différent d'information sur la performance de l'instrument de mesure. L'estimation de la fidélité repose sur trois aspects de la fidélité : la stabilité, l'homogénéité et l'équivalence.

TYPES DE FIDÉLITÉ

La stabilité

La stabilité est la constance ou la stabilité dans les réponses quand un instrument est appliqué plusieurs fois à des sujets. L'estimation de la stabilité d'un instrument de mesure se fait à l'aide d'une technique qui évalue la fidélité test-retest.

L'intervalle de temps entre les applications du même instrument de mesure à un groupe de sujets s'établit approximativement deux semaines après le premier test. La technique du test-retest est plus appropriée pour apprécier la fidélité des traits stables comme les mesures affectives (WALTZ, STRICKLAND et LENZ, 1991). Le postulat de base de la technique de stabilité sous-entend que le facteur à mesurer demeure constant à deux temps de mesure. Le chercheur doit s'assurer qu'il n'y a pas d'activités qui interviennent entre les deux administrations du test qui pourraient affecter la stabilité de la caractéristique qui est mesurée (WALTZ, STRICKLAND et LENZ, 1991).

Pour produire un coefficient de stabilité entre deux ensembles de scores, la corrélation de PEARSON (r) est utilisée ou la corrélation par rang de SPEARMAN s'il s'agit de données de niveau nominal ou ordinal. Un coefficient de stabilité élevé signifie que les mesures ont peu changé entre le passage du test et du retest, c'est-à-dire que l'instrument évalue le même phénomène.

La consistance interne

La consistance interne réfère à l'homogénéité de l'instrument de mesure. L'évaluation de la consistance interne repose sur le postulat que l'instrument est unidimensionnel, c'est-à-dire qu'il mesure un seul concept. Si l'instrument de mesure **contient** plusieurs dimensions (sous-concepts), la consistance interne devra être estimée pour chaque dimension. Les principales méthodes pour estimer la consistance interne sont : l'alpha de CRONBACH, le coefficient KUDER-RICHARDSON, les corrélations inter-item total et la fidélité moitié-moitié.

L'alpha de CRONBACH est l'approche la plus souvent utilisée pour estimer la consistance interne d'une

échelle de mesure quand il y a plusieurs choix dans l'établissement des scores, telle l'échelle de type Likert. L'alpha varie de ,00 à 1,00 : ,00 dénote une absence d'homogénéité et 1,00 une homogénéité parfaite. Généralement, on s'attend à un certain degré de fidélité, c'est pourquoi les valeurs de ,00 et 1,00 se retrouvent rarement comme résultat d'un test. Le KUDER-RICHARDSON (KR-20) est utilisé pour estimer la consistance interne des échelles dichotomiques, c'est-à-dire celles qui ont deux réponses seulement (oui ou non). Le KR-20 varie de ,00 à 1,00, la valeur la plus élevée dénote une plus grande consistance interne.

Les corrélations inter-item total. Les corrélations inter-item total sont une fonction du degréd'inter-relation de tous les items d'une échelle mesurant un seul concept. Il s'agit d'évaluer un groupe de corrélations simples entre chaque item et le score total pour l'échelle au complet. Les corrélations indiquent dans quelle mesure les items individuels sont reliés au score total.

La fidélité moitié-moitié. Selon cette technique, les items sont divisés en deux moitiés, soit à l'aide d'une répartition aléatoire, soit à l'aide d'une répartition des items pairs/impairs. Les moitiés sont appliquées, aux sujets et les scores sont corrélés les uns aux autres. Si les scores des deux moitiés ont une corrélation élevée, c'est une bonne évidence de la consistance interne de l'échelle de mesure.

L'équivalence

L'équivalence réfère au degré de similarité entre deux versions ou formes parallèles ou plus d'un instrument de mesure ou encore entre deux observateurs ou plus mesurant le même phénomène. Les versions doivent être équivalentes. L'équivalence de l'instrument est établie en corrélant les scores des deux formes d'une échelle.

La fidélité entre les observateurs réfère au degré d'exactitude entre deux observateurs ou plus qui ont appliqué les mêmes instruments de mesure à un même groupe de sujets. Le coefficient de fidélité reflète l'exactitude entre les estimations ou scores des observateurs plutôt que l'exactitude de l'instrument lui-même. La fidélité entre les observateurs est estimée selon un pourcentage d'accord quand deux individus ou plus appliquent les mêmes instruments de mesure ou observent un même phénomène.

LA VALIDITÉ DES MESURES

La validité d'un instrument de mesure réfère au degré auquel l'instrument mesure ce qu'il est supposé mesurer, c'est-à-dire le concept. La validité est concernée par l'erreur systématique. Tout comme la fidélité, la validité peut être appréciée de différentes façons ; cependant, elle demeure plus difficile à établir (POLIT et HUNCLER, 1991). Les trois principales façons d'estimer la validité, discutées dans les écrits sont : la validité de contenu, la validité de critère et la validité de construit. Bien que la littérature plus récente considère la validité dans un sens plus global, incluant les trois types de validité en lien avec la validité de construit (BERK, 1990 ; REW, STUPPY et BECKER, 1988 ; APA, 1985), nous distinguerons dans ce texte les trois types de validité les plus souvent discutés.

TYPES DE VALIDITÉ

La validité de contenu

La validité de contenu réfère à la représentation adéquate de l'échantillon d'items utilisés pour mesurer un comportement ou une caractéristique. Est-ce que l'instrument de mesure et les items qu'il contient représentent vraiment le domaine étudié ?A titre d'exemple, si on prend un concept à plusieurs facettes tel que l'aliénation, ce concept peut être étudié sous différents aspects tels impuissance, délinquance, isolement social, etc. ; la validation de contenu exige que toutes ces facettes soient représentées par plusieurs items dans le test. GREEN et LEWIS (1986) considèrent cinq étapes dans l'établissement de la validité de contenu : la recension des écrits sur le domaine, les réflexions personnelles sur la signification du concept, l'identification des composantes du concept et leur utilisation dans d'autres travaux, l'identification des items et les analyses empiriques des items par l'étude des inter-relations entre les items. La validité de contenu, en général, n'a pas de fondement empirique et s'appuie surtout sur des jugements.

Validité nominale et validité par consensus. Deux types de validité de contenu sont basés sur le jugement des pairs. La validité nominale est le processus par lequel un expert juge de la validité d'un instrument de mesure en ce qu'il semble relié au concept d'intérêt. Cette approche ne représente pas l'évidence de la validité de

contenu, bien qu'elle soit reliée à la perception des sujets quant au concept à mesurer (LYNN, 1986 ; THOMAS, 1992). La validité par consensus est le processus par lequel un panel d'experts juge de la validité d'un instrument de mesure. Les juges doivent s'entendre sur l'étendue des items et dans quelle mesure les items reflètent le concept à l'étude et sont représentatifs.

La validité de critère

La validité de critère représente le degré d'association entre un instrument de mesure ou une technique et un autre instrument indépendant susceptible de mesurer le même phénomène ou concept. Le critère, c'est la deuxième mesure qui sert à évaluer le même concept et à déterminer le degré de corrélation. La difficulté réside dans le choix d'un critère fidèle et approprié pour fins de comparaisons avec les scores de l'instrument de mesure.

Les deux formes de validité de critère sont la validité concomitante et la validité prédictive. La *validité concomitante* représente le degré de corrélation entre deux mesures du même concept appliquées en même temps. Par exemple, l'échelle SCL-90-R des symptômes primaires de santé mentale et l'échelle MMPI mesurant des symptômes similaires administrées aux mêmes sujets *en même temps.* La *validité prédictive représente* le degré de corrélation entre une mesure du concept et une mesure ultérieure du même concept. Par exemple, les scores obtenus à un test d'aptitudes par des étudiants à leur entrée dans un programme d'études et leur rendement évalué par les scores subséquents. En fait, tous les tests ont un potentiel de prédiction en ce qu'ils prédisent un certain type de résultat, qu'il soit présent ou futur. Exemples : les tests d'aptitudes prédisent la réussite future ; les tests d'intelligence prédisent l'habileté d'apprendre maintenant et dans l'avenir. Les corrélations s'établissent à l'aide du coefficient alpha de CRONBACH.

Alors que la validité avec corrélation de critère est examinée au cours d'une seule expérience, la validité de construit est estimée à travers un long processus d'effets réciproques entre l'observation, le raisonnement et l'imagination.

La validité de construit

La validité de construit tente de valider le corps théorique sous-jacent à la mesure et *de* vérifier *des* hypothèses d'associations. Le point central de la validité de « construit » repose sur le concept abstrait qui est mesuré et sa relation avec d'autres concepts. C'est aussi l'analyse de la signification des scores d'un test par rapport au concept *mesuré.*

Il existe trois parties dans le processus de validation des concepts : 1) on identifie les concepts qui pourraient expliquer le rendement d'un test, 2) on tire des hypothèses vérifiables de la théorie sous-jacente au concept et 3) on conduit une étude afin de vérifier les hypothèses formulées (THORNDIKE et HACEN, 1977 ; CRONBACH, 1971 ,1984 ; KERLINCER, 1986 ; NUNALLY, 1978).

Les deux principales approches pour évaluer la validité de construit sont la validité de convergence et la validité de différenciation.

La *validité de* convergence. Différentes échelles de mesure du même construit sont appliquées à des sujets. Les analyses de corrélation effectuées sur ces mesures doivent produire des résultats similaires.

La validité de différenciation. Des échelles de mesure reliées au concept d'intérêt sont appliquées à des sujets ; on évalue la capacité de ['instrument de mesure à différencier le construit mesuré des autres qui lui sont similaires.

Il existe plusieurs façons, entre autres, de vérifier la validité de convergence et de différenciation : l'approche *multi-trait-multi-méthode* (CAMPBELL et FISKE, 1959), l'analyse de facteurs, l'approche du groupe contraste, la spécificité et la sensibilité. *La spécificité* fait référence à la capacité de l'instrument à détecter les cas dans lesquels le phénomène n'est pas présent. *La sensibilité* fait référence à la capacité de l'instrument à détecter les vrais *cas* qui *manifestent le* phénomène.

Interprétation des coefficients

Les coefficients de fidélité sont d'importants indicateurs de la qualité d'un instrument de mesure. Il n'existe pas de standard précis pour juger d'un coefficient acceptable. Si un chercheur veut établir des comparaisons entre les scores de deux groupes de sujets, un coefficient de fidélité autour de 0,70 **serait** probablement suffisant (POLIT et HUNCLER, 1991). L'interprétation du coefficient de fidélité repose sur le degré de variabilité des scores entre les sujets qui possèdent plus ou moins de la caractéristique à mesurer.

La connaissance de la fidélité d'un instrument est utile non seulement pour l'interprétation des résultats, mais aussi pour suggérer des modifications. La fidélité des

échelles de mesure est tributaire du nombre d'items. L'ajout d'items peut contribuer à améliorer la fidélité de l'instrument.

En ce qui a trait à l'interprétation de la validité, *il* faut se référer au degré de validité qu'un instrument de mesure possède. La vérification de la validité d'un instrument de mesure n'est jamais « prouvée » en soi ; ce sont les applications de l'instrument qui sont validées dans des contextes différents. Cependant, il existe des échelles standardisées qui sont le produit de tests de validité répétés auprès d'une variété d'échantillons de sujets.

Il faut se rappeler aussi qu'un instrument de mesure qui a été traduit dans une autre langue a perdu de ses qualités métrologiques.

RÉFÉRENCES

American Psychological Association's Committee to Develop Standards (1985). — *Standards for educational and psychological testing.* Washington, DC : American Psychological Association.

BERK (R.A.) (1990). — Importance of expert judgment in content-related validity evidence. *Western Journal of Nursing Research, 12*(5), 659-671.

BRINBERG (D.), McGRATH (J.E.) (1985). — *Validity and the research process.* Beverly Hills, CA : Sage Publications.

CAMPBELL (D.), FISKE (D.) (1959). -Convergent and discriminant validation by the multi-trait-multi-method matrix. *Psychology Bulletin,* 53, 273-302.

CRONBACH (L.J.) (1984). — *Essentials of psychological testing,* 4e édition. New York : Harper et Row, Publishers.

CRONBACH (L.J.) (1971). -Test validation, dans R.L. Thorndike (ed.), *Educational Measurement,* 2e édition. Washington, DC : American Council of Education.

GREEN (L.), LEWIS (F.) (1986). — *Measurement and evaluation education and health promotion.* Palo Alto, CA : Mayfield.

KERLINGER (F.N.) (1986). — *Foundations of behavioral research,* 3e édition. New York : Holt, Renehart et Winston Inc.

LYNN (M.R.) (1986). — Determination and quantification of content validity. Nursing *Research, 35,* 382-385.

NUNALLY (J.C.) (1978). — *Psychometric Theory.* New York : McGraw-Hill Book Company.

POLIT (D.F.), HUNCLER (B.P.) (1991). — *Nursing research : Princip/es and methods,* 4e édition. Philadelphie : J.B. Lippincott.

REW (L.), STUPPY (D.), BECKER (H.) (1988). — Construct validity in instrument development : A vital Link between nursing practice, research, and theory. *Advance in Nursing Science, 10*(4), 1 O-22.

THOMAS (S.) (1992). -Face validity. *Western Journal of Nursing Research, 14*(1), 109-I 12.

THORNDIKE (R.L.), HACEN (E.) (1977). — *Measurement and evaluation in psychology and education,* 4e édition. New York : John Wiley and Sons.

WALTZ (C.F.), STRICKLAND (O.L.), LENZ (E.R.) (1991). — *Measurement in nursing research,* chap. 5. Philadelphie : F.A. Davis.