Faculté de Médecine Paris-Sud
ED 420 - Ecole Doctorale de Santé Publique

Département de médecine sociale et préventive
Ecole de Santé Publique

Mesures subjectives et épidémiologie

Problèmes méthodologiques liés à l'utilisation des techniques psychométriques

par

Alexandra ROUQUETTE

Thèse de doctorat en cotutelle France-Canada
présentée et soutenue publiquement le 16 décembre 2014 à la Faculté de Médecine Paris-Sud
en vue de l'obtention du grade de Philosophiae Doctor (Ph.D.)
en **Santé Publique** option **Epidémiologie**
sous la direction de **Sylvana CÔTÉ** et **Bruno FALISSARD**

Septembre, 2014

JURY :

Jean BOUYER (Ph.D., Université Paris-Sud, France), Président
Lise GAUVIN (Ph.D., Université de Montréal, Canada), Président-rapporteur
Serge BRIANÇON (M.D., Ph.D., Université de Nancy, France), Rapporteur / Examinateur Externe
Marie-Hélène MAYRAND (M.D., Ph.D., Université de Montréal, Canada), Rapporteur
Pascale TUBERT-BITTER (Ph.D., Université Paris-Sud, France), Examinateur
Sylvana CÔTÉ (Ph.D., Université de Montréal, Canada), Directeur
Bruno FALISSARD (M.D., Ph.D., Université Paris-Sud, France), Directeur

RESUME

L'utilisation des mesures subjectives en épidémiologie s'est intensifiée récemment, notamment avec la volonté de plus en plus affirmée d'intégrer la perception qu'ont les sujets de leur santé dans l'étude des maladies et l'évaluation des interventions. La psychométrie regroupe les méthodes statistiques utilisées pour la construction des questionnaires et l'analyse des données qui en sont issues. Ce travail de thèse avait pour but d'explorer différents problèmes méthodologiques soulevés par l'utilisation des techniques psychométriques en épidémiologie. Trois études empiriques sont présentées et concernent 1/ la phase de validation de l'instrument : l'objectif était de développer, à l'aide de données simulées, un outil de calcul de la taille d'échantillon pour la validation d'échelle en psychiatrie ; 2/ les propriétés mathématiques de la mesure obtenue : l'objectif était de comparer les performances de la différence minimale cliniquement pertinente d'un questionnaire calculée sur des données de cohorte, soit dans le cadre de la théorie classique des tests (CTT), soit dans celui de la théorie de réponse à l'item (IRT) ; 3/ son utilisation dans un schéma longitudinal : l'objectif était de comparer, à l'aide de données simulées, les performances d'une méthode statistique d'analyse de l'évolution longitudinale d'un phénomène subjectif mesuré à l'aide de la CTT ou de l'IRT, en particulier lorsque certains items disponibles pour la mesure différaient à chaque temps. Enfin, l'utilisation de graphes orientés acycliques a permis de discuter, à l'aide des résultats de ces trois études, la notion de biais d'information lors de l'utilisation des mesures subjectives en épidémiologie.

Mots-clefs : mesures subjectives, psychométrie, questionnaire, variable latente, épidémiologie, longitudinal, biais d'information, graphe acyclique orienté

ABSTRACT

Title: Subjective measurements and epidemiology: methodological issues raised by the use of psychometric techniques

Recently, subjective measurements have increasingly been used in epidemiology, alongside the growing will to integrate individuals' point of view on their health in studies on diseases or health interventions. Psychometrics includes statistical methods used to develop questionnaires and to analyze questionnaire data. This doctoral dissertation aimed to explore methodological issues raised by the use of psychometric techniques in epidemiology. Three empirical studies are presented and cover 1 / the validation stage of a questionnaire: the objective was to develop, using simulated data, a tool to determine sample size for internal validity studies on psychiatric scale; 2 / the mathematical properties of the subjective measurement: the objective was to compare the performances of the minimal clinically important difference of a questionnaire, assessed on data from a cohort study, computed using the classical test theory (CTT) framework or the item response theory framework (IRT); 3 / its use in a longitudinal design: the objective was to compare, using simulated data, the performances of a statistical method aimed to analyze the longitudinal course of a subjective phenomenon measured using the CTT or IRT framework, especially when some of the available items used for its measurement differ at each time of data collection. Finally, directed acyclic graphs were used to discuss the results from these three studies and the concept of information bias when subjective measurements are used in epidemiology.

Keywords: subjective measurement, psychometrics, questionnaire, latent variable, epidemiology, longitudinal design, information bias, directed acyclic graphs

LABORATOIRES DE RATTACHEMENT

Unité Inserm U669 – Paris-Sud Innovation Group in adolescent Mental Health (PSIGIAM)

Directeur : Bruno Falissard

97 boulevard de Port-Royal,

75679 Paris cedex 14

France

Groupe de Recherche sur l'Inadaptation Psychosociale de l'enfant

(GRIP)

Directeur : Richard Tremblay

Université de Montréal,

3050 Édouard-Montpetit,

Montréal (Québec), H3T 1J7

Canada

LISTE DES PRODUCTIONS SCIENTIFIQUES

Publications scientifiques

Rouquette A, Falissard B. (2011) Sample Size Requirements for the Internal Validation of Psychiatric Scales. *International Journal of Methods in Psychiatric Research.* Vol. 20(4): 235-249

Rouquette A, Blanchin M, Sébille V, Guillemin F, Côté S, Falissard B, Hardouin JB. (2014) The Minimal Clinically Important Difference determined using Item Response Theory Models: an attempt to solve the issue of the association with baseline score. *Journal of Clinical Epidemiology*; 67(4):433-440.

Rouquette A, Côté S, Hardouin JB, Falissard B. Item Response Theory (IRT) used to enhance accuracy of data analyses in longitudinal studies: a simulation study. *En finalisation.*

Communications orales

Rouquette A, Côté S, Hardouin JB, Falissard B. Item Response Theory (IRT) used to enhance accuracy of data analyses in longitudinal studies of child development: a simulation study. Society for Research in Child Development, 2014 special topic meeting: Developmental Methodology. September 11-13, 2014, San Diego, CA, USA.

Rouquette A, Blanchin M, Sébille V, Guillemin F, Côté SM, Falissard B, Hardouin JB. Utilisation des modèles issus de la Théorie de Réponse à l'Item (IRT) pour la détermination de la Différence Minimale Cliniquement Pertinente d'un questionnaire. EPICLIN 7 - 7ème Conférence Francophone d'Épidémiologie Clinique, - Paris, 16 & 17 mai 2013

Rouquette A, Côté SM, Falissard B. Psychometry and epidemiology: use of Item Response Theory (IRT) models in longitudinal studies of child development. 2nd meeting of Marie Curie International Exchange Program, Montreal, Canada, June 3-4, 2012

Communications affichées

Rouquette A, Blanchin M, Sébille V, Guillemin F, Côté SM, Falissard B, Hardouin JB. Determination of the Minimal Clinically Important Difference (MCID) using Item Response Theory (IRT) models: an attempt to solve the issue of the association with baseline score. International workshop on "Response-Shift and subjective measures in health science", Nantes, 7 juin 2013

TABLE DES MATIERES

LISTE DES ABRÉVIATIONS

1PLM	One Parameter Logistic Model (modèle logistique à un paramètre)
2PLM	Two Parameters Logistic Model (modèle logistique à deux paramètres)
ACP	Analyse en Composantes Principales
AFC	Analyse Factorielle Confirmatoire
AFE	Analyse Factorielle Exploratoire
BDI	Beck Depression Inventory (inventaire de dépression de Beck)
BIC	Bayesian Information Criteria (critère d'information bayesien)
CIF	Classification Internationale du Fonctionnement, du handicap et de la maladie
CTT	Classical Test Theory (théorie classique des tests)
DAG	Directed Acyclic Graph (graphe acyclique orienté)
ddl	Degrés de liberté
DIF	Differential Item Functioning (fonctionnement différentiel d'item)
DMCP	Différence Minimale Cliniquement Pertinente
DMCP-Sc	Différence Minimale Cliniquement Pertinente déterminée à l'aide du Score
DMCP-Sc_{SI}	Différence Minimale Cliniquement Pertinente déterminée à l'aide du Score composée de plusieurs valeurs en fonction du Score Initial
DMCP-TL	Différence Minimale Cliniquement Pertinente déterminée à l'aide du Trait Latent
ET	Ecart-Type
$\mathcal{F}$	Loi de Fisher
GLM	Generalized Linear Model (modèle linéaire generalisé)
HAMD	HAMilton rating scale for Depression (échelle de dépression de Hamilton)
IC	Intervalle de Confiance
IIQ	Intervalle InterQuartiles
IRT	Item Response Theory (théorie de réponse à l'item)
LCGA	Latent Class Growth Analysis (analyse en classe latent de trajectoires)
LCGA-Sc	Latent Class Growth Analysis appliquée sur le Score

LCGA-θ_{est}	Latent Class Growth Analysis appliquée sur l'estimation du trait latent par un modèle de Rasch
LCGA-θ_{sim}	Latent Class Growth Analysis appliquée sur les valeurs simulées du trait latent
Méd	Médiane
MOS-SF36	Medical Outcomes Study, Short Form, 36 items
$\mathcal{N}$	Loi normale
NSN	Nombre de Sujets Nécessaire
OMS	Organisation Mondiale de la Santé
PCM	Partial Credit Model
PRO	Patient Reported Outcome
QT	Question de Transition
RMSEA	Root Mean Square Error Approximation
SDS	Zung Self-Rating Depression Scale
Se	Sensibilité
SEM	Strucural Equation Modeling (modèle d'équation structurelle)
SI	Score Initial
SP	Santé Perçue
Sp	Spécificité
T1	Premier temps de collecte
T2	Deuxième temps de collecte
TL	Trait Latent
VP	Valeur Prédictive
VPN	Valeur Prédictive Négative
VPP	Valeur Prédictive Positive

LISTE DES TABLEAUX

LISTE DES FIGURES

REMERCIEMENTS

En premier lieu, bien-sûr, à mes directeurs de thèse, Sylvana et Bruno, pour leurs enseignements, leur soutien, leur disponibilité, leur ouverture d'esprit qui m'ont permis d'avoir ce parcours si enrichissant et passionnant. Même si, d'accord, les démarches administratives étaient parfois un peu lourdes ; mais je referais le même, sans hésiter, parce que tout ce que j'ai pu apprendre à vos côtés, tout ce que vous m'avez permis de découvrir, toutes les rencontres que j'ai pu faire grâce à vous, tout ça me fait largement oublier les quelques situations déconcertantes que nous avons pu vivre grâce à cette cotutelle. Merci, sincèrement, d'avoir accepté de m'encadrer et de m'avoir accompagnée tout au long de ce projet.

Mes remerciements vont aussi aux équipes de l'Ecole Doctorale ED420 de l'université Paris-Sud (Audrey Bourgeois, Jean Bouyer), du programme de Ph.D. en Santé Publique (Louise Dubuc, Angélique de Chatigny, Maria-Victoria Zunzunegui, Louise Potvin) et de la Faculté des Etudes Supérieures et Postdoctorales de l'Université de Montréal (Stéphanie Tailliez, Yannick Tremblay) qui ont œuvré sans relâche et avec la plus grande gentillesse pour simplifier au maximum mon cheminement doctoral dans les deux pays.

Il m'est impossible de nommer tous les collègues qui, à Montréal, comme à Paris, ont été d'un soutien précieux à bien des moments mais j'aimerais remercier les équipes de l'unité Inserm U669 et du GRIP qui m'ont accueillie chaleureusement. Une mention spéciale, quand même, aux voisines de bureau : Nadine Provençale, Marie-Claude Salvas, Christine Hassler (un grand merci pour ta relecture !), Caro Barry. Et puis aussi aux groupes des « docs & post-docs » avec qui les lunchs à HEC, les cinq-à-sept, les RDV académiques et autres épluchades de blé d'inde étaient tellement agréables… Ah oui, et surtout Laura Pryor ! Merci infiniment pour ton aide et ton amitié pendant ces quatres années, tes livres, tes documents de cours, tes conseils pour les examens… Et voilà que je te demande encore de faire, à ma place, le dépôt de ma thèse à Montréal… Un grand merci à Lionel Riou-Franca aussi pour sa relecture attentive, même demandée à la dernière minute !

Deux autres équipes ont été très importantes dans ce parcours de doctorante et je tiens à leur signifier : l'EA4275 à Nantes et l'équipe de Biostatistique et d'Epidémiologie de l'Hôpital Hôtel-Dieu à Paris. Merci à Véronique Sébille, Myriam Blanchin et, bien-sûr, à Jean-Benoit

Hardouin, le Raschiste, co-auteurs de certains articles issus de ce travail de thèse et avec qui j'ai beaucoup de plaisir à travailler. Merci aussi à Francis Guillemin de Nancy qui a collaboré au deuxième travail présenté dans cette thèse et avec qui, je l'espère, d'autres travaux en collaboration suivront. A Paris, je tiens à remercier Joël Coste pour sa patience devant le temps que j'ai consacré à ce travail de thèse dans son unité et aussi pour avoir partagé ses connaissances sur la mesure, j'ai beaucoup appris ces deux dernières années. Merci à Sophie Grabar pour tous ses conseils, son oreille attentive et pour avoir bien voulu, avec Jimmy Mullaert, faire les cobayes pour l'introduction de ce manuscrit et me rassurer sur la voie que j'avais décidée de prendre.

Merci aux amis du Québec et de France. Le seul souci avec cette cotutelle, c'est que maintenant, où que j'aille, je suis toujours trop éloignée d'une partie d'entre vous… Agnès, merci beaucoup (et à ta maman !) pour ce travail ingrat qu'est la recherche des fautes d'orthographes. Enfin, toute cette aventure n'aurait pas pu avoir lieu sans le soutien inconditionnel et la patience de ma famille que je ne remercierai jamais assez.

****

Avant le dépôt final de ce manuscrit, j'aimerais exprimer mes plus sincères remerciements à Mesdames et Messieurs les membres du Jury. A Madame Lise Gauvin et Monsieur Jean Bouyer qui m'ont fait l'honneur de bien vouloir présider ce Jury de part et d'autre de l'Atlantique. A Madame Marie-Hélène Mayrand et Monsieur Serge Briançon qui ont accepté d'être les rapporteurs de ce travail et à Madame Pascale Tubert-Bitter qui a accepté de juger cette thèse. Veuillez trouver ici le témoignage de ma profonde gratitude.

CHAPITRE I : INTRODUCTION GENERALE

En épidémiologie, l'utilisation des questionnaires et échelles de mesure est allée crescendo dans la deuxième moitié du XX^{ème} siècle, en même temps qu'évoluait la définition de cette discipline (Clancy et Eisenberg, 1998; Evans, 2001; Rumeau-Rouquette *et al.*, 1986). En 1970, l'épidémiologie était encore définie de manière classique comme « l'étude de la distribution et des déterminants de la fréquence des maladies dans les populations humaines » (MacMahon et Pugh, 1970). Une profonde réflexion sur la définition de la santé et sur la compréhension de ses déterminants était pourtant déjà en cours depuis quelques décennies comme en témoigne l'intégration de la notion de « bien-être » par l'Organisation Mondiale de Santé (OMS) dans sa définition de la santé dès 1946 (Organisation Mondiale de la Santé, 1946). Par la suite, l'émergence d'une vision plus globale des déterminants de la santé a mis en avant le rôle que pouvaient jouer l'environnement, les habitudes de vie, l'organisation des systèmes de soins, etc. en plus de la biologie humaine dans le chemin causal vers la santé (Lalonde, 1974; Organisation Mondiale de la Santé, 1986). Au fur et à mesure, l'épidémiologie a donc adopté cette vision globale et positive de la santé et en 2008, Rothman et Greenland proposaient d'élargir la définition de l'épidémiologie à « l'étude de la distribution des états et évènements liés à la santé dans les populations ». L'intention était de ne plus centrer l'attention uniquement sur les maladies mais aussi sur « les états physiologiques tels que la tension artérielle, les mesures psychologiques telles que les scores de dépression et les évènements positifs tels que l'acquisition d'une immunité contre une maladie » (Rothman *et al.*, 2008).

Cet élargissement du champ de l'épidémiologie a aussi été largement favorisée par une volonté de plus en plus affirmée chez les cliniciens et les décideurs politiques de mieux tenir compte de la perspective « interne » des patients concernant leur état de santé (Boini *et al.*, 2010; Clancy et Eisenberg, 1998; Evans, 2001; Roger, 2011; Sullivan, 2003). En effet, la perception que les individus ont de la santé et de la maladie influence leurs comportements ; l'étude des déterminants et des conséquences de la santé perçue pourrait donc permettre la mise en évidence de cibles potentielles d'interventions pour la prévention des maladies ou la promotion de la santé, au niveau individuel ou populationnel (Briançon *et al.*, 2011; National

Institutes of Health - Office of Behavioral & Social Sciences Research, 2010). Cependant, mesurer des phénomènes tels que perçus par les individus est encore, à l'heure actuelle, hors de portée des instruments de mesure physiques classiquement utilisés en épidémiologie pour mesurer par exemple, le poids, la tension artérielle ou encore pour diagnostiquer des maladies génétiques, bactériennes, endocriniennes, etc. Ces phénomènes perçus sont abstraits, donc non directement mesurables, et subjectifs, donc difficilement évaluables autrement que rapportés par le sujet lui-même. L'intérêt croissant pour l'étude de ces états et évènements liés à la santé tels que perçus par les individus a donc nécessité le développement et l'utilisation par les épidémiologistes d'instruments permettant de mesurer ces phénomènes subjectifs.

Ces instruments se présentent généralement sous la forme de questionnaires composés d'une liste de questions, les items, ayant un dispositif de réponse à plusieurs modalités cotées. La mesure du phénomène étudié pour un individu est habituellement représentée par un score, c'est-à-dire la somme, éventuellement pondérée, des réponses de cet individu à chacun des items du questionnaire. Dans le champ de l'épidémiologie clinique, par exemple, l'utilisation de questionnaires est devenue très fréquente avec le développement des instruments de type « Patient Reported Outcome » (PRO). Ils permettent, dans les études évaluant une intervention, de mesurer les effets de l'intervention testée rapportés par le patient, donc prenant en compte son jugement pour statuer sur son efficacité (US Department of Health and Human Services, 2009). Plus largement, de nombreux instruments de mesure subjective[1] sont maintenant couramment utilisés en épidémiologie pour étudier les liens qui peuvent exister entre la santé et des concepts tels que la qualité de vie, la dépression, l'estime de soi, le soutien social, la douleur, etc. (Evans et Marmor, 1996; Organisation Mondiale de la Santé, 1996).

Les méthodes statistiques utilisées pour la construction de ces instruments et pour l'analyse des données qui en sont issues sont regroupées sous le nom de méthodes psychométriques. La psychométrie est définie comme l'étude de la mesure des caractéristiques

[1] Cette métonymie, remplaçant l'expression « mesure d'un phénomène subjectif », sera utilisée tout au long du texte pour en simplifier la lecture.

psychologiques telles que les capacités, les aptitudes, les performances intellectuelles, les traits de personnalité, la connaissance (Armitage et Colton, 1998). Cette discipline s'est développée à la fin du XIXème siècle avec le recours de plus en plus fréquent à la quantification, à la mesure numérique des phénomènes en psychologie (Martin, 1997). Francis Galton (1822-1911), James Cattel (1860-1940), Charles Spearman (1863-1945), Louis Thurstone (1887-1955), Lee Cronbach (1916-2001) sont quelques-uns des scientifiques issus de la psychométrie et dont les noms sont indissociables de l'histoire des statistiques. Que l'origine du développement des méthodes permettant la mesure de phénomènes subjectifs se situe en psychologie n'est pas surprenant et, dans le domaine médical, c'est d'ailleurs en psychiatrie que leur utilisation est la plus répandue. Encouragés par l'évolution de leur discipline au cours de la seconde moitié du XX^{ème} siècle, les épidémiologistes se sont naturellement intéressés à ces techniques psychométriques leur permettant d'élargir la palette des variables prises en compte dans leurs modèles statistiques aux phénomènes subjectifs.

Comme tout instrument de mesure, les questionnaires doivent faire l'objet d'une évaluation métrologique ayant pour but d'apprécier les propriétés de la mesure obtenue. Trois grandes propriétés sont exigées : la mesure doit être valide (elle mesure bien ce qu'elle est censée mesurer), elle doit être fiable (elle est reproductible aussi longtemps que les conditions de mesure ne changent pas) et elle doit être sensible au changement (Falissard, 2008; Paolaggi et Coste, 2001). Cependant, lorsque le phénomène mesuré est subjectif, la signification de ces trois propriétés est bien plus équivoque que pour une mesure objective (i.e. n'impliquant aucun jugement personnel, ni de la personne qui mesure, ni de la personne objet de la mesure) (de Vet *et al.*, 2011). En effet, la définition du phénomène (le « construit » dans la tradition psychométrique) est rarement unanime et peut varier d'un expert à l'autre, d'un pays à l'autre, d'une époque à l'autre, etc., or la validité de la mesure est dépendante de cette définition (Nunnaly, 1978). Aussi, un effet d'apprentissage peut survenir lors de l'administration répétée d'un questionnaire à un individu ; l'évaluation de la fiabilité et de la sensibilité au changement de l'instrument devient ainsi plus complexe. Un autre exemple de phénomène complexifiant l'évaluation métrologique de ce type d'instrument est celui du « response shift » où l'interprétation des items du questionnaire est modifiée par certaines des expériences vécues

par les sujets ; la comparabilité des différentes mesures obtenues avec ce questionnaire au cours du temps chez un même sujet devient dans ce cas questionnable (Sprangers et Schwartz, 1999).

En épidémiologie, les qualités métrologiques des instruments de mesure utilisés sont primordiales pour assurer la validité interne d'une étude. Les erreurs faites lors de la mesure des informations nécessaires à l'estimation de l'effet étudié peuvent être à l'origine d'un biais appelé classiquement « biais d'information » ou « biais de mesure » (Bouyer, 2009; Rothman *et al.*, 2008; Szklo et Nieto, 2007). Ce travail de thèse a pour but d'explorer différents problèmes méthodologiques soulevés par l'utilisation des instruments de mesure subjective en épidémiologie. En particulier, une mise en perspective de la notion de biais d'information dans ce cadre sera discutée à l'aide des résultats des trois études empiriques présentées dans ce travail et concernant 1/ la phase de validation de l'instrument, 2/ les propriétés mathématiques de la mesure obtenue, 3/ son utilisation dans un schéma longitudinal.

Dans la première partie de cette introduction seront présentés les différents modèles statistiques utilisés en psychométrie. Ensuite, une deuxième partie exposera les trois questions de recherche auxquelles s'est intéressé ce travail de thèse, illustrées par les trois études présentées dans les chapitres 2 à 4. Enfin, une troisième partie présentera les matériels et méthodes utilisés pour étudier ces questions de recherche, ainsi que le cadre méthodologique emprunté pour la discussion générale des résultats des trois études présentée au chapitre 5.

1. Contexte théorique

Le modèle de mesure est l'algorithme utilisé pour transformer les réponses obtenues à l'ensemble des items du questionnaire en une seule valeur numérique, la mesure (Falissard, 2008). Depuis la fin du XIX^{ème} siècle, les psychométriciens ont développé différentes théories décrivant les relations entre les items d'un questionnaire et le construit à mesurer. Ces théories sont utilisées comme cadre conceptuel lors de l'établissement du modèle de mesure d'un questionnaire.

a. La théorie classique des tests

Pendant longtemps, la Théorie Classique des Tests (CTT : Classical Test Theory) a, de loin, été la plus utilisée en psychométrie. Cette théorie postule que le score observé (s) pour un individu *i* à un test psychologique peut être décomposé en deux termes : son vrai score (*t*) et un terme d'erreur (e).

$$s_i = t_i + e_i$$

Le vrai score d'un individu est donc conceptualisé dans cette théorie comme la moyenne des scores observés à ce test administré de manière répétée à cet individu, la moyenne des termes d'erreurs sur les J répétitions étant supposée nulle.

$$t_i = \frac{\sum_{j=1}^{J} s_{ij}}{J} \text{ avec } \frac{\sum_{j=1}^{J} e_{ij}}{J} = 0 \text{ et } j = 1 \text{ à } J$$

La précision du test au niveau individuel est représentée par la variance des erreurs sur l'ensemble des répétitions (Novick, 1966).

Lors de la généralisation de cette équation à l'échelle d'une population, le vrai score devient lui aussi une variable aléatoire de variance $var(t_i)$. A ce niveau populationnel, la précision du test est évaluée à l'aide de son indice de fiabilité (« reliability ») défini comme le carré de la corrélation entre les vrais scores et les scores observés des sujets de la population, c'est-à-dire le rapport de la variance du vrai score sur la variance du score observé :

$$\rho_{st}^2 = \frac{var(t_i)}{var(s_i)}$$

Or, si les termes d'erreurs sont supposés indépendants du vrai score des individus, la variance du score observé peut être décomposée en :

$$var(s_i) = var(t_i) + var(e_i)$$

Ainsi, plus la variance de l'erreur dans la population diminue, plus l'indice de fiabilité du test augmente (Mellenbergh, 1996). Comme il est impossible d'avoir accès empiriquement à la valeur de $var(t_i)$, de nombreux travaux ont été menés pour obtenir des estimations

empiriques de cet indice de fiabilité dont le célèbre coefficient alpha de Cronbach (Borsboom, 2005; Cronbach, 1951; T. Kline, 2005).

De nombreuses critiques ont été formulées dans la littérature contre l'utilisation de la CTT (Embretson et Reise, 2000; T. Kline, 2005). En particulier, les hypothèses concernant les termes d'erreur ne permettent pas de traiter les phénomènes d'erreurs systématiques tels que, par exemple, les phénomènes d'apprentissage lors de l'administration répétée d'un test. Une autre critique est que, dans la pratique, le respect des hypothèses de la CTT est rarement discuté dans les études et que, bien souvent, sous cette théorie, la seule justification fournie pour l'utilisation d'un questionnaire est la valeur de son coefficient alpha de Cronbach. Il est pourtant bien connu que la valeur de ce coefficient augmente avec le nombre d'items dans le questionnaire ; plus un questionnaire est long, plus son utilisation devient donc justifiée lorsque la CTT est utilisée. Enfin, un autre argument évoqué contre l'utilisation de la CTT concerne la définition opérationaliste du vrai score qu'elle en donne. Dans la théorie opérationaliste, un terme théorique est synonyme des opérations qui servent à le mesurer. Dans la CTT, le vrai score est défini comme la moyenne des scores observés au test administré de manière répétée à l'individu ; le vrai score est donc défini par les items servant à calculer le score observé. Ainsi, différents questionnaires mesurant, par exemple, le concept « dépression » (Inventaire de dépression de Beck, Echelle de dépression de Hamilton, etc.) mesureraient, selon cette théorie, différents concepts « dépression » définis chacun par les items du questionnaire servant à le mesurer. La mesure d'un même concept par deux questionnaires différents est donc conceptuellement impossible dans cette théorie (Borsboom, 2005).

b. Les modèles à variables latentes

i. Le modèle en facteurs communs et spécifique

Les modèles à variables latentes ont émergé dès le début du XX^{ème} siècle avec les travaux de Charles Spearman portant sur l'analyse factorielle, formalisés par Louis Thurstone en 1947 dans son « modèle en facteurs communs et spécifique » (Brown, 2006; Spearman, 1904). Un facteur, ou variable latente, est un construit hypothétique non-observable ayant la

capacité d'expliquer les associations existant entre des variables observées, les indicateurs. Dans ce modèle, chaque indicateur est une fonction linéaire d'un ou plusieurs facteurs communs et d'un facteur spécifique. Les facteurs communs sont de type continu et influencent plusieurs indicateurs. Le facteur spécifique, continu lui aussi, n'influence qu'une seule variable et est supposé comporter deux parties : une partie spécifique à l'indicateur et une partie due à l'erreur de mesure. L'équation fondamentale du modèle s'écrit de la manière suivante :

$$y_j = \sum_{m=1}^{M} \lambda_{jm} F_m + \varepsilon_j$$

L'indice j représente les items (ou indicateurs, $j = 1$ à p) et l'indice m les facteurs communs ($m = 1$ à M). y_j est le vecteur de longueur N comportant les réponses des N sujets à l'item j. Chaque facteur commun F_m est un vecteur de longueur N comportant les niveaux non-observables des N sujets sur ce facteur. Ils chargent sur chaque item avec un coefficient spécifique à cet item, la charge λ_{jm} dont la valeur est comprise entre 0 et 1, équivalent à un coefficient de corrélation entre l'item j et le facteur m. Le facteur spécifique à chaque item est aussi un vecteur de longueur N représenté par ε_j, indépendant des facteurs communs et des facteurs spécifiques aux autres items et dont les valeurs sont normalement distribuées (Brown, 2006).

L'Analyse Factorielle a d'abord été utilisée à visée Exploratoire (AFE), c'est-à-dire dans le but de déterminer le nombre et la nature des variables latentes permettant d'expliquer les variations et co-variations d'un ensemble de variables observées. Dans ce sens, la variable latente prend sa signification *a posteriori*, c'est-à-dire déduite de l'analyse des données (Bollen, 2002). Par la suite, l'Analyse Factorielle à visée Confirmatoire (AFC) a été développée (Jöreskog, 1969). En AFC, le nombre et la signification des variables latentes sont décidés *a priori* puis, l'adéquation aux données du modèle d'analyse factorielle ainsi décidé est ensuite évaluée.

Alors que dans la grande majorité des questionnaires et tests existants, les dispositifs de modalités de réponse aux items sont de type catégoriel, les variables observées prises en

compte dans le modèle en facteurs communs et spécifique sont supposées être continues. Récemment, certains estimateurs, autres que le classique estimateur du maximum de vraisemblance, ont donc été développés afin de permettre son application sur des données de type catégorielles (Brown, 2006; Muthén et Muthén, 2012). Cependant, bien avant cela, vers la moitié du XX^{ème} siècle, un autre cadre conceptuel s'appliquant au cas des variables observées de type catégorielles et mettant toujours en jeu des variables latentes, a été imaginé par des chercheurs en sciences de l'éducation.

ii. Les modèles issus de la théorie de réponse à l'item

Les modèles issus de la théorie de réponse à l'item (IRT : Item Response Theory) ont d'abord été développés pour des items binaires. Ils modélisent la probabilité que le sujet i ($i = 1$ à N) réponde positivement à l'item j ($j = 1$ à p). Dans cette théorie, la variable latente est continue, est nommée « trait latent » (TL) et est habituellement représentée par la lettre grecque θ. Le modèle logistique à deux paramètres d'item (2PLM : two parameters logistic model) est très souvent utilisé et s'écrit, dans le cas d'une variable binaire Y dont la modalité positive est codée 1 :

$$P(Y_{ij} = 1/\theta_i, \nu_j, \delta_j) = \frac{\exp\left[\nu_j(\theta_i - \delta_j)\right]}{1 + \exp\left[\nu_j(\theta_i - \delta_j)\right]}$$

où $\theta_i \sim \mathcal{N}(0, \sigma_\theta^2)$est le niveau de l'individu sur le TL et les deux paramètres spécifiques d'item sont :

- δ_j : le coefficient de difficulté de l'item j représentant le niveau sur le TL qu'un sujet doit atteindre pour avoir une probabilité de 50% de répondre positivement à l'item
- ν_j : le coefficient de discrimination de l'item j représentant sa capacité à différencier deux individus ayant un niveau sur le TL proche de la valeur δ_j

Le modèle à un paramètre d'item (1PLM), le plus connu, considère un pouvoir discriminant identique (égal à 1 pour le modèle de Rasch) pour tous les items de l'échelle ($\nu_j = constante$ $\forall\, j$). Des extensions au 1PLM (Partial Credit Model, Rating Scale Model, etc.) et au 2PLM

(Graded Response Model, Modified Graded Response Model) ont été développées pour l'application aux variables observées nominales ou ordinales (Embretson et Reise 2000).

Ces modèles IRT reposent sur trois hypothèses fondamentales : l'indépendance locale (les réponses aux items sont indépendantes conditionnellement au TL), l'unidimensionnalité (une seule variable latente est suffisante pour expliquer la variance commune entre les réponses aux items), la monotonicité (la probabilité d'une réponse non-nulle à l'item augmente avec le TL) (Embretson et Reise, 2000). Le respect de ces trois hypothèses a longtemps été un frein à l'utilisation des modèles IRT. De nombreux cas de non-respect de l'hypothèse d'indépendance locale peuvent être retrouvés dans les questionnaires existants. Par exemple, la sous-échelle activité physique du très utilisé questionnaire de qualité de vie MOS-SF36 (Medical Outcomes Study, Short Form, 36 items) contient plusieurs items ayant montré une dépendance locale (i.a. le sujet doit indiquer s'il est limité en raison de son état de santé actuel pour « Monter plusieurs étages par l'escalier » puis pour « Monter un étage par l'escalier ») (Horton et Tennant, 2011; Leplège *et al.*, 2001; Ware et Gandek, 1998; Ware et Sherbourne, 1992). De même, il n'est pas rare, le questionnaire MOS-SF36 en est aussi un exemple, que les instruments de mesure subjective existants évaluent un concept multidimensionnel. Bien que des extensions des modèles IRT aient été développées pour le cas des concepts multidimensionnels, l'attitude la plus fréquente dans ce cas est d'appliquer un modèle IRT à chacune des dimensions du questionnaire (Embretson et Reise, 2000). Enfin, un autre frein à l'utilisation des modèles IRT est la complexité calculatoire qu'implique la résolution des p équations, correspondant à chacun des p items du questionnaire, pour l'estimation des différents paramètres d'items et individuels. Les récents progrès en matière d'informatique et d'électronique ont largement diminué ce frein et les modèles IRT sont maintenant de plus en plus utilisés dans le domaine de la santé (Hays et Lipscomb, 2007; Hays *et al.*, 2000; Revicki et Cella, 1997)

iii. Les modèles d'analyse en classe latente

Que ce soit dans le modèle en facteurs communs et spécifique ou dans les modèles IRT, la variable latente est une variable continue représentant donc un construit hypothétique, par exemple la dépression, sous la forme d'un continuum allant de $-\infty$ à $+\infty$ sur lequel

peuvent être ordonnés les sujets, en fonction de leur niveau de dépression par exemple. Cette représentation du construit hypothétique n'est cependant pas la seule utilisée en pratique. Dans le cas de la dépression, en plus du niveau d'intensité, la pratique clinique distingue bien souvent des sous-types, « atypique », « mélancolique », « sévère », etc., sous-entendant ainsi une représentation catégorielle de ce concept. Venant du domaine de la sociologie, des modèles à variable latente discrète ont aussi été développés parallèlement aux modèles à variable latente continue au cours du XX^{ème} siècle (Bartholomew *et al.*, 2011). Leur but est de permettre la détermination de sous-groupes de sujets (les classes) dans un échantillon à l'aide de variables observées. L'hypothèse est qu'il existe un nombre adéquat de catégories (classes) pour la variable latente permettant d'obtenir l'indépendance conditionnelle parmi les variables observées (Muthén, 2002).

Par exemple, dans le cas de p variables observées dichotomiques Y_j, $(j = 1 \text{ à } p)$ et d'une variable latente catégorielle C à c classes latentes, la probabilité qu'un individu de la classe latente k ait la variable Y_j égale à 1 $(k = 1 \text{ à } c)$ est modélisée sous une forme logistique et est notée $P(Y_j = 1/C = k)$. Soit $P(C = k)$, la probabilité *a priori* d'appartenir à la classe latente k, l'hypothèse d'indépendance conditionnelle donne la probabilité jointe suivante :

$$P(Y_1, Y_2, \dots, Y_p) = \sum_{k=1}^{c} P(C = k) \prod_{\mathrm{j}=1}^{\mathrm{p}} P(Y_j = 1/C = k)$$

Ainsi, la probabilité *a posteriori* qu'un individu ayant le vecteur de variables observées $(Y_1, Y_2, \dots, Y_p)$ appartienne à la classe latente k s'obtient à l'aide du théorème de Bayes :

$$P(C = k/Y_1, Y_2, \dots, Y_p) = \frac{P(C = k) \prod_{\mathrm{j}=1}^{\mathrm{p}} P(Y_j = 1/C = k)}{P(Y_1, Y_2, \dots, Y_p)}$$

Chaque individu de l'échantillon peut donc être ainsi affecté à la classe latente la plus probable en fonction de son vecteur de variables observées (Droesbeke *et al.*, 2005; Muthén, 2002).

De la même manière que dans le cas d'une variable latente continue, le modèle d'analyse en classe latente peut être étendu aux variables observées polytomiques nominales

ou ordinales. Lorsque les variables observées sont de type continu, ce modèle est appelé modèle d'analyse en profil latent ou modèle de mélange (Borsboom, 2008; Goodman, 1974; Muthén, 2002). Ces modèles ne sont pas utilisés en pratique comme cadre conceptuel pour l'établissement d'un modèle de mesure lors du développement d'un questionnaire, cependant certains auteurs ont récemment envisagé un cadre global regroupant l'ensemble des modèles utilisant des variables latentes (Bartholomew *et al.*, 2011; Borsboom, 2008, 2005; Moustaki et Knott, 2000; Muthén, 2002; Muthén et Muthén, 2012).

c. Les modèles linéaires généralisés à variables latentes

Les différents modèles présentés dans la partie 1.b de ce chapitre ont leurs origines dans différentes disciplines, la psychologie pour l'analyse factorielle, l'éducation pour l'IRT, la sociologie pour les analyses en classes latentes. Pour cette raison, leur développement s'est fait parallèlement dans chacune de ces disciplines au cours du XX^{ème} siècle et l'idée et la formalisation de leur unification dans le cadre des modèles linéaires généralisés (GLM : Generalized Linear Model) pour variables latentes n'est apparue que progressivement vers la fin de ce siècle (Bartholomew *et al.*, 2011).

i. La formalisation mathématique

Comme le sont les modèles linéaires généralisés ne comportant que des variables observées, les GLM à variables latentes ont été définis mathématiquement à l'aide des trois composantes classiques (Moustaki et Knott, 2000; Nelder et Wedderburn, 1972) :

- La composante aléatoire dans laquelle chacune des p variables observées Y_j, ($j = 1$ à p) provient d'une distribution dont la fonction de densité appartient à la famille exponentielle de la forme

$$f_j(Y_j; \theta_j, \phi_j) = \exp\left(\frac{Y_j\theta_j - b_j(\theta_j)}{\phi_j} + d_j(Y_j, \phi_j)\right)$$

  où $b_j(\theta_j)$ et $d_j(Y_j, \phi_j)$ sont des fonctions spécifiques prenant une forme dépendante de la distribution de Y_j, θ_j est le paramètre canonique de la distribution exponentielle et $\phi_j$est un paramètre d'échelle ;

- La composante systématique dans laquelle les M covariables sont des variables latentes notées F_m, $(m = 1 \text{ à } M)$ et produisent un prédicteur linéaire $\eta_j$correspondant à chaque Y_j ;

$$\eta_j = \alpha_{j0} + \sum_{m=1}^{M} \alpha_{jm} F_m$$

- La fonction de lien reliant l'espérance μ_j de Y_j au prédicteur linéaire par la relation

$$\eta_j = \upsilon_j(\mu_j(\boldsymbol{F}))$$

où $\upsilon_j(.)$ est une fonction monotone différentiable et peut être différente pour différentes variables observées Y_j.

Ainsi, lorsque les variables observées Y_j sont distribuées normalement, la fonction de lien $\upsilon_j(.)$ est l'identité et l'équation prend la forme de l'équation fondamentale du modèle en facteurs communs et spécifique. Lorsque les variables observées Y_j sont binaires, prenant la forme d'une distribution de Bernouilli de paramètre $\pi_j(\boldsymbol{F})$ où $m = 1$ (unidimensionalité), la fonction de lien $\upsilon_j(.)$ est logistique et l'équation prend la forme de l'équation d'un modèle IRT 2PLM.

Cet effort pour fournir un cadre global regroupant l'ensemble des différents modèles à variable latente développés parallèlement au cours du XX^{ème} siècle a permis de formaliser les liens existant entre les modèles utilisés en analyse factorielle, les modèles IRT et aussi les modèles à variables latentes discrètes qui sont prises en charge dans ces GLM sous la forme de variables latentes indicatrices (Bartholomew *et al.*, 2011; Borsboom, 2005; Muthén, 2002). Un autre intérêt notable est la flexibilité qu'il offre pour l'intégration simultanée dans un même modèle de mesure de variables observées continues, de comptage, binaires ou polytomiques (Moustaki et Knott, 2000). Enfin, le statut de la variable latente (facteur ou trait latent ou classe latente) y est homogénéisé. Contrairement au vrai score dans la CTT défini par les items servant à mesurer le score observé, la variable latente est une entité définie *a priori* dans un modèle théorique dont les paramètres, reliant la variable latente aux items, seront estimés à partir des données observées. L'adéquation de ce modèle théorique aux données pourra être évaluée et, en cas de mauvaise adéquation, le modèle sera rejeté indiquant ainsi que les items

utilisés ne sont pas les indicateurs adéquats pour évaluer la variable latente telle que définie dans ce modèle théorique. Par ailleurs, si l'adéquation est bonne, il est possible d'affirmer que si le niveau sur la variable latente est évalué à partir de 10 items dans le modèle théorique, il sera aussi possible de l'estimer, avec une moins bonne précision, à partir d'un sous-ensemble de ces 10 items. Deux questionnaires différents peuvent donc être utilisés pour évaluer une même variable latente dans cette théorie (Borsboom, 2005).

ii. Représentation graphique

Les modèles à variables latentes peuvent être représentés graphiquement à l'aide de symboles communs à l'ensemble de ces modèles. Ces représentations graphiques sont appelées diagrammes de chemin (path diagrams) et les symboles utilisés sont représentés dans la figure 1. Représenter un modèle à variables latentes à l'aide de ces symboles est équivalent à son écriture sous forme mathématique. Un exemple simple de transcription d'un diagramme de chemin en formulation mathématique est fourni dans la figure 2.

Les variables observées sont symbolisées par des rectangles, les variables latentes par des ellipses. Un effet supposé causal d'une variable sur une autre est symbolisé par une flèche droite unidirectionnelle partant de la variable « cause » et se dirigeant vers la variable « effet ». Toute variable ayant une flèche unidirectionnelle pointant vers elle est appelée variable endogène (variable dépendante, ayant une cause représentée dans le diagramme). Deux variables corrélées sans qu'il y ait de relation causale supposée entre les deux sont reliées par une flèche courbe bidirectionnelle. Seules les variables exogènes (variables indépendantes, sans cause représentée dans le diagramme) sont pointées par ce type de flèche bidirectionnelle. Enfin, les résidus (erreurs de mesure) sont représentés soit par une ellipse (variable latente) avec une flèche pointant vers une (et une seule) variable endogène, soit seulement par une flèche pointant sur une variable endogène, l'ellipse étant omise. La variance des variables exogènes (dont les résidus) peut être symbolisée par une flèche courbe bidirectionnelle pointant de chaque côté sur la même variable (R. B. Kline, 2005).

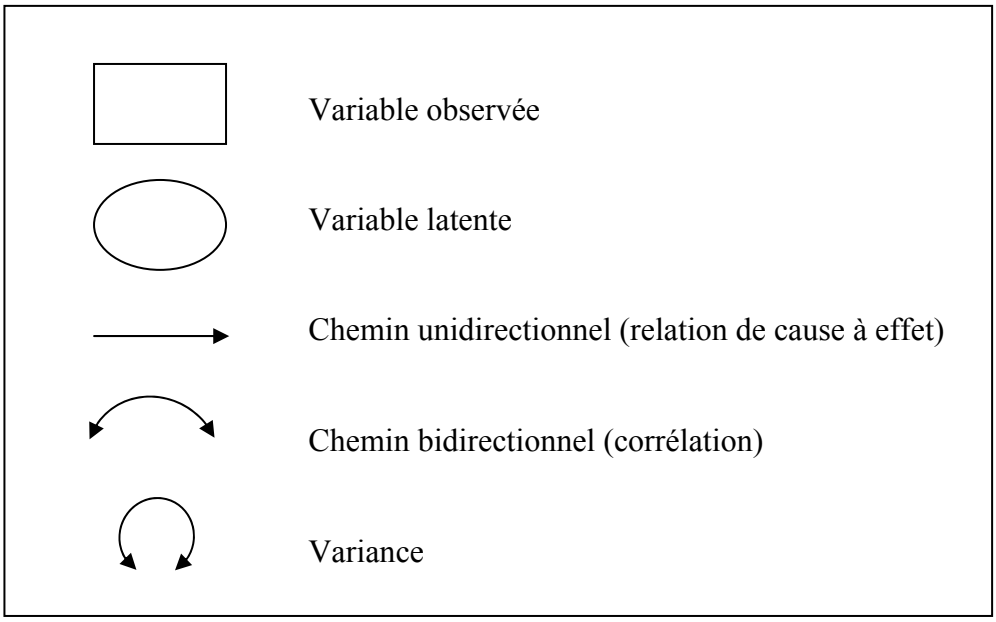
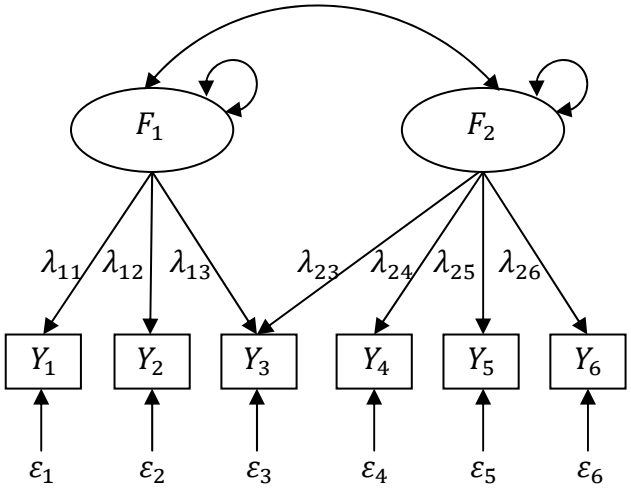


Figure 1 : Symboles utilisés dans les diagrammes de chemin (path diagrams)



$$\begin{cases} Y_1 = \alpha_1 + \lambda_{11}F_1 + \varepsilon_1 \\ Y_2 = \alpha_2 + \lambda_{12}F_1 + \varepsilon_2 \\ Y_3 = \alpha_3 + \lambda_{13}F_1 + \lambda_{23}F_2 + \varepsilon_3 \\ Y_4 = \alpha_4 + \lambda_{24}F_2 + \varepsilon_4 \\ Y_5 = \alpha_5 + \lambda_{25}F_2 + \varepsilon_5 \\ Y_6 = \alpha_6 + \lambda_{26}F_2 + \varepsilon_6 \end{cases}$$

avec $\varepsilon_j \sim \mathcal{N}(0, \sigma_j),\ j = 1 \text{ à } 6$

et $\boldsymbol{F} = \begin{pmatrix} F_1 \\ F_2 \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & \psi_{12} \\ \psi_{12} & 1 \end{pmatrix}\right), m = 1 \text{ à } 2$

Figure 2 : Exemple de diagramme de chemin et de transcription mathématique du modèle en facteurs communs (F) et spécifique (ε) pour une échelle bidimensionnelle à six items (Y). Les charges factorielles sont notées λ et les constantes α.

2. Questions de recherche

A l'occasion de ce travail de thèse, trois axes, caractéristiques des études épidémiologiques, ont été choisis afin d'explorer les problèmes méthodologiques potentiellement soulevés par l'utilisation des instruments de mesure subjective dans cette discipline.

a. La phase de validation de l'instrument de mesure

Quels que soient les instruments utilisés pour mesurer les variables dépendantes ou indépendantes dans son étude, un des premiers reflexes de l'épidémiologiste est de vérifier la validité de la mesure obtenue dans sa population cible. Dans le cas des mesures subjectives, cette phase de validation concerne différents types de validité (Mokkink *et al.*, 2010). Une première catégorisation en « validité intra-concept » et « validité inter-concept » peut aider à classer ces différents types. La validité intra-concept concerne tout ce qui est relié à la définition du concept à mesurer (théorie définitoire) ; elle regroupe, par exemple, la validité de contenu (bonne représentation du concept mesuré dans le contenu de l'instrument, les items), la validité interne (bonne homogénéité des items de l'échelle pour mesurer le même concept) et la validité convergente (bonne cohérence entre les résultats recueillis chez le sujet, un proche ou un soignant). La validité inter-concept concerne l'ensemble des relations existant entre les différents concepts d'une même discipline (théorie nomologique) ; elle regroupe, par exemple, la validité divergente (faible corrélation avec des concepts différents), la validité concourante (forte corrélation avec une mesure de référence, un gold standard) et la validité prédictive (si le gold standard ne peut être obtenu qu'ultérieurement). Enfin, englobant ces différents types de validité, la validité de face est l'impression globale que l'instrument mesure bien ce qu'il est censé mesurer (Falissard, 2008; Terwee *et al.*, 2007).

Après la phase de construction ou de traduction d'un questionnaire où la validité de contenu peut être évaluée, la première étape de cette phase de validation est bien souvent la planification d'une étude pour évaluer, au minimum, la validité interne du questionnaire. Pour cela, l'évaluation de la validité structurelle est planifiée, le plus souvent par analyse factorielle pour déterminer la structure dimensionnelle de l'instrument. L'évaluation de la consistance

interne par le calcul du coefficient alpha de Cronbach fait aussi partie de la validité interne et est aussi habituellement évaluée à cette occasion. Une question inévitable lors de la mise en place de telles études est celle de la taille d'échantillon nécessaire. Pour la validité interne, cette taille doit permettre l'obtention d'une précision suffisante du coefficient alpha de Cronbach et des charges λ_{jm} estimées lors de l'analyse factorielle afin d'approcher au mieux la structure dimensionnelle de l'instrument. Cependant, contrairement à ce que l'on observe d'habitude en recherche biomédicale, il n'existe pas dans la littérature de règle simple de calcul du nombre de sujets nécessaire (NSN) basée sur des fondements théoriques rigoureux. C'est néanmoins un point essentiel de leur planification ; un effectif inadéquat pouvant mener à des conclusions erronées quant à la structure dimensionnelle ou à la consistance interne de l'instrument et ainsi conduire à des biais d'information lors de son utilisation ultérieure dans des études épidémiologiques.

La première question de recherche de ce travail de thèse concernera la détermination du nombre de sujets suffisant pour pouvoir conclure quant à la validité interne d'un instrument de mesure subjective. Elle sera traitée, dans le cadre des échelles de mesure en psychiatrie, dans l'étude présentée au chapitre 2.

b. Les propriétés de la mesure obtenue

La comparaison est un des actes essentiels de l'épidémiologie étiologique : comparaison de moyennes ou de pourcentages entre deux groupes, comparaison d'une grandeur chez un même sujet au cours du temps, etc. (Rumeau-Rouquette *et al.*, 1986). Le test statistique permettant cette comparaison prend différentes formes en fonction du type de mesure en question. Stevens en 1946 a établi une classification des mesures dont l'idée initiale était de préciser les opérations et méthodes statistiques applicables à chaque niveau de mesure (Stevens, 1946). Quatre niveaux ont été définis : les *échelles nominales* où les codes assignés à chaque catégorie de réponse ne sont que des étiquettes sans signification numérique ; les *échelles ordinales* où les codes assignés permettent d'ordonner les catégories de réponse les

unes par rapport aux autres ; les *échelles d'intervalle*, non bornées, dont le zéro est arbitrairement fixé et où il est possible d'affirmer que l'écart existant entre deux unités adjacentes est le même quel que soit leur niveau sur l'échelle ; enfin, les *échelles de ratio* où il est possible d'affirmer que le rapport qui existe entre les valeurs 2 et 1, par exemple, est le même que celui qui existe entre les valeurs 4 et 2 ou 8 et 4, etc. Bien que cette affirmation ait été remise en question par la suite, selon Stevens, le calcul de moyennes, d'écart-types et l'application de procédures statistiques paramétriques nécessite au minimum le niveau de mesure d'intervalle (Gaito, 1980; Michell, 1986; Stevens, 1946; Townsend et Ashby, 1984; Velleman et Wilkinson, 1993).

Dans le domaine des mesures subjectives, si le modèle de mesure est issu de la CTT, le score observé (simple somme ou somme pondérée des réponses aux items du questionnaire) est utilisé comme approximation du vrai score du sujet. Si la propriété d'échelle d'intervalle n'est pas atteinte au niveau de chaque item individuellement, la pondération des items a pour but de rapprocher le score observé du niveau d'échelle d'intervalle. Cependant, cette propriété est difficile à démontrer de manière indiscutable que ce soit au niveau de l'item ou au niveau du score (Falissard, 2008). En revanche, si le modèle de mesure est issu de l'IRT, la mesure obtenue est supposée être d'intervalle car l'échelle du TL, dans cette théorie, a le niveau d'échelle d'intervalle (Embretson et Reise, 2000).

La deuxième question de recherche de ce travail de thèse s'intéressera aux performances des procédures statistiques paramétriques appliquées sur le score observé issu de la CTT *versus* sur une mesure issue d'un modèle de mesure développé à partir d'un modèle à variable latente. Elle sera traitée, dans le cadre de l'évaluation de la sensibilité au changement d'un instrument de mesure subjective, dans l'étude présentée au chapitre 3.

c. Utilisation dans un schéma longitudinal

Le schéma longitudinal est le schéma privilégié en épidémiologie pour de nombreuses raisons dont la principale est l'opportunité qu'il offre de mettre en évidence la séquence

temporelle « La cause précède l'effet », chère à Bradford Hill dans ses critères de causalité (Hill, 1965). Cependant, l'utilisation d'instruments de mesure subjective peut poser plusieurs types de problèmes dans un tel schéma. Par exemple, lors de la mise en place d'une étude de cohorte, le choix des questionnaires validés les plus adaptés aux phénomènes subjectifs à mesurer doit être fait dès la phase de planification (Rothman *et al.*, 2008). Au cours du suivi, en particulier lorsqu'il dure plusieurs dizaines d'années, certains questionnaires peuvent subir des améliorations, voire être substitués par un autre questionnaire considéré comme plus adapté que celui qui avait été choisi initialement. Dans cette situation, les items disponibles pour évaluer le même phénomène aux différents temps de collecte ne sont pas tous identiques.

Cette problématique est particulièrement courante dans le domaine de l'épidémiologie développementale. Cette approche intègre les principes, théories et méthodes de la psychologie développementale dans la recherche épidémiologique afin d'éclaircir les mécanismes selon lesquels les processus développementaux affectent les risques de survenue de problèmes de santé (Costello *et al.*, 2006; Pillemer et White, 2005). L'attention est donc portée sur l'analyse de données issues de cohortes suivies au cours de stades du développement tels que la petite enfance, l'enfance, l'adolescence, etc. L'évaluation des phénomènes subjectifs doit donc être adaptée au cours du suivi et tous les items utilisés ne sont pas identiques à chaque stade développemental.

Dans de tels cas, si la CTT est utilisée, le score observé n'est pas comparable d'un temps à l'autre car il n'est pas calculé sur les mêmes items ; sauf si, pratique habituelle pour contourner ce problème, seuls les items présents à tous les temps de collecte sont utilisés pour

calculer le score observé (i.e. les items absents[2] à au moins un temps de collecte ne sont pas utilisés pour ce calcul). Cette pratique, bien que rendant possible la comparaison du phénomène étudié entre chaque temps, est responsable d'une perte d'information étant donné que les réponses des sujets à certains items sont tout simplement mises de côté. En découle une perte de précision à la fois de l'estimation du phénomène étudié et de l'ensemble des paramètres estimés dans les modèles statistiques l'utilisant. La validité des inférences statistiques issues de ce type d'études est donc affaiblie.

Dans ce type de situation, la possibilité qu'offrent les modèles à variables latentes de mesurer un même phénomène à partir d'échantillons d'items différents trouve tout son intérêt. Par exemple, dans le modèle de Rasch, une des propriétés les plus importantes est l'objectivité spécifique, c'est-à-dire que l'échelle du TL est identique quelles que soient la population étudiée ou les conditions de la mesure (Embretson et Reise, 2000; Rupp et Zumbo, 2006). Autrement dit, l'erreur d'échantillonnage mise à part, les estimations des paramètres d'items obtenues dans différents échantillons ou à différentes occasions de mesure sont identiques. De même, l'erreur de mesure mise à part, les estimations des niveaux sur le TL des individus obtenues à partir de groupes d'items différents sont identiques (Embretson et Reise, 2000; Hambleton *et al.*, 1991).

La troisième question de recherche de ce travail de thèse concernera les performances des analyses statistiques étudiant l'évolution d'un phénomène subjectif au cours du temps mesuré à l'aide de la CTT *versus* un modèle de mesure issu de l'IRT, en particulier lorsque le

[2] Une distinction sera faite tout au long de ce texte entre l'expression « item absent » désignant un item qui n'a pas été posé à un temps de collecte, i.e. aucun des sujets de l'échantillon n'a eu la possibilité de répondre à cet item car il ne faisait pas partie du questionnaire, et l'expression « items/donnée/réponse manquant(e) » désignant un item posé (faisant partie du questionnaire) à tous les sujets de l'échantillon mais pour lequel certains sujets n'ont pas donné de réponse.

groupe d'items disponibles pour la mesure diffère à chaque temps de collecte. Elle sera traitée dans l'étude présentée au chapitre 4, dans le cadre de la détection de groupes de sujets ayant des trajectoires homogènes d'évolution du phénomène subjectif étudié au cours du temps par analyse en classe latente de trajectoires (LCGA : Latent Class Growth Analysis), analyses très fréquemment utilisées en épidémiologie développementale (Berlin *et al.*, 2014; Jung et Wickrama, 2008; Nagin, 2005).

3. Matériels et méthodes

a. Matériels et considérations d'ordre éthique

Dans ce travail de thèse, deux types de données ont été utilisées : des données simulées par ordinateur et des données réelles. Le Comité d'Ethique de la Recherche En Santé de l'université de Montréal a donné son approbation pour ce travail de recherche.

i. Données artificielles

La première et la troisième étude sont des études par simulation. Ces études sont de plus en plus utilisées dans la littérature médicale. Leur principe repose sur la génération de bases de données artificielles selon un modèle de simulation dont les paramètres peuvent être contrôlés. L'intérêt est donc que « la vérité » (les paramètres du modèle sous-jacent) est connue. Le but est d'analyser ces bases de données et de comparer les résultats obtenus avec « la vérité » connue. En faisant varier les paramètres du modèle de simulation (taille de l'échantillon, nombre d'items, paramètres de difficultés d'items, etc.), il est donc possible d'évaluer leur influence sur les performances de l'analyse pour retrouver « la vérité » (Burton *et al.*, 2006). Par exemple, dans la première étude, le paramètre d'intérêt à faire varier dans le modèle de simulation est la taille des échantillons simulés pour pouvoir étudier son influence sur les performances de l'analyse factorielle et sur la précision du coefficient alpha de Cronbach.

ii. Données réelles

La deuxième étude repose sur l'analyse de données réelles issues d'une cohorte multicentrique (centres hospitaliers universitaires de Nancy, Besançon et Metz en France) de 1520 patients âgés de moins de 75 ans, hospitalisés en milieu médical et chirurgical entre octobre 2008 et septembre 2010 pour pathologie chronique dans les domaines cardio-vasculaires, locomoteur, nephro-urologique, digestif, pneumologique et endocrinologique. Chaque participant ou son représentant légal a donné son consentement libre et éclairé à l'inclusion puis à chaque temps de suivi. Le comité d'éthique de Lorraine en France a donné son approbation pour cette étude.

b. Méthodes

La méthodologie utilisée dans chacune des trois études présentées est détaillée dans les chapitres correspondants. Lors de la discussion générale, les conséquences de l'utilisation des mesures subjectives en épidémiologie en termes de biais de mesure seront examinées à l'aide de graphes acycliques orientés (DAG : Directed Acyclic Graph).

i. Les graphes acycliques orientés

Ces graphes sont un outil d'analyse graphique des liens causaux potentiels entre les variables d'une étude, sans qu'il y ait besoin d'hypothèse sur les paramètres de la distribution des variables ou la forme des liens entre chaque variable (linéarité par exemple) (Greenland *et al.*, 1999). Ils ont beaucoup en commun avec la représentation graphique utilisée dans les modèles à variables latentes (cf. paragraphe 1.c.ii de ce même chapitre), en dehors de :

- La représentation des variables : elles sont habituellement représentées par leur nom sans autre symbolisme
- La représentation des liens : seule la flèche unidirectionnelle représentant un lien causal est utilisée (d'où le terme « orienté » dans graphe acyclique orienté).

Aucune flèche bidirectionnelle n'est utilisée ce qui implique que deux variables exogènes sont toujours supposées indépendantes. Les corrélations entre deux variables sont interprétées comme la présence d'au moins une cause commune inconnue entre les deux variables et sont

représentées comme dans la figure 3. Enfin, un diagramme causal est dit « acyclique » lorsqu'il ne contient pas de boucle de rétroaction (« feedback loop »), c'est-à-dire que si une variable X est la cause de la variable Y, la variable Y ne peut en aucun cas être la cause de la variable X au même moment. L'hypothèse d'une telle rétroaction peut-être posée mais nécessitera une représentation de type longitudinale : la valeur de X au temps 1 (X_1) cause la valeur de Y au temps 2 (Y_2) qui cause la valeur de X au temps 3 (X_3), etc. (Rothman *et al.*, 2008).

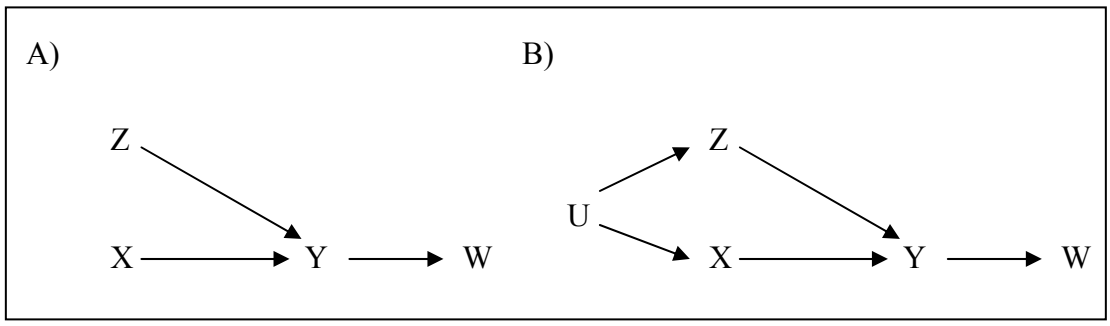


Figure 3 : Graphe acyclique orienté représentant une séquence causale où la variable Y causée par X et Z, est la cause de la variable W. Dans le graphe A), X et Z sont des variables exogènes indépendantes l'une de l'autre. Dans le graphe B), X et Z sont corrélées, i.e. ce sont des variables endogènes ayant une cause commune inconnue U

Les DAG sont de plus en plus employés en épidémiologie car ils permettent de reconnaître et d'éviter des erreurs de raisonnement en analyse causale et de repérer la présence de biais potentiels qu'ils soient de sélection, de confusion ou d'information (Greenland *et al.*, 1999; Hernán et Cole, 2009; Hernán *et al.*, 2004, 2002; Rothman *et al.*, 2008).

ii. Les biais d'information et leur représentation en DAG

Lors de l'évaluation d'un phénomène à l'aide d'un instrument de mesure, deux types d'erreur de mesure peuvent survenir : l'erreur aléatoire (dont la moyenne tend asymptotiquement vers zéro) et l'erreur systématique (dont la moyenne tend asymptotiquement vers une valeur non nulle). Les biais d'information surviennent lors de la présence d'une erreur systématique de mesure (Rothman *et al.*, 2008). Si l'erreur systématique faite sur la mesure de d'une variable X dépend de la « vraie valeur » (sans erreur) de la variable Y, le biais d'information sur l'estimation du lien entre les variables X et Y est qualifié

de « différentiel » ; sinon, il est qualifié de « non-différentiel ». Si l'erreur systématique faite sur la mesure de la variable X dépend de l'erreur systématique faite sur la mesure de la variable Y, elle est qualifiée d'erreur systématique « dépendante » ou « corrélée », sinon elle est qualifiée d' « indépendante ». La représentation de ces quatre grands types de biais à l'aide de DAG a nécessité l'introduction d'une distinction entre la « vraie valeur » d'une variable et sa valeur observée, indiquée par un astérisque (Hernán et Cole, 2009). A l'image de la représentation utilisée dans les modèles à variables latentes, un terme d'erreur ε (représentant les facteurs responsables d'erreur de mesure, autres que les variables représentées) ayant un effet sur la valeur de la variable observée est symbolisé (VanderWeele et Hernán, 2012).

Dans la figure 4, sont représentés les quatre grands types de biais d'information pouvant survenir lors de l'estimation de l'effet d'une variable X sur une variable Y. Le choix a été fait dans cette représentation de poser l'hypothèse que X et Y sont indépendantes (hypothèse nulle) afin que le biais potentiellement créé par l'erreur de mesure soit mieux mis en valeur visuellement sur ces schémas. Dans chaque cas, un effet de la vraie valeur de la variable sur son erreur de mesure a été symbolisé car cela est quasi-systématiquement le cas, en particulier pour des valeurs extrêmes (VanderWeele et Hernán, 2012). Le premier type de biais représenté (A) est non-différentiel avec erreurs systématiques indépendantes. Dans ce cas, si une association significative est retrouvée entre X* et Y* alors la conclusion pourra être qu'il existe une association entre X et Y. Le deuxième type de biais représenté (B) est toujours non-différentiel mais avec erreurs sytématiques corrélées. Ceci peut se présenter, par exemple dans une étude rétrospective où X et Y seraient auto-rapportés et où un biais de mémoire affecterait leur mesure. Les deux derniers cas présentés dans la figure 4 sont des exemples de biais différentiels avec erreurs systématiques indépendantes (C) ou corrélées (D). Dans ces trois derniers cas, une association significative entre X* et Y* pourrait être mise en évidence même en cas d'absence d'association entre X et Y. Dans le cas C, qui pourrait, par exemple, correspondre au cas d'une étude où la recherche de la présence de la maladie Y serait effectuée sans aveugle concernant l'exposition X de chacun des sujets, une association entre X* et Y* sera systématiquement retrouvée. Enfin, dans le cas D, la présence d'une association

entre X* et Y* dépendra du signe des effets de U et Y sur les erreurs systématiques qui pourraient se compenser (Hernán et Cole, 2009; VanderWeele et Hernán, 2012).

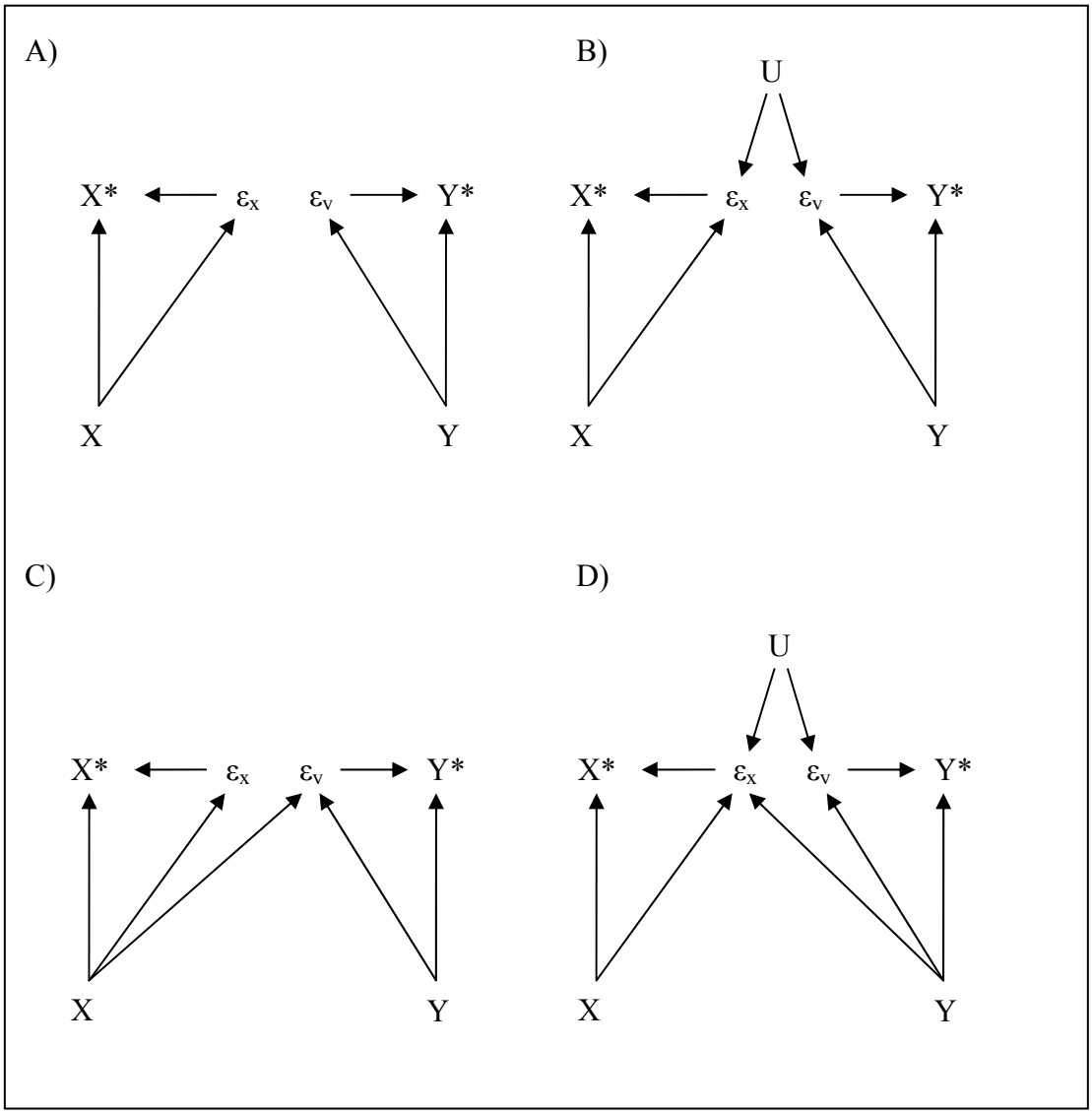


Figure 4 : Les quatre grands types de biais d'information représentés à l'aide de graphes orientés acycliques : A) biais non-différentiel, erreurs systématiques indépendantes, B) biais non-différentiel, erreurs systématiques corrélées, C) biais différentiel, erreurs systématiques indépendantes, D) biais différentiel, erreurs systématiques corrélées. X et Y sont les vraies valeurs des deux variables d'intérêt, X* et Y* sont les mesures observées de X et Y, ε_x et ε_y sont les erreurs faites sur la mesure de X et Y, U est l'ensemble des causes communes à ε_x et ε_y (adapté de la figure 2 dans Hernan et Cole, 2009 et de la figure 1 de l'appendice dans VanderWeele et Hernan, 2012 avec l'autorisation des éditeurs)

CHAPITRE 2 : NOMBRE DE SUJETS NECESSAIRES POUR LA VALIDATION INTERNE D'ECHELLES EN PSYCHIATRIE

Traduction de l'article suivant, présenté en langue originale en annexe 1 :

Rouquette A, Falissard B. (2011) Sample Size Requirements for the Internal Validation of Psychiatric Scales. *International Journal of Methods in Psychiatric Research.* Vol. 20(4): 235-249

Contribution des co-auteurs :

Alexandra Rouquette : Conception de l'étude, revue de la littérature, programmation des analyses statistiques, interprétation des résultats, rédaction, soumission et révision de l'article.

Bruno Falissard : Conception de l'étude, contribution à la programmation des analyses statistiques et à l'interprétation des résultats, révision de l'article.

Résumé:

La règle du ratio « Nombre de sujets à inclure sur nombre d'items dans l'échelle (N/p)» utilisée pour le calcul du nombre de sujets nécessaire (NSN) dans les études de validation interne de questionnaire a été énoncée dans la littérature sans justification rigoureuse. L'objectif de ce travail était de développer un outil de détermination du NSN pour ces études, dans le domaine de la psychiatrie. Une revue de la littérature a d'abord été effectuée afin de déterminer les particularités des échelles utilisées dans cette discipline. Ensuite, deux études de simulations ont été menées utilisant 1/ un modèle de génération de données développé selon le modèle de structure dimensionnelle retrouvé par la revue de la littérature, 2/ une base de données réelles. La qualité des solutions obtenues sur des échantillons de tailles différentes par analyse en composantes principales ou Analyse Factorielle Exploratoire (AFE) a ainsi pu être étudiée. Enfin, l'influence de l'effectif sur la précision du coefficient alpha de Cronbach a été examinée. La règle du ratio *N/p* ne trouve aucune justification dans cette étude : la validation d'échelles constituées d'un faible nombre d'items ne permet pas d'inclure moins de sujets. Un effectif minimum de 300 sujets doit être la règle et doit être augmenté si le nombre attendu de facteurs dans l'échelle est grand, si l'AFE est utilisée et si le nombre d'items est petit.

Mots-clefs: Taille d'échantillon, études de validation, analyse factorielle, questionnaire, psychiatrie

1. Introduction

Une étape inévitable lors de la mise en place d'un projet de recherche quantitative, en particulier lors de la planification des analyses statistiques, est celle de la détermination du nombre de sujets à inclure. Dans une étude analytique, la taille de l'échantillon est déterminée par la puissance souhaitée pour le test d'hypothèse. Dans les études descriptives, cette approche ne peut pas être utilisée et c'est la précision souhaitée pour l'estimation du paramètre d'intérêt qui sert pour le calcul du nombre de sujets nécessaire.

Les études de validation interne d'échelles de mesure sont des études descriptives où deux types de paramètres sont habituellement estimés : le coefficient alpha de Cronbach qui évalue la consistance interne et les charges factorielles, généralement estimées lors de l'étude de la structure dimensionnelle de l'instrument par Analyse en Composantes Principales (ACP) ou Analyse Factorielle Exploratoire (AFE). Pour le coefficient alpha de Cronbach, une formule de l'intervalle de confiance a été proposée par Feldt dans les années 1960 (Fan et Thompson, 2001; Feldt, 1965). Le nombre de sujets nécessaire à l'estimation de ce coefficient avec une précision fixée peut donc être facilement calculé. La difficulté à établir la taille de l'échantillon pour les études de validation interne vient, en réalité, de l'utilisation de l'analyse factorielle (ACP ou AFE) pour l'étude de la structure dimensionnelle de l'instrument.

De nombreuses recommandations existent dans la littérature concernant le nombre de sujets nécessaire à une analyse factorielle, cependant, aucune n'est basée sur des fondements empiriques ou théoriques rigoureux. La règle la plus utilisée est celle du ratio du nombre de sujets (N) sur le nombre d'items (p) dans l'instrument qui doit être égal à trois ou 10, voire 20 selon les auteurs (Cattell, 1978; Everitt, 1975; Gorsuch, 1983; Nunnaly, 1978). D'autres recommandations portent sur une taille d'échantillon minimum variant de 50 à 500 sujets selon les auteurs (Aleamoni, 1973; Comrey et Lee, 1992; Comrey, 1978; Loo, 1983). Devant la multiplicité de ces recommandations et l'insuffisance de leurs justifications, certains auteurs se sont intéressés aux conséquences d'une taille d'échantillon insuffisante sur les résultats de l'analyse factorielle. En dehors de N, deux autres paramètres se sont révélés importants pour la fiabilité et la stabilité de la solution factorielle : le ratio du nombre d'items sur le nombre de

facteurs (M) présents dans l'instrument et la valeur des charges factorielles. L'importance du ratio p/M pour la qualité de la solution factorielle est justifiée par le fait que c'est un indicateur de la « détermination » des facteurs, i.e. le degré avec lequel le facteur est clairement représenté par un nombre suffisant d'items, au moins trois ou quatre selon Mac Callum *et al.* (MacCallum *et al.*, 1999). En ce qui concerne les charges factorielles, leurs valeurs reflètent le niveau des communautés des variables, i.e. les proportions de variance que chaque variable partage avec les facteurs communs. Plus le ratio p/M et la valeur des charges factorielles sont faibles, plus le nombre de sujets à inclure devra être important pour assurer une fiabilité et une stabilité données à la solution de l'analyse factorielle (Guadagnoli et Velicer, 1988; Hogarty *et al.*, 2005; MacCallum *et al.*, 1999; Mundfrom *et al.*, 2005; Velicer et Fava, 1998).

Ces études ont donc montré que la taille d'échantillon dépend en partie de la nature des données à analyser, de leur « force » : des données « fortes » en analyse factorielle impliquent un haut niveau de communautés sans charges croisées (deux facteurs chargeant sur le même item) et des facteurs chargeant fortement sur plusieurs variables (Costello et Osborn, 2005; Fabrigar *et al.*, 1999). Plus les données sont fortes, plus le nombre de sujets nécessaire à l'analyse factorielle est petit. Dans ces conditions, étant donnée l'extrême variété des données rencontrées dans les différents champs disciplinaires où les analyses factorielles sont utilisées, il devient difficile d'envisager une règle générale pour le calcul de la taille d'échantillon valide dans l'ensemble de ces champs. Cependant, champs par champs, il est possible de mettre en évidence des caractéristiques communes aux échelles utilisées.

En psychiatrie, les charges factorielles sont habituellement proches de 0,6, le ratio p/M varie entre 2 et 20 ou plus en fonction des échelles et le nombre d'items chargés par chacun des facteurs est souvent différent à l'intérieur d'une même échelle (Dawkins *et al.*, 2006; Gabryelewicz *et al.*, 2004; Iwata *et al.*, 2000; Loza *et al.*, 2003). Une autre caractéristique observée dans les échelles utilisées en psychiatrie est la forme du diagramme des valeurs propres. L'unidimensionnalité est rare et il existe en général une première dimension représentant une grande part de la variance contenue dans les données (30 à 35%) puis une ou plusieurs autres dimensions expliquant des parts de variance de plus en plus petites (de 15 à

5%) (Chapman *et al.*, 2009; Sanchez-Lopez Mdel et Dresch, 2008; Uslu *et al.*, 2008; Villalta-Gil *et al.*, 2006). Une telle structure peut être interprétée comme la présence de facteurs corrélés ou, de manière semblable, comme la présence d'un facteur de second ordre expliquant la structure des corrélations entre les facteurs de premier ordre.

L'inclusion d'un nombre insuffisant de sujets dans une étude de validation d'échelle de mesure peut être à l'origine de conclusions erronées. Inversement, si le nombre de sujets inclus est inutilement grand, il en découle une perte de temps et de ressources. L'objectif principal de cette étude est d'utiliser les caractéristiques spécifiques des échelles rencontrées dans le champ de la psychiatrie afin de développer un outil de détermination de la taille d'échantillon nécessaire à l'étude de leur validité interne pour garantir une précision acceptable du coefficient alpha de Cronbach et, surtout, la stabilité et la fiabilité de la solution factorielle. Un objectif secondaire est d'évaluer l'influence du choix entre ACP et AFE sur le nombre de sujets nécessaire et sur la fiabilité de la solution factorielle.

2. Matériel et méthodes

Cette étude a comporté trois étapes. En premier lieu, une revue de la littérature a eu pour but de déterminer les caractéristiques communes des échelles en psychiatrie. Ensuite, une étude par simulation a été menée pour évaluer l'influence de la taille d'échantillon sur la stabilité et la fiabilité de la solution obtenue par ACP ou AFE. Une première série de simulations a été réalisée selon le modèle de structure dimensionnelle retrouvée en psychiatrie par la revue de la littérature, puis une seconde série a été réalisée à l'aide de données réelles. Enfin, l'influence de la taille de l'échantillon sur la précision du coefficient alpha de Cronbach dans les conditions rencontrées en psychiatrie a été étudiée.

a. Revue de la littérature

Dix échelles ont été sélectionnées sur leur fréquence d'utilisation en pratique clinique et pour représenter un échantillon des grandes pathologies rencontrées en psychiatrie :

- la Positive And Negative Syndrome Scale (PANSS - 30 items)
- la Brief Psychiatric Rating Scale (BPRS – 18 items)

- le Beck Anxiety Inventory (BAI - 21 items)
- le State-Trait Anxiety Inventory (STAI - 40 items)
- la Hamilton Anxiety Rating Scale (HAMA - 14 items)
- la Hamilton Rating Scale for Depression (HAMD - 17 items)
- la Montgomery-Asberg Depression Rating Scale (MADRS - 10 items)
- le Beck Depression Inventory (BDI - 21 items)
- la Hospital Anxiety and Depression Scale (HADS - 14 items)
- le General Health Questionnaire (GHQ – 12 items)

Une recherche des articles incluant des résultats d'ACP ou AFE d'une de ces dix échelles a été effectuée dans la base de données Medline en utilisant les mots-clefs suivants : pour chaque échelle, le "nom de l'échelle en anglais" et/ou son "abréviation", les expressions "factor analysis" et/ou "components analysis" et la langue de l'article "English" et/ou "French". Une présélection a été réalisée sur la base des résumés puis les articles présentant les trois critères suivants ont été inclus: la détermination de la structure dimensionnelle d'une des dix échelles par ACP ou AFE, l'indication des valeurs propres ou du pourcentage de variance expliquée par chaque facteur avant rotation, l'effectif de l'échantillon supérieur ou égal à 100.

Dans chaque article étaient relevés la méthode d'extraction des facteurs (ACP ou AFE), la méthode de rotation utilisée (orthogonale ou oblique), le nombre de facteurs retenus, les valeurs propres ou le pourcentage de variance expliquée par chaque facteur avant rotation, le nombre d'items par facteur et la valeur des coefficients de corrélation inter-facteurs. Lorsque la matrice des charges était reproduite dans l'article, la moyenne des charges principales (i.e. la charge la plus élevée parmi toutes celles qui existent entre un item et les facteurs de l'échelle) était calculée. Si l'étude portait sur plusieurs groupes de sujets, ce sont les résultats issus du groupe avec le plus grand effectif qui étaient pris en compte. De même, si l'analyse était effectuée sur des données relevées à des temps différents, les résultats pris en compte étaient ceux obtenus au temps initial. L'ensemble de ces données a été consigné sur tableur Microsoft® Office Excel version 2007 et les analyses statistiques descriptives pour

chacune de ces variables ont été effectuées à l'aide du logiciel R version 2.6.2 (R Development Core Team, 2008).

b. Etudes de simulation

i. Simulations basées sur des données artificielles

La méthode de simulation développée repose sur le modèle en facteurs communs et spécifique et est décrite dans l'appendice. Certains points importants doivent être soulignés. Dans ce modèle de simulation, deux hypothèses ont été faites : l'existence d'une structure factorielle simple (i.e. chaque item est chargé par un seul facteur et la valeur de toutes les charges secondaires a été fixée à zéro) et la même valeur λ pour l'ensemble des charges principales. Lorsque le modèle en facteurs communs et spécifique est utilisé, les réponses aux items ont une distribution normale. Pour se rapprocher des instruments rencontrés en pratique, ces réponses ont été catégorisées en quatre classes ordonnées tel un dispositif de réponse de type Likert. La distribution des réponses était différente et non-symétrique pour chacun des items de l'échelle afin de simuler des effets plancher et plafond. Enfin, les paramètres pouvant être contrôlés dans cette méthode de simulation étaient donc : l'effectif de l'échantillon N, le nombre de facteurs M, la valeur des charges principales λ, le nombre total d'items p, le nombre d'items chargés par chacun des facteurs de l'échelle p_m et la valeur des corrélations inter-facteurs ($cor(F_m, F_{m'}), m \neq m'$).

Pour M et p, les valeurs habituellement retrouvées dans les échelles en psychiatrie ont été étudiées, i.e. échelles à deux, trois ou quatre facteurs ayant un nombre d'items variant entre 10 et 45 items ($p = 10, 15, 20, 25, 30, 35, 40\ ou\ 45$). Les valeurs prises par λ et p_m ont été déterminées par la revue de la littérature. Enfin, les valeurs des corrélations inter-facteurs ont été choisies parmi les valeurs rencontrées dans la littérature mais aussi de manière à ce que le pourcentage de variance expliquée par chaque facteur soit au plus proche de la moyenne du pourcentage retrouvé dans la littérature. Une fois que l'ensemble des valeurs prises par les paramètres fut déterminé, deux séries de 10 000 échantillons ont été simulées pour chaque valeur de N étudiée ($N = 50, 100, 150, 200, 300, 500, 1000$) et pour chaque condition définie par M et par p. Ensuite, une ACP a été réalisée sur chaque échantillon d'une série et une AFE

sur les échantillons de l'autre série. Ces deux méthodes d'extraction des facteurs ont été suivies d'une rotation oblique (de type promax) comme recommandé lorsque les facteurs de l'échelle sont corrélés (Costello et Osborn, 2005; Fabrigar *et al.*, 1999; Floyd et Widaman, 1995). Pour déterminer le nombre de sujets nécessaires, trois critères ont été utilisés comme seuils de bonnes fiabilité et stabilité de la solution factorielle obtenue :

- Un écart-type de la valeur des charges principales obtenues après rotation sur les 10000 simulations (σ_λ) inférieur à 0,05 (soit une demi-amplitude de l'intervalle de confiance des charges principales proche de 0,1)
- Un pourcentage de simulations où l'ensemble des items de l'échelle sont chargés par le bon facteur (celui qui est déterminé dans le modèle de simulation) après rotation ($R_{\%}$) supérieur à 90%
- Une moyenne du pourcentage d'items chargés par le mauvais facteur après rotation sur les 10 000 simulations ($W_{\%}$) inférieur à 1%.

Lorsqu'une AFE était réalisée, le pourcentage de simulations avec survenue de phénomènes d'Heywood (i.e. estimation de la valeur d'une charge supérieure à 1, phénomène ne survenant pas dans le cas de l'ACP) était aussi calculé. Enfin, pour chacune des deux méthodes, la moyenne des charges principales sur les 10000 simulations (μ_λ) était estimée.

ii. Simulations basées sur des données réelles

Pour apporter une perspective complémentaire, une étude par simulation a été conduite à l'aide d'une base contenant les données de 1009 patients hospitalisés pour la première fois pour trouble du comportement alimentaire à la Clinique des Maladies Mentales et de l'Encéphale de l'hôpital Sainte-Anne à Paris entre janvier 1988 et juillet 2004 (Fedorowicz *et al.*, 2007). Les données concernant deux des questionnaires passés lors de l'inclusion ont été utilisées : la version à 13 items du BDI (Beck *et al.*, 1961) et la version à 21 items de l'HAMD (Hamilton, 1960). Une analyse parallèle a permis la détermination du nombre de facteurs à extraire pour chacune de ces deux échelles, puis, deux séries de 10000 échantillons ont été simulées par tirage au sort avec remise dans cette base de données pour chacune des tailles d'échantillon suivantes : 100, 200, 300, 400, 500, 600, 700 et 800. Une ACP a ensuite été

réalisée sur une série et une AFE sur l'autre. Ces analyses étaient chacune suivies d'une rotation promax dans le cas d'un instrument multidimensionnel. La moyenne des écarts-type des charges était ensuite calculée sur les 10000 échantillons simulés par taille d'échantillon étudiée.

Les analyses ont été effectuées grâce au logiciel R version 2.6.2. La fonction `princomp` a été utilisée pour les ACP et la rotation de la matrice des charges obtenue a été effectuée à l'aide de la fonction `promax` avec une constante fixée à quatre (Costello et Osborn, 2005; Jackson, 1991). Pour les AFE, la fonction `factanal` (avec l'argument `rotation=promax`) dont la procédure d'estimation est le maximum de vraisemblance, a été choisie pour deux raisons : la solution trouvée est celle qui a les propriétés statistiques optimales et c'est la méthode la plus largement utilisée en pratique (Revelle, 2008). Enfin, le tirage au sort a été réalisé à l'aide de la fonction `sample` et l'analyse parallèle à l'aide de la fonction `scree.plot` du package `psy`.

c. Etude de la précision du coefficient alpha de Cronbach

La valeur du coefficient alpha de Cronbach (α) la plus souvent citée comme étant la minimum acceptable est 0,7 (Fedorowicz *et al.*, 2007; Nunnaly, 1978; Peterson, 1994). La relation entre p, N (mêmes valeurs que précédemment) et la demi-amplitude de l'intervalle de confiance de ce coefficient pour trois valeurs attendues (α = 0,7 ; 0,8 et 0,9) a été étudiée. La formule de calcul de cet Intervalle de Confiance (IC) établie par Feldt a été utilisée avec une valeur du risque de première espèce fixée à 0,05 (Fan et Thompson, 2001) :

Borne supérieure : $IC_{sup} = 1 - \left[(1-\alpha) \times \mathcal{F}_{(0{,}025;ddl_1;ddl_2)}\right]$

Borne inférieure : $IC_{inf} = 1 - \left[(1-\alpha) \times \mathcal{F}_{(0{,}975;ddl_1;ddl_2)}\right]$

où $ddl_1 = N - 1$, $ddl_2 = (N-1)(p-1)$ et $\mathcal{F}$ représentant la valeur des percentiles 0,025 et 0,975 d'une distribution de Fisher.

3. Résultats

a. Caractéristiques des échelles en psychiatrie

Les mots-clefs utilisés pour la recherche dans la base de données Medline ont permis l'identification de 827 études. Parmi elles, 232 ont été présélectionnées à l'aide du résumé et 56 articles répondaient aux critères d'inclusion. Cinq de ces articles montraient des résultats d'analyse factorielle sur deux des dix échelles sélectionnées, ce qui augmenta finalement le nombre de références inclues à 61. Le tableau 1 présente pour chacune des échelles sélectionnées, le nombre total d'études inclues et le nombre d'études où un même nombre de facteurs était identifié dans l'échelle.

Afin d'estimer un modèle moyen des structures factorielles rencontrées en psychiatrie, les analyses descriptives ont porté sur l'ensemble des références inclues sans tenir compte du nombre de facteurs retrouvés dans les échelles. Dans le tableau 2 sont indiquées les moyennes du pourcentage de variance expliquée par chaque facteur avant rotation pour chaque échelle. Un diagramme en boîtes à moustaches de ces pourcentages en fonction du rang du facteur dans l'échelle sur l'ensemble des références inclues est représenté dans la figure 5.

Tableau 1 : Références des études inclues et nombre d'études dans lesquelles le nombre de facteurs extraits était identique pour chaque échelle sélectionnée

Echelle	Références	Total	Nombre de facteurs					
			2	3	4	5	6	7
PANSS	(Bell *et al.*, 1994; Fresan *et al.*, 2005; Honey *et al.*, 2003; Kay et Sevy, 1990; Lancon *et al.*, 1999; Lee *et al.*, 2003; Lindenmayer *et al.*, 2004; Loza *et al.*, 2003; Lykouras *et al.*, 2000; Salokangas *et al.*, 2002; Villalta-Gil *et al.*, 2006)	11	-	-	1	8	-	2
BPRS	(Adachi *et al.*, 2000; Harvey *et al.*, 1996; Lachar *et al.*, 2001; Ventura *et al.*, 2000)	4	-	-	2	1	1	-
BAI	(Beck, 1991; Chapman *et al.*, 2009; Kabacoff *et al.*, 1997; Steer *et al.*, 1995, 1993)	5	4	-	1	-	-	-
STAI	(Iwata *et al.*, 2000, 1998; Kabacoff *et al.*, 1997)	3	2	1	-	-	-	-
HAMA	(Beck, 1991; Serretti *et al.*, 1999)	2	2	-	-	-	-	-
HAMD	(Grunebaum *et al.*, 2005; Olden *et al.*, 2009)	2	-	-	1	1	-	-
MADRS	(Gabryelewicz *et al.*, 2004; Galinowski et Lehert, 1995; Lee *et al.*, 2003; Parker *et al.*, 2003; Serretti *et al.*, 1999)	5	3	2	-	-	-	-
BDI	(Basker *et al.*, 2007; Bonicatto *et al.*, 1998; Bonilla *et al.*, 2004; Gorenstein *et al.*, 1999; Grunebaum *et al.*, 2005; Helm et Boward, 2003; Jo *et al.*, 2007; Killgore, 1999; Munoz *et al.*, 2007; Powell, 2003; Salamero *et al.*, 1994; Shek, 1990; Uslu *et al.*, 2008; Wang *et al.*, 2005)	15	9	2	3	-	-	1
HADS	(Dagnan *et al.*, 2008; Dawkins *et al.*, 2006; Friedman *et al.*, 2001; Pallant et Bailey, 2005; Smith *et al.*, 2002; Woolrich *et al.*, 2006)	6	4	2	-	-	-	-
GHQ12	(Castro-Costa *et al.*, 2008; Farrell, 1998; Hankins, 2008; Hu *et al.*, 1992; Kilic *et al.*, 1997; Lopez-Castedo et Fernandez, 2005; Sanchez-Lopez Mdel et Dresch, 2008; Werneke *et al.*, 2000)	8	5	3	-	-	-	-
Total		61	29	10	8	10	1	3

PANSS : Positive And Negative Syndrome Scale à 30 items, BPRS : Brief Psychiatric Rating Scale à 18 items, BAI : Beck Anxiety Inventory à 21 items, STAI : State-Trait Anxiety Inventory à 40 items, HAMA : Hamilton Anxiety Rating Scale à 14 items, HAMD : Hamilton Rating Scale for Depression à 17 items, MADRS : Montgomery-Asberg Depression Rating Scale à 10 items, BDI : Beck Depression Inventory à 21 items, HADS : Hospital Anxiety and Depression Scale à 14 items, GHQ12 : General Health Questionnaire à 12 items)

Tableau 2 : Pourcentage de variance expliquée par facteur avant rotation, par échelle et sur l'ensemble des références

Echelle		Facteurs						
		F_1	F_2	F_3	F_4	F_5	F_6	F_7
PANSS	Moyenne [Nombre références]	25,8 [11]	12,8 [11]	8,8 [11]	6,8 [11]	5,8 [10]	3,6 [2]	3,6 [2]
	(Minimum – Maximum)	(14,5 – 41,2)	(8,7 – 18,6)	(6,1 – 13,4)	(3,9 – 11,1)	(3,6 – 9,3)	(3,6 – 3,7)	(3,6 – 3,7)
BPRS	Moyenne [Nombre références]	18,9 [4]	14,0 [4]	10,4 [4]	8,3 [4]	6,9 [2]	6,7 [1]	
	(Minimum – Maximum)	(12,8 – 23,3)	(9,3 – 17,2)	(8,7 – 11,7)	(6,7 – 10,0)	(6,1 – 7,8)	(. - .)	
BAI	Moyenne [Nombre références]	37,6 [5]	7,2 [5]	6,2 [1]	5,2 [1]			
	(Minimum – Maximum)	(36,3 – 39,5)	(4,4 – 7,7)	(. - .)	(. - .)			
STAI	Moyenne [Nombre références]	32,2 [3]	9,6 [3]	6,0 [1]				
	(Minimum – Maximum)	(29,8 – 34,3)	(7,4 – 11)	(. - .)				
HAMA	Moyenne [Nombre références]	26,9 [2]	8,2 [2]					
	(Minimum – Maximum)	(20,4 – 33,5)	(6,4 – 10)					
HAMD	Moyenne [Nombre références]	12,8 [2]	11,3 [2]	10,7 [2]	8,5 [2]	9,4 [1]		
	(Minimum – Maximum)	(12,6 – 13,0)	(11,2 - 11,4)	(10,4 – 11,0)	(7,3 – 9,8)	(. - .)		
MADRS	Moyenne [Nombre références]	33,7 [5]	15,6 [5]	10,6 [2]				
	(Minimum – Maximum)	(25,1 – 41,1)	(10,4 – 26,9)	(10,2 – 11,0)				
BDI	Moyenne [Nombre références]	29,4 [15]	8,8 [15]	6,8 [6]	5,4 [4]	6,0 [1]	5,5 [1]	4,9 [1]
	(Minimum – Maximum)	(22,9 – 34,5)	(5,9 – 25,1)	(5,0 – 6,1)	(4,9 – 6,1)	(. - .)	(. - .)	(. - .)
HADS	Moyenne [Nombre références]	34,3 [6]	13,1 [6]	8,4 [2]				
	(Minimum – Maximum)	(23,6 – 41,4)	(11,4 – 16,4)	(8,1 – 8,6)				
GHQ12	Moyenne [Nombre références]	39,6 [8]	13,0 [8]	9,2 [3]				
	(Minimum – Maximum)	(30,3 – 50,9)	(8,5 – 25,9)	(8,6 – 9,8)				
Total	Moyenne [Nombre références]	30,4 [61]	11,3 [61]	8,7 [32]	6,9 [22]	6,3 [14]	4,9 [4]	4,0 [3]
	(Minimum – Maximum)	(12,6 - 50,8)	(4,4 – 26,9)	(5,0 – 13,4)	(3,9 – 11,1)	(3,6 – 9,4)	(3,6 – 6,7)	(3,6 – 4,9)

PANSS : Positive And Negative Syndrome Scale à 30 items, BPRS : Brief Psychiatric Rating Scale à 18 items, BAI : Beck Anxiety Inventory à 21 items, STAI : State-Trait Anxiety Inventory à 40 items, HAMA : Hamilton Anxiety Rating Scale à 14 items, HAMD : Hamilton Rating Scale for Depression à 17 items, MADRS : Montgomery-Asberg Depression Rating Scale à 10 items, BDI : Beck Depression Inventory à 21 items, HADS : Hospital Anxiety and Depression Scale à 14 items, GHQ12 : General Health Questionnaire à 12 items

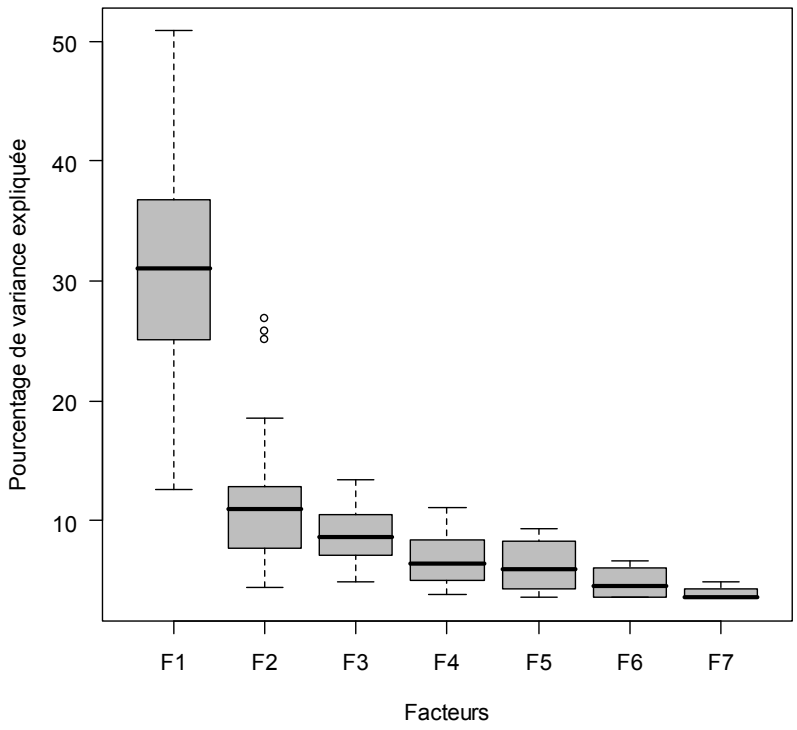


Figure 5 : Boîtes à moustaches du pourcentage de variance expliquée par chaque facteur en fonction de son rang dans l'échelle, sur l'ensemble des références inclues

La matrice des charges était reportée dans 95,1% (58) des cas. La moyenne des charges principales était égale à 0,626 avec une médiane (méd) à 0,636 et un intervalle interquartiles (IIQ) égal à [0,587 - 0,662]. Cette moyenne était de 0,635 (méd=0,642 ; IIQ : [0,601 - 0,671]) lorsque la méthode d'extraction des facteurs utilisée était l'ACP (80,3% des cas, soit 49 références) et de 0,593 (méd=0,601 ; IIQ : [0,545 - 0,637]) dans le cas de l'AFE. Une rotation des facteurs de type orthogonale était utilisée dans 63,9% (39) des cas et les corrélations inter-facteurs étaient indiquées dans 34,4% (21) des cas. En tout, 51 coefficients de corrélation étaient indiqués et leur moyenne était estimée à 0,356 (méd=0,33 ; IIQ : [0,155 - 0,535]). Enfin, concernant le ratio p/M, sa moyenne était de 7,1 (méd=6 ; IIQ : [5 - 10,5]) mais le nombre d'items sur chacun des facteurs étant inégal selon les échelles, le pourcentage moyen du nombre d'items par facteur a été calculé en fonction du nombre de facteurs mis en évidence dans l'échelle, comme indiqué dans le tableau 3.

Tableau 3 : Pourcentage moyen du nombre d'items par facteur (IIQ : Intervalle InterQuartiles)

Nombre de facteurs dans l'échelle		**Facteurs**						
		F_1	F_2	F_3	F_4	F_5	F_6	F_7
2	Moyenne	55,7	39,0					
	IIQ	[50,0 - 59,2]	[33,3 - 42,9]					
3	Moyenne	43,2	34,8	20,2				
	IIQ	[40,4 - 49,4]	[27,8 - 41,3]	[16,7 - 24,1]				
4	Moyenne	29,1	26,6	19,3	20,3			
	IIQ	[27,0 - 33,0]	[22,9 - 32,2]	[16,3 - 20,6]	[14,8 - 23,7]			
5	Moyenne	22,6	20,1	15,7	16,7	15,1		
	IIQ	[20,0 - 25,8]	[16,7 - 22,5]	[13,3 - 19,2]	[16,7 - 19,7]	[12,7 - 16,7]		
6	Moyenne	16,7	22,2	16,7	16,7	11,1	16,7	
	IIQ	[. - .]	[. - .]	[. - .]	[. - .]	[. - .]	[. - .]	
7	Moyenne	24,1	22,4	12,5	15,2	12,1	6,5	7,1
	IIQ	[19,5 - 26,7]	[20,2 - 25,2]	[10,5 - 15,5]	[13,3 - 16,9]	[9,8 - 13,3]	[5,0 - 8,1]	[4,0 - 9,0]

b. *Influence de la taille de l'échantillon sur la qualité des solutions obtenues par ACP ou AFE*

i. Résultats des simulations basées sur des données artificielles

Choix des valeurs pour les paramètres des modèles de simulation

En se basant sur les résultats de la revue de la littérature, la valeur de λ a été fixée à 0,6 dans le modèle de simulation. Les pourcentages moyens du nombre d'items par facteur indiqués dans le tableau 3 ont été utilisés pour déterminer les p_m. Par exemple, pour une échelle à trois dimensions, l'entier inférieur le plus proche de $p \times 0{,}45$ était pris comme valeur de p_1, l'entier inférieur le plus proche de $p \times 0{,}35$ comme valeur de p_2 et le nombre d'items restants comme valeur de p_3. Concernant les valeurs des corrélations inter-facteurs, elles étaient fixées à 0,45 dans une échelle à deux dimensions. Dans une échelle à trois dimensions, $cor(F_1, F_2)$ était fixée à 0,45 et les deux autres corrélations à 0,35. Dans une échelle à quatre dimensions, $cor(F_1, F_2)$, $cor(F_2, F_3)$ et $cor(F_1, F_3)$ était fixées à 0,35 et les trois autres coefficients à 0,45. Le diagramme de chemin du modèle d'analyse factorielle en facteurs communs et spécifique utilisé pour simuler les échantillons dans le cas d'une échelle à 3 dimensions et 10 items est représenté dans la figure 6.

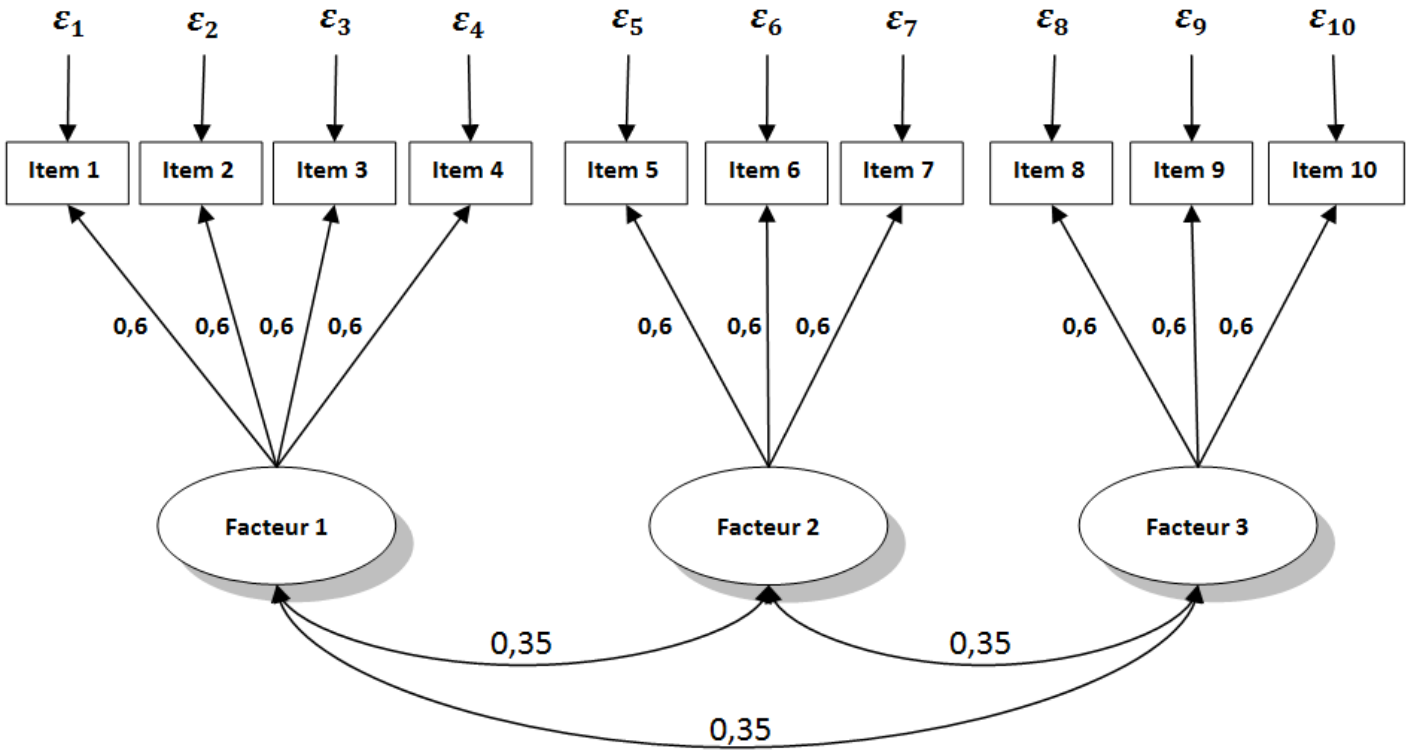


Figure 6 : Diagramme de chemin du modèle de simulation à trois facteurs et 10 items

Critères de qualité des solutions factorielles

Afin de faciliter la lecture des résultats, seules les données concernant σ_λ, $R_{\%}$ et $W_{\%}$ dans le cas d'une échelle à trois facteurs sont présentées. Le tableau 4 présente ces critères lorsqu'une ACP était effectuée et le tableau 5 dans le cas d'une AFE. Les trois critères de qualité ($\sigma_\lambda < 0{,}05$, $R_{\%}$>90% et $W_{\%} < 1\%$) étaient remplis pour $N \geq 500$ si l'échelle était constituée de moins de 25 items et pour $N \geq 300$ si elle était constituée de 25 items ou plus dans le cas de l'ACP. Lorsqu'une AFE était réalisée, les seuils de qualité étaient atteints par les trois critères pour des valeurs de N supérieures : 1000 pour une échelle constituée de moins de 20 items, 500 pour une échelle de 20 items ou plus.

Tableau 4 : Critères de jugement de la qualité des solutions d'une analyse en composantes principales dans le cas d'une échelle à trois dimensions

Effectif		**Nombre d'items**							
		10	**15**	**20**	**25**	**30**	**35**	**40**	**45**
	σ_λ	0,182	0,161	0,144	0,136	0,130	0,127	0,124	0,123
50	$R_{\%}$	48,4	48,5	51,1	50,7	51,3	51,5	50,6	49,5
	$W_{\%}$	9,3	6,5	4,4	3,4	2,7	2,3	2,1	1,9
	σ_λ	0,111	0,097	0,092	0,088	0,087	0,086	0,084	0,083
100	$R_{\%}$	88,8	92,6	94,6	95,9	96,4	97,1	96,9	97,1
	$W_{\%}$	1,5	0,6	0,3	0,2	0,1	0,1	0,1	0,1
	σ_λ	0,081	0,075	0,072	0,071	0,069	0,069	0,068	0,068
150	$R_{\%}$	97,8	99,4	99,5	99,7	99,8	99,7	99,8	99,8
	$W_{\%}$	0,3	0,1	-	-	-	-	-	-
	σ_λ	0,067	0,063	0,062	0,061	0,060	0,059	0,059	0,059
200	$R_{\%}$	99,5	99,9	99,8	99,8	99,9	99,9	99,9	99,8
	$W_{\%}$	0,1	-	-	-	-	-	-	0,1
	σ_λ	0,052	0,050	0,051	0,049	0,049	0,049	0,048	0,048
300	$R_{\%}$	99,9	100,0	99,8	99,9	99,9	99,8	100,0	99,8
	$W_{\%}$	0,1	-	0,1	-	-	0,1	-	-
	σ_λ	0,039	0,039	0,039	0,039	0,038	0,037	0,037	0,038
500	$R_{\%}$	99,9	100,0	99,9	99,9	99,9	99,9	99,9	99,9
	$W_{\%}$	-	-	-	-	-	-	-	-
	σ_λ	0,029	0,027	0,027	0,027	0,027	0,027	0,026	0,026
1000	$R_{\%}$	99,9	100,0	100,0	99,9	100,0	99,9	100,0	100,0
	$W_{\%}$	0,1	-	-	-	-	-	-	-

σ_λ : écart-type de la valeur des charges principales obtenues après rotation sur les 10000 simulations, $R_{\%}$: pourcentage de simulations où l'ensemble des items de l'échelle sont chargés par le bon facteur, $W_{\%}$: moyenne du pourcentage d'items chargés par le mauvais facteur après rotation sur les 10000 simulations, - : inférieur à 5.10^{-2}

Tableau 5 : Critères de jugement de la qualité des solutions d'une analyse factorielle exploratoire dans le cas d'une échelle à trois dimensions

Effectif		Nombre d'items							
		10	**15**	**20**	**25**	**30**	**35**	**40**	**45**
	σ_λ	0,226	0,187	0,164	0,153	0,144	0,138	0,134	0,131
50	$R_{\%}$	31,1	34,9	40,9	43,7	45,3	47,3	47,1	46,7
	$W_{\%}$	14,9	10,5	6,6	4,8	3,5	2,8	2,4	2,1
	σ_λ	0,159	0,125	0,109	0,101	0,096	0,093	0,091	0,089
100	$R_{\%}$	70,7	86,3	92,7	95,0	95,8	96,5	96,6	96,9
	$W_{\%}$	4,4	1,3	0,4	0,2	0,1	0,1	0,1	0,1
	σ_λ	0,128	0,098	0,086	0,080	0,077	0,075	0,073	0,072
150	$R_{\%}$	89,8	98,7	99,4	99,7	99,6	99,8	99,7	99,8
	$W_{\%}$	1,3	0,1	0,0	0,0	0,0	0,0	0,0	0,0
	σ_λ	0,109	0,082	0,073	0,069	0,066	0,064	0,063	0,062
200	$R_{\%}$	96,4	99,8	99,8	99,9	99,9	99,9	99,9	99,9
	$W_{\%}$	0,4	-	-	-	-	-	-	-
	σ_λ	0,086	0,065	0,059	0,056	0,054	0,052	0,051	0,051
300	$R_{\%}$	99,6	99,9	99,9	99,9	99,9	99,9	100,0	99,9
	$W_{\%}$	-	-	-	-	-	-	-	-
	σ_λ	0,063	0,050	0,045	0,043	0,041	0,040	0,039	0,040
500	$R_{\%}$	99,9	99,9	100,0	99,9	100,0	99,9	100,0	99,8
	$W_{\%}$	-	-	-	-	-	-	-	0,1
	σ_λ	0,043	0,034	0,032	0,030	0,029	0,028	0,028	0,027
1000	$R_{\%}$	100,0	100,0	99,9	100,0	100,0	100,0	100,0	100,0
	$W_{\%}$	-	-	-	-	-	-	-	-

σ_λ : écart-type de la valeur des charges principales obtenues après rotation sur les 10000 simulations, $R_{\%}$: pourcentage de simulations où l'ensemble des items de l'échelle sont chargés par le bon facteur, $W_{\%}$: moyenne du pourcentage d'items chargés par le mauvais facteur après rotation sur les 10000 simulations, - : inférieur à 5.10^{-2}

Dans le cas d'une échelle à deux facteurs, des valeurs inférieures de N permettaient d'atteindre les seuils de qualité : en général 300 sujets étaient suffisants sauf si l'échelle était constituée de moins de 30 items ou si une AFE était utilisée auxquels cas 500 sujets étaient nécessaires. A l'inverse, dans le cas d'une échelle à quatre facteurs, quel que soit le type d'analyse utilisée, la valeur de N nécessaire était supérieure (500), voire même, dans le cas d'une AFE et d'une échelle constituée de moins de 20 items, les critères de qualité n'étaient pas obtenus avec la valeur maximale de N étudiée (1000). Enfin, le pourcentage de simulations avec survenue de phénomènes d'Heywood dans le cas de l'AFE était toujours inférieur à 2% quels que soient la valeur de N et le nombre de facteurs dans l'échelle.

Une interpolation à partir des courbes représentant σ_λ en fonction de N pour chaque valeur de p et de M dans le cas de l'ACP et de l'AFE a permis de préciser la taille d'échantillon nécessaire à l'obtention d'une solution factorielle respectant les critères de qualité. L'intersection entre ces courbes et la droite correspondant à la valeur $\sigma_\lambda = 0{,}05$ a permis la détermination des tailles d'échantillons avec une précision de 50 sujets. Ces nombres, résumés dans le tableau 6, étaient toujours surestimés et permettaient toujours l'obtention des deux autres critères de qualité.

Tableau 6 : Taille d'échantillon nécessaire pour l'obtention des trois critères de qualité de la solution factorielle

Type d'analyse	**Nombre de facteurs**	**Nombre d'items**							
		10	**15**	**20**	**25**	**30**	**35**	**40**	**45**
ACP	**2**	300	300	300	300	300	300	250	250
	3	350	350	350	300	300	300	300	300
	4	400	400	350	350	350	350	350	350
AFE	**2**	500	400	350	300	300	300	300	300
	3	800	500	450	400	350	350	350	350
	4	-	-	600	500	450	400	400	400

ACP : analyse en composantes principales, AFE : analyse factorielle exploratoire, - : >1000

Exactitude de la solution factorielle

La figure 7 représente, pour une échelle à trois dimensions dans le cas de l'ACP et de l'AFE, μ_λ en fonction de N et pour chaque valeur de p. Dans le cas de l'ACP, la moyenne des charges principales était d'autant plus éloignée de la valeur attendue (λ=0,6) que le nombre d'items dans l'échelle était faible et la taille de l'échantillon était peu influente sur cette moyenne. Inversement, dans le cas de l'AFE, l'influence de N était plus importante pour l'exactitude de la solution factorielle : quel que soit le nombre d'item, chaque courbe se rapprochait de manière asymptotique de la valeur attendue. Ces courbes se comportaient de manière identique pour des échelles à deux dimensions ou à quatre dimensions mais dans ce dernier cas, la surestimation des valeurs moyennes des charges par l'ACP était plus importante et l'effectif nécessaire pour se rapprocher de l'asymptote était supérieur lors d'une AFE.

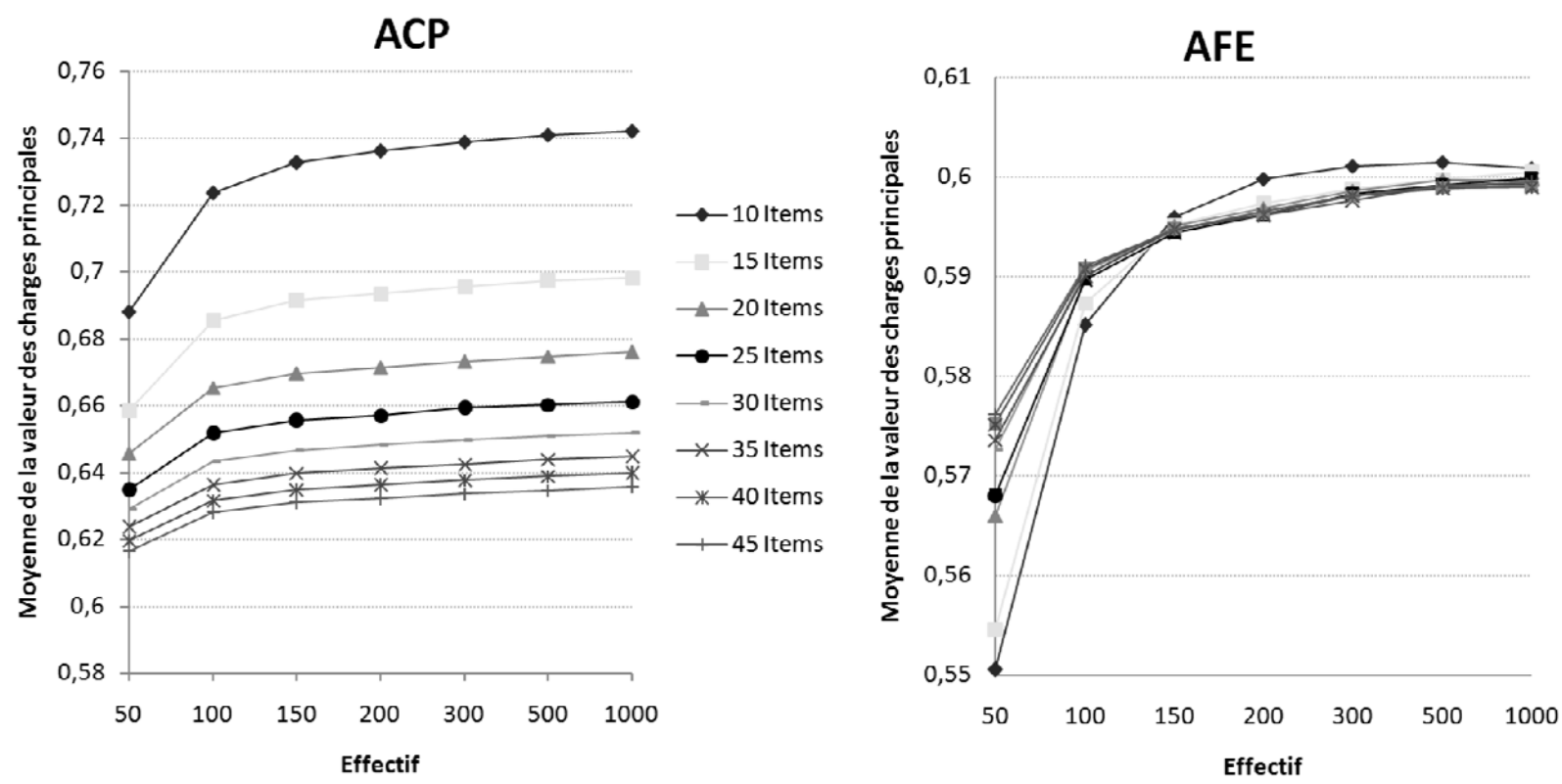


Figure 7 : Moyenne de la valeur des charges principales après rotation sur les 10000 simulations dans le cas d'une échelle à trois dimensions en fonction de l'effectif (ACP : Analyse en Composantes Principales, AFE : Analyse Factorielle Exploratoire)

ii. Résultats de l'étude de simulations basées sur des données réelles

Les analyses ont porté sur les 960 (95,1%) sujets ayant répondu à l'ensemble des 13 items du BDI et sur les 817 (81,0%) sujets ayant répondu à l'ensemble des 21 items de l'HAMD. Les résultats de l'analyse parallèle suggéraient l'unidimensionnalité du BDI alors que trois facteurs étaient retrouvés dans l'HAMD. La figure 8 présente la moyenne de l'écart-type des charges sur les 10000 échantillons simulés, estimées par ACP ou AFE suivies d'une rotation promax pour chacune des deux échelles. Dans le cas du BDI, cette moyenne était inférieure à 0,05 lorsque l'effectif était supérieur ou égal à 100 pour l'ACP, supérieur ou égal à 250 environ pour l'AFE. Dans le cas de l'HAMD, pour les valeurs d'effectif étudiées, ce seuil de 0,05 n'était jamais atteint par la moyenne de l'écart-type des charges.

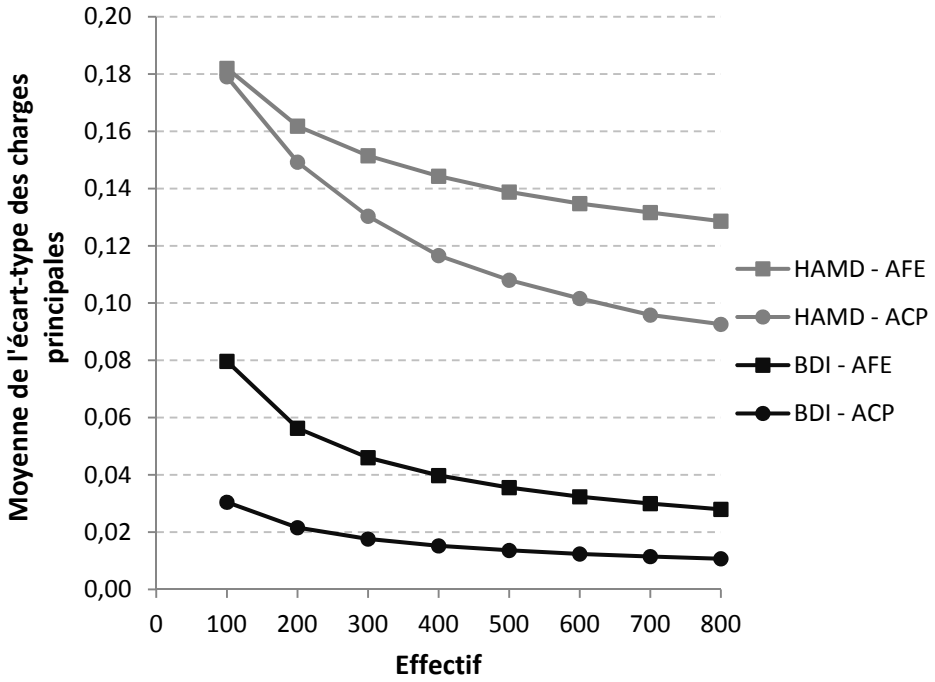


Figure 8 : Moyenne de l'écart-type des charges sur les 10000 échantillons simulés estimées par analyse en composantes principales (ACP) ou analyse factorielle exploratoire (AFE) suivies d'une rotation promax dans les cas de l'inventaire de dépression de Beck (BDI) et de l'échelle de dépression de Hamilton (HAMD).

Ce dernier résultat peu satisfaisant dans le cas de l'HAMD méritait d'être exploré plus attentivement. Notre hypothèse était que cette valeur obtenue pour la moyenne de l'écart-type des charges, bien supérieure à la valeur attendue, était due à l'existence possible de structures factorielles différentes dans certains échantillons parmi les 10000 simulés. Pour tester cette hypothèse, un modèle de mélange gaussien (fonction `Mclust` de la librairie `mclust` du logiciel R version 2.6.2) a été recherché dans la distribution de chaque charge principale de l'HAMD obtenue sur les 10000 échantillons pour un effectif de 400 sujets. L'hypothèse d'une composante unique était systématiquement rejetée et le nombre de composantes optimisant le critère d'information bayesien (BIC : Bayesian Information Criteria) était compris entre deux et six avec un mode égal à trois. Le programme de simulation éliminait la possibilité d'un phénomène artificiel de « label switching ».

c. Influence de la taille d'échantillon sur la précision du coefficient alpha de Cronbach

La figure 9 représente la demi-amplitude de l'intervalle de confiance à 95% du coefficient alpha de Cronbach en fonction de N pour les trois valeurs attendues (α = 0,7 ; 0,8 et 0,9). Seules les deux valeurs extrêmes de p étudiées (10 et 45 items) sont représentées car, comme le montre la figure 9, l'influence de p sur la précision du coefficient alpha de Cronbach était faible dans les conditions étudiées. Une demi-amplitude de 0,05 était atteinte pour un effectif de 300 lorsque la valeur attendue était 0,7, pour un effectif de 150 lorsque la valeur attendue était 0,8 et était atteinte dès 50 sujets si la valeur attendue était 0,9.

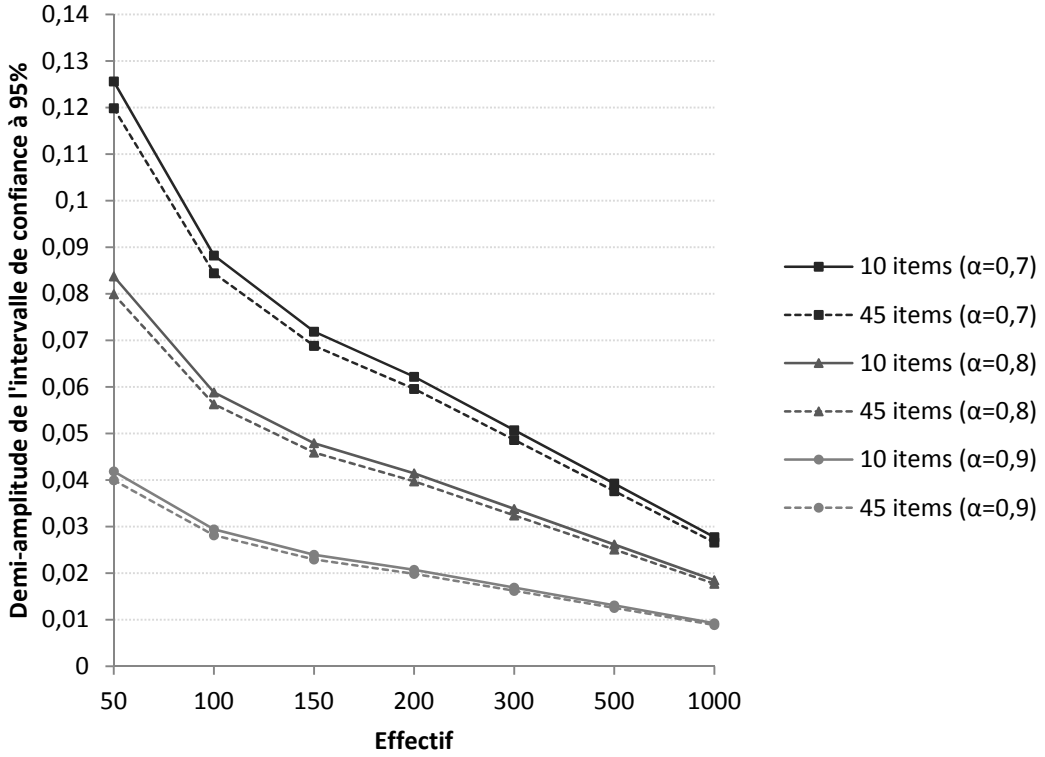


Figure 9 : Demi-amplitude de l'intervalle de confiance à 95% du coefficient alpha de Cronbach pour quatre valeurs attendues (α) en fonction de l'effectif et du nombre d'items dans l'échelle

4. Discussion

En se rapprochant le plus possible des conditions habituellement rencontrées en pratique, ces études de simulation ont permis d'apporter une réponse à la question de la taille d'échantillon nécessaire à la validation interne d'échelle en psychiatrie. En effet, si la structure factorielle de l'instrument est nette, l'estimation du nombre de sujets nécessaire à l'obtention d'une solution factorielle stable et précise est indiquée dans le tableau 6 pour différentes conditions définies par la méthode d'analyse choisie et par le nombre d'items et de facteurs dans l'échelle. Afin d'obtenir une précision souhaitée du coefficient alpha de Cronbach, cette estimation est ensuite à adapter à l'aide de la figure 9.

D'après les résultats de l'étude par simulation de données artificielles, 300 sujets sont habituellement suffisants mais ce nombre doit être augmenté dans trois cas : lorsque le nombre de facteurs dans l'échelle est grand, lorsque l'AFE est choisie comme méthode d'analyse et lorsque le nombre d'items est petit. Ce dernier point est l'un des résultats les plus importants de cette étude car il montre à quel point l'utilisation de la règle du ratio N/p peut être délétère, en particulier pour des échelles à faible nombre d'items. Plusieurs études récentes sur le sujet de la taille d'échantillon nécessaire à une analyse factorielle ont abouti à la même conclusion, cependant, l'étendue des valeurs des paramètres qu'elles étudiaient (λ, p, M) ne permettait pas de dégager une règle simple de détermination du nombre de sujets nécessaire (Guadagnoli et Velicer, 1988; Hogarty *et al.*, 2005; MacCallum *et al.*, 1999; Mundfrom *et al.*, 2005; Velicer et Fava, 1998).

Un autre résultat important concerne le choix entre ACP et AFE pour l'analyse de ce type de données. L'utilisation de l'ACP dans ce champ de recherche a été très souvent critiquée dans la littérature. En effet, le modèle d'analyse en facteurs communs et spécifiques est basé sur l'hypothèse de la présence de variables latentes qui expliquent les corrélations observées entre les items d'une échelle. Certains auteurs ont fait remarquer que l'ACP n'est pas totalement compatible avec cette hypothèse (Costello et Osborn, 2005; Fabrigar *et al.*, 1999; Floyd et Widaman, 1995). Par ailleurs, dans le modèle d'analyse en facteurs communs, la variance de chacun des items est décomposée en une part de variance commune et une part de variance unique, cette dernière comprenant la variance due à l'erreur de mesure et la variance spécifique à chaque item. La matrice de corrélations utilisée en AFE pour l'estimation des charges est la matrice « réduite » comprenant sur sa diagonale les communautés des items alors que dans l'ACP, aucune distinction n'est faite entre variance commune et unique et c'est la variance totale qui est représentée sur la diagonale de la matrice de corrélation (Fabrigar *et al.*, 1999; Ford *et al.*, 2006; Widaman, 1993). Les relations entre items sont par ce fait surestimées et les valeurs de charges retrouvées par ACP sont supérieures à celles estimées par l'AFE. Dans les conditions rencontrées en psychiatrie, cette surestimation des charges par ACP est d'autant plus importante que p est petit et M est grand ; ce biais n'étant pas diminué lorsque N augmente comme le montre la figure 7. L'utilisation de

l'AFE est donc recommandée dans ce champ afin d'obtenir des solutions factorielles moins biaisées.

Devant la difficulté à recommander une règle générale de calcul de la taille d'échantillon valide dans l'ensemble des champs où les techniques psychométriques s'appliquent, une revue de la littérature a permis d'identifier les caractéristiques des structures factorielles rencontrées en psychiatrie. Dans ce domaine, les facteurs sont en général corrélés entre eux, la valeur des charges principales est en moyenne de l'ordre de 0,6 et il existe une assez bonne détermination des facteurs avec un ratio p/M supérieur à sept. A l'aide de ce modèle « moyen » de structure factorielle, des données catégorielles intégrant différents niveaux d'effet plancher ou plafond pour chaque item ont pu être simulées. Les conditions rencontrées en psychiatrie ont donc été reproduites au plus près dans ces données artificielles, permettant ainsi d'obtenir des résultats mieux transposables à la pratique en conditions réelles que ceux des précédentes études de simulation sur ce sujet (Guadagnoli et Velicer, 1988, 1988; Hogarty *et al.*, 2005; Mundfrom *et al.*, 2005; Velicer *et al.*, 1982).

Il faut, néanmoins, noter que deux hypothèses ont été nécessaires au programme de simulation et pourraient être responsables d'une augmentation de la force des données simulées par comparaison aux données réelles. Une de ces hypothèses concerne l'égalité des charges principales. L'absence d'influence significative de cette hypothèse sur la qualité de la solution factorielle a été soulignée par Velicer et Fava dans leur étude de simulation conduite en 1998 (Velicer et Fava, 1998). L'autre hypothèse concerne la structure simple des données : absence de charges croisées et charges secondaires fixées à zéro. Pour permettre d'évaluer l'influence de ces hypothèses sur les tailles d'échantillon estimées dans le tableau 6, une étude de simulation basée sur données réelles a été réalisée. Les résultats de cette étude pourraient suggérer que les tailles d'échantillon recommandées sont sous-estimées, cependant, des structures factorielles différentes ont été observées dans les échantillons simulés par tirage au sort avec remise dans un seul et même échantillon de données réelles. L'importance de la moyenne de l'écart-type des charges observée sur ces 10000 échantillons résultait donc du mélange de différentes structures factorielles existant dans ces différents échantillons. De nouvelles études sont nécessaires pour explorer ce phénomène, il est cependant possible à

cette étape de conclure que les tailles d'échantillon présentées dans le tableau 6 sont les minima nécessaires, déterminées à partir d'une situation idéalisée dans laquelle le modèle d'analyse en facteurs communs est vrai. En pratique, la stabilité de la solution obtenue à partir de données réelles pourrait requérir un échantillon plus grand.

Les résultats obtenus dans cette étude sont basés sur une échelle « modèle » en psychiatrie et peuvent donc varier en fonction des caractéristiques propres à un instrument particulier. Des éléments de connaissances préalables sur p et M peuvent être utilisés pour préciser l'effectif nécessaire à partir du tableau 6. Par exemple, l'étude de la validité interne d'une échelle à cinq facteurs nécessite au minimum 400 sujets si l'ACP est choisie comme méthode d'analyse ou 450 sujets si c'est l'AFE qui est choisie. Enfin, dans cette étude, c'est l'influence de N sur la précision du coefficient alpha de Cronbach qui a été étudiée alors que des développements récents suggèrent que d'autres méthodes pourraient être plus appropriées pour l'évaluation de la consistance interne (Green et Yang, 2009; Sijtsma, 2008). Cependant, le débat concernant la meilleure méthode perdure et le coefficient alpha de Cronbach est encore de loin le plus utilisé en pratique.

5. Conclusion

Déjà mise à mal dans de précédentes études, la règle du ratio N/p ne trouve aucune justification dans les résultats obtenus ici et doit être abandonnée. La validation d'une petite échelle ne nécessite pas moins de sujets que la validation d'une grande échelle, au contraire. Si le but est de mettre en évidence la structure factorielle, sous l'hypothèse que le modèle d'analyse en facteurs communs et spécifique est vrai, un minimum de 300 sujets est en général acceptable dans les conditions rencontrées en psychiatrie. Cet effectif doit néanmoins être augmenté si le nombre de facteurs attendus dans l'échelle est grand. Par ailleurs, cette étude montre que pour obtenir des solutions factorielles plus exactes, l'AFE devrait être choisie comme méthode d'analyse.

6. Appendice

Dans le modèle en facteurs communs et spécifiques, chaque variable observée (item) est une combinaison linéaire d'un ou plusieurs facteurs communs et d'un facteur unique. L'équation fondamentale du modèle s'écrit de la manière suivante :

$$y_j = \lambda_{j1}F_1 + \lambda_{j2}F_2 + \cdots + \lambda_{jm}F_m + \cdots \lambda_{jM}F_M + \varepsilon_j$$

L'indice j représente les items ($j = 1$ à p) et l'indice m les facteurs ($m = 1$ à M). y_j est le vecteur de longueur N comportant les réponses des N sujets à l'item j. Chaque facteur commun F_m est un vecteur de longueur N comportant les niveaux non-observables des N sujets sur ce facteur. Ils chargent sur chaque item avec un coefficient spécifique à cet item, la charge λ_{jm}. Le facteur unique à chaque item est aussi un vecteur de longueur N qui est représenté par ε_j et est indépendant des facteurs communs et des facteurs uniques aux autres items (Brown, 2006).

Dans le modèle de simulation développé pour cette étude, deux hypothèses ont été posées : l'existence d'une structure simple (chaque item n'est chargé que par un seul facteur) et les charges principales (λ) sont toutes égales. Ainsi, si les p_1 premiers items ne sont chargés que par le premier facteur F_1, les p_2 items suivants par F_2, …, les p_m suivants par F_m, …, et les p_M derniers items par F_M ($\sum_{m=1}^{M} p_m = p$), les réponses à l'ensemble des p items peuvent être modélisées de la manière suivante :

$$\begin{cases} \forall j \in [1\,, p_1]\,, & y_j = \lambda' F_1 + \varepsilon_j \\ \forall j \in [(p_1+1), p_2]\,, & y_j = \lambda' F_2 + \varepsilon_j \\ \quad \vdots & \\ \forall j \in [(p_{m-1}+1), p_m]\,, & y_j = \lambda' F_m + \varepsilon_j \\ \quad \vdots & \\ \forall j \in [(p_{M-1}+1), p_M], & y_j = \lambda' F_M + \varepsilon_j \end{cases}$$

avec $\forall j \in [1, p],\ \varepsilon_j \sim \mathcal{N}(0,1)$ et $\varepsilon_j \perp \varepsilon_{(j' \neq j)}$

et $\forall m \in [1, M],\ F_m \sim \mathcal{N}(0,1)$ et $F_m \perp \varepsilon_j$

A noter que le coefficient λ' n'est pas directement égal à λ. En effet, afin de préserver les variances des y_j égales à l'unité, une standardisation est nécessaire par le facteur $\frac{1}{\sqrt{1+\lambda'^2}}$.

A partir de ce modèle, les données individuelles ont donc été simulées sous la forme d'une matrice où chaque ligne contient les réponses d'un individu à chacun des p items de l'échelle et chaque colonne contient les réponses des N individus à un item. Si i est l'indice des individus ($i = 1$ à N), la réponse de l'individu i à l'item j est donc de la forme :

$$\forall i \in [1, N], \qquad \forall m \in [1, M], \qquad \forall j \in [(p_{m-1}+1), p_m], \quad y_{ij} = \frac{\lambda' F_{mi} + \varepsilon_{ij}}{\sqrt{1+\lambda'^2}}$$

L'introduction d'une corrélation inter-facteurs a été rendue possible en modélisant chaque facteur à l'aide d'un terme spécifique à chacun des facteurs ($f_m \sim \mathcal{N}(0,1)$) et d'un terme commun à tous les facteurs ($C \sim \mathcal{N}(0,1)$) :

$$F_m = a_m f_m + b_m C \quad \text{avec } f_m \perp C$$

Les proportions de chacun de ces termes, a_m et b_m, permettent donc le contrôle des corrélations inter-facteurs avec pour seule contrainte que la somme de leurs carrés soit égale à 1, là aussi dans le but de conserver les variances des facteurs communs égales à l'unité. Enfin, une dernière étape a été nécessaire pour que les données simulées soient de type catégoriel et

de distribution non-symétrique à l'image de celles rencontrées en conditions réelles. En prenant pour exemple un dispositif de réponse de type Likert à quatre points, la transformation des y_{ij} en nombres entiers entre un et quatre a été réalisée en utilisant trois seuils dans leur distribution : $(-1+\delta_j)$, $(0+\delta_j)$ et $(1+\delta_j)$ où δ_j est tiré dans une distribution uniforme entre [-0,5 ; 0,5] pour introduire une asymétrie et ainsi simuler des effets planchers et plafonds.

La génération des données a été effectuée grâce au logiciel R version 2.6.2. Les vecteurs ε_j , f_m , et C ont été générés à l'aide de la fonction `rnorm` de ce logiciel et δ_j à l'aide de la fonction `runif`.

CHAPITRE 3 : UTILISATION DES MODELES ISSUS DE LA THEORIE DE REPONSE A L'ITEM POUR LA DETERMINATION DE LA DIFFERENCE MINIMALE CLINIQUEMENT PERTINENTE D'UN QUESTIONNAIRE

Traduction de l'article suivant, présenté en langue originale en annexe 2 :

Rouquette A, Blanchin M, Sébille V, Guillemin F, Côté S, Falissard B, Hardouin JB. The Minimal Clinically Important Difference determined using Item Response Theory Models: an attempt to solve the issue of the association with baseline score. *Journal of Clinical Epidemiology*; 67(4):433-440

Contribution des co-auteurs :

Alexandra Rouquette : Conception de l'étude, revue de la littérature, programmation des analyses statistiques, interprétation des résultats, rédaction, soumission et révision de l'article.

Myriam Blanchin : Contribution à la programmation des analyses statistiques et à l'interprétation des résultats, révision de l'article.

Véronique Sébille : Contribution à l'interprétation des résultats, à la révision de l'article.

Francis Guillemin : Contribution à l'interprétation des résultats, à la révision de l'article.

Sylvana Côté : Contribution à l'interprétation des résultats, à la révision de l'article.

Bruno Falissard : Contribution à l'interprétation des résultats, à la révision de l'article.

Jean-benoit Hardouin : Conception de l'étude, contribution à la programmation des analyses statistiques, à l'interprétation des résultats, à la rédaction et révision de l'article.

Résumé

Objectif. L'évaluation de la Différence Minimale Cliniquement Pertinente (DMCP) d'un questionnaire par la méthode recommandée dans la littérature (« anchor-based ») est problématique car elle mène à une valeur différente selon la sévérité initiale des sujets inclus dans l'étude. Sa détermination sur une échelle d'intervalle, l'échelle du Trait Latent (TL), en utilisant un modèle issu de la Théorie de Réponse à l'Item (Item Response Theory, IRT) pourrait éviter ce problème. L'objectif de cette étude était de comparer la sensibilité (Se), la spécificité (Sp) et les valeurs prédictives (VP) de la DMCP déterminée sur l'échelle du score (DMCP-Sc) ou sur l'échelle du TL (DMCP-TL).

Cadre et schéma de l'étude. La DMCP-Sc et la DMCP-TL de la sous-échelle santé générale du questionnaire MOS-SF36 ont été déterminées sur une cohorte de 1170 patients, dans le cas d'une aggravation ou d'une amélioration de l'état de santé perçu, en utilisant la méthode anchor-based et un modèle IRT de la famille de Rasch. Les Se, Sp et VP ont été calculées en prenant la question de transition du questionnaire MOS-SF36 comme gold standard.

Résultats. L'amplitude de la DMCP-Sc dans le cas de l'amélioration (1,58 points de score) était inférieure à celle retrouvée dans le cas d'une détérioration (-7,91 points). Les Se, Sp et VP étaient similaires pour la DMCP-Sc et la DMCP-TL dans les deux cas. Cependant, lorsque la DMCP était définie sur l'échelle du score par plusieurs valeurs en fonction de la sévérité initiale, les Se, Sp et VP étaient systématiquement plus élevées.

Conclusion. Les résultats de cette étude renforcent les recommandations récentes concernant l'utilisation d'une DMCP définie par plusieurs valeurs en fonction de la sévérité initiale.

Mots-clefs: Différence minimale cliniquement pertinente, Questionnaires, Sensibilité et spécificité, Théorie de réponse à l'item, Mesure d'intervalle, Question de transition, Trait latent

1. Introduction

Les échelles et les questionnaires sont de plus en plus utilisés dans les études longitudinales pour mesurer l'état de santé perçu par les sujets et évaluer son évolution au cours du temps. En effet, la perception que les individus ont de la santé et de la maladie influence leurs comportements ; les cliniciens, chercheurs et décideurs ont donc un intérêt croissant pour l'intégration de ce genre de mesures dans l'évaluation des traitements, interventions ou politiques de Santé Publique (Clancy et Eisenberg, 1998; Ellwood, 1988; McHorney, 1997; Roger, 2011; US Department of Health and Human Services, 2009). Une des principales limites à leur utilisation en recherche clinique ou épidémiologique est l'interprétation des mesures obtenues (Beaton *et al.*, 2011, 2002, 2001; Cook, 2008; Copay *et al.*, 2007; de Vet *et al.*, 2010; Ferreira *et al.*, 2011; Guyatt et Cook, 1994; Liang, 2000; Norman *et al.*, 1997; Revicki *et al.*, 2008; Stucki *et al.*, 1996).

Par exemple, que signifie une diminution de deux points d'anxiété sur une période de six mois lorsqu'elle est mesurée à l'aide d'une échelle dont le score varie de zéro à 20 ? Sur le plan clinique, cette différence traduit-elle une évolution du niveau d'anxiété perceptible par le patient, ses proches ou le clinicien ou est-elle « négligeable », c'est-à-dire sans répercussion clinique ? Le concept de Différence Minimale Cliniquement Pertinente (DMCP) a été initialement décrit en 1989 pour faciliter l'interprétation d'une différence observée lors de l'utilisation d'un questionnaire dans une étude longitudinale. Sa définition est « la plus petite différence de score dans le domaine d'intérêt que les patients perçoivent comme bénéfique et qui mènerait, en l'absence d'effets secondaires ou de coût excessif, à une modification de la prise en charge des patients » (Jaeschke *et al.*, 1989). Dans les cas où l'évaluation par les patients eux-mêmes est compliquée, cette définition a été adaptée au point de vue du clinicien comme « la plus petite taille d'effet qui le mènerait à recommander un traitement au patient » (van Walraven *et al.*, 1999).

Il n'existe pas encore de consensus clair dans la littérature concernant la meilleure méthode de détermination de la DMCP d'un questionnaire. D'un côté, certaines méthodes utilisent des indices statistiques basés sur la distribution de la différence de score dans la

population comme, par exemple, le d de Cohen (Cohen, 1988). De l'autre côté, les méthodes dites « anchor-based methods » utilisent un critère externe (« anchor ») ayant une pertinence clinique pour caractériser les différences de score observées (Beaton *et al.*, 2002; Copay *et al.*, 2007; Crosby *et al.*, 2003; Liang, 2000; Revicki *et al.*, 2008). Ce critère peut être un indicateur (dosage sanguin par exemple) de la réponse clinique à une intervention ou de l'évolution de la maladie mais le plus utilisé est le jugement global que porte le patient lui-même ou le clinicien sur l'évolution du phénomène mesuré par le questionnaire. En effet, cette dernière méthode, dite « du jugement global », est de plus en plus recommandée dans la littérature car elle est la seule à fournir une mesure de la signification du changement telle que perçu par l'individu (Cook, 2008; Crosby *et al.*, 2004; Revicki *et al.*, 2008; Terwee *et al.*, 2007; US Department of Health and Human Services, 2009). En pratique, de nombreuses études utilisent maintenant plusieurs types de critères externes (Purcell *et al.*, 2010; Sloan, 2005; Yost *et al.*, 2005).

Quelle que soit la méthode utilisée, la détermination de la DMCP reste problématique car, en fonction de l'échantillon utilisé, du critère externe utilisé, etc. la valeur de DMCP obtenue pour un même questionnaire est variable. L'existence d'une seule et unique valeur spécifique d'un questionnaire a d'ailleurs été largement remise en question dans la littérature (Beaton, 2003; Beaton *et al.*, 2002; Hays et Woolley, 2000; Revicki *et al.*, 2006; Terwee *et al.*, 2010). Par exemple, un phénomène particulier est régulièrement rencontré lors de l'utilisation des méthodes « anchor-based » : la dépendance de la valeur de la DMCP estimée au Score Initial (SI) des sujets de l'échantillon (Bird et Dickson, 2001; Crosby *et al.*, 2004; de Vet *et al.*, 2007; Jensen *et al.*, 2003; Lauridsen *et al.*, 2006; Stratford *et al.*, 1998, 1996; Stucki *et al.*, 1996; Ten Klooster *et al.*, 2006; Terwee *et al.*, 2010; Tubach *et al.*, 2005). Il a donc été recommandé dernièrement de définir la DMCP par plusieurs valeurs dépendantes du SI plutôt que par une seule et unique valeur (Copay *et al.*, 2007; Crosby *et al.*, 2004, 2003; Revicki *et al.*, 2008; Tubach *et al.*, 2005). En pratique, une telle définition implique que pour pouvoir conclure sur la pertinence clinique de l'évolution d'un score chez un sujet, différentes valeurs de DMCP devront être utilisées en fonction de son SI.

Dans la littérature, quatre explications à ce phénomène de dépendance au SI ont été avancées (Copay *et al.*, 2007). La première tient compte de la nature subjective de la DMCP :

un changement peut être perçu différemment par un sujet en fonction de son degré de sévérité initiale (Baker *et al.*, 1997). La deuxième explication porte sur la nature statistique de la DMCP qui est soumise, par ce fait, au phénomène de « régression vers la moyenne », i.e. la tendance observée des scores extrêmes à devenir moins extrêmes lors d'une mesure répétée (Crosby *et al.*, 2004, 2003). Enfin, les deux dernières explications concernent l'échelle de mesure utilisée pour déterminer la DMCP : le score, somme éventuellement pondérée des réponses aux items du questionnaire. Premièrement, cette échelle étant bornée, les bornes inférieures et supérieures sont responsables d'effets plancher et plafond, c'est-à-dire que la mesure d'un changement important pour les sujets dont le SI est proche d'une borne est souvent impossible car un tel changement dépasse les bornes de l'échelle (Baker *et al.*, 1997; Bird et Dickson, 2001; Copay *et al.*, 2007; Hays et Woolley, 2000). Deuxièmement, l'échelle du score n'a pas nécessairement les propriétés d'une échelle d'intervalle : sur une échelle d'intervalle, toutes les unités présentes le long de l'échelle sont égales les unes aux autres (Bird et Dickson, 2001; McHorney, 1997; Samsa *et al.*, 1999; Stevens, 1946). C'est sur ce défaut de propriété d'intervalle de l'échelle du score et sur son rôle potentiel dans la dépendance de la DMCP au SI que va porter cette étude. En effet, si l'échelle du score n'est pas une échelle d'intervalle, l'interprétation d'une différence de score peut varier selon la portion de l'échelle où elle se situe.

Les modèles issus de la théorie de réponse à l'item (Item Response Theory, IRT) permettent l'analyse de données issues de questionnaires et l'expression des résultats sur une échelle d'intervalle. Dans cette théorie, le concept mesuré par le questionnaire est estimé par une variable quantitative ayant les propriétés d'une mesure d'intervalle, le Trait Latent (TL) (Embretson et Reise, 2000). Ainsi, si un questionnaire mesure, par exemple, le niveau d'anxiété des sujets, une différence de x unités représente la même quantité, quelle que soit sa position sur l'échelle du TL (bas, moyen ou haut niveau d'anxiété). Si notre hypothèse sur le rôle de la propriété d'intervalle d'une échelle dans la dépendance de la DMCP au score initial est vraie, sa détermination sur l'échelle du TL par un modèle IRT pourrait éviter ce phénomène. Par rapport à la DMCP déterminée sur l'échelle du score, un meilleur classement

des sujets ayant vécu « aucun changement » versus « un changement cliniquement pertinent » pourrait donc être obtenu.

L'objectif de cette étude est de comparer la sensibilité (Se), la spécificité (Sp), les valeurs prédictives positive (VPP) et négative (VPN) de la DMCP déterminée sur l'échelle du score (DMCP-Sc) et de la DMCP déterminée sur l'échelle du trait latent (DMCP-TL) à l'aide d'un modèle IRT et de la méthode « anchor-based » dont le critère externe est pris pour test de référence.

2. Méthodes

a. Données

Les données sont issues de l'étude multicentrique longitudinale prospective française SATISQOL (satisfaction and quality of life) composée de 1709 patients hospitalisés, âgés de moins de 75 ans, inclus entre octobre 2008 et septembre 2010 et pour lesquels une intervention chirurgicale ou médicale était programmée dans le cadre de la prise en charge d'une maladie chronique cardiovasculaire, musculo-squelettique, néphro-urologique, digestive, pulmonaire ou endocrinologique. Pour être inclus, les patients devaient comprendre et parler le français, avoir une capacité cognitive suffisante pour remplir un questionnaire auto-administré et présenter des symptômes de leur maladie chronique depuis au moins six mois. Les sujets étaient exclus de l'étude en cas d'absence d'intervention thérapeutique au cours de leur hospitalisation.

Des informations sociodémographiques (âge, sexe, diagnostic, etc.), la satisfaction des soins auto-rapportée (version française du questionnaire Patient Judgements of Hospital Quality (Nguyen Thi *et al.*, 2002; Rubin *et al.*, 1990)) et la qualité de vie (version française du questionnaire Medical Outcomes Study Short Form-36 – MOS-SF36 (Leplège *et al.*, 2001; Ware et Sherbourne, 1992)) étaient relevées au cours de l'hospitalisation. Six mois plus tard, lors d'une consultation médicale programmée, il était demandé aux patients de remplir à nouveau le questionnaire MOS-SF36. Tous les patients ont donné par écrit leur consentement

éclairé pour participer à cette étude pour laquelle un avis positif a été émis par le comité d'éthique de Lorraine.

b. Questionnaire

Le questionnaire générique MOS-SF36 est composé de 36 items dont un évaluant l'évolution temporelle de la santé perçue et 35 regroupés en huit sous-échelles concernant la santé physique, mentale et sociale. Une des conditions d'application des modèles IRT étant l'unidimensionnalité, les analyses ont concerné les cinq items de la dimension « santé perçue » (SP)[3]. Le dispositif de réponse de chacun de ces items était une échelle de Likert à cinq modalités. Le score, variant entre 0 (niveau de santé perçue le plus bas) et 100, était calculé tel que recommandé par le manuel de l'utilisateur de la version française du questionnaire MOS-SF36 (Leplège *et al.*, 2001). Ainsi, si le nombre de réponses manquantes pour un même sujet était inférieur à trois, une imputation par la moyenne de l'individu aux autres items de la dimension SP été effectuée comme conseillé dans ce manuel.

L'item évaluant l'évolution de la santé perçue à six mois post-hospitalisation a été choisi comme question de transition (QT) évaluant le jugement global du sujet : « Par rapport à il y a six mois, comment trouvez-vous votre état de santé en ce moment ? ». Cinq réponses étaient possibles : « Bien meilleur », « Plutôt meilleur », « A peu près pareil », « Plutôt moins bon » et « Beaucoup moins bon ».

Les sujets ayant plus de deux réponses manquantes aux items de la dimension SP à l'un des deux temps de collecte ou n'ayant pas répondu à la QT lors de la collecte à six mois de l'hospitalisation ont été exclus des analyses.

c. Analyses

Il est maintenant reconnu que le changement, en termes de quantité et de qualité, est perçu différemment dans le sens d'une amélioration et dans le sens d'une détérioration

[3] Cette sous-échelle a été choisie car mesurant le même concept que celui évalué par l'item « Evolution temporelle de la santé perçue » choisi comme question de transition

(Beaton *et al.*, 2011; Crosby *et al.*, 2003; Hays et Woolley, 2000; Revicki *et al.*, 2008; Stratford *et al.*, 1998). L'ensemble des analyses de cette étude a donc été effectué dans chacune des deux circonstances.

i. Détermination de la DMCP-Sc

L'évolution sur six mois de la santé perçue telle que mesurée par la sous-échelle SP a été calculée comme la différence entre le score obtenu lors de l'hospitalisation (T1) et celui obtenu à six mois post-hospitalisation (T2). La DMCP-Sc a ensuite été calculée comme la moyenne de la différence de score entre T1 et T2 dans le sous-groupe de patients ayant répondu « Plutôt meilleur » à la QT dans le cas d'une amélioration (groupe amélioration) et dans le sous-groupe de patients ayant répondu « Plutôt moins bon » à la QT dans le cas d'une détérioration (groupe détérioration). La dépendance de l'évolution du score entre T1 et T2 au SI a été évaluée à l'aide d'un coefficient de corrélation de Pearson. Des coefficients de corrélation polychorique ont été utilisés pour étudier l'association entre l'évolution du score et la réponse donnée à la QT à 6 mois post-hospitalisation.

Afin de suivre les recommandations récentes, une DMCP-Sc_{SI} composée par plusieurs valeurs fonction du SI a aussi été calculée (Copay *et al.*, 2007; Crosby *et al.*, 2004, 2003; Revicki *et al.*, 2008; Tubach *et al.*, 2005). Concrètement, la DMCP-Sc_{SI} était composée des trois moyennes de la différence de score entre T1 et T2 pour les sujets ayant un SI dans le premier tiers ([0 - 33]), le second tiers (]33 – 67[) ou le dernier tiers ([67 – 100]) dans le groupe amélioration pour la DMCP- Sc_{SI} amélioration et dans le groupe détérioration pour la DMCP- Sc_{SI} détérioration.

ii. Détermination de la DMCP-TL

Hypothèses fondamentales de l'IRT

Trois hypothèses fondamentales sont nécessaires à l'utilisation des modèles IRT : l'unidimensionalité (la variance commune à un échantillon d'items est entièrement expliquée par une seule variable latente, le TL), l'indépendance locale (les réponses aux items sont indépendantes les unes aux autres une fois pris en compte le niveau sur le TL) et la

monotonicité (la probabilité d'une réponse non-nulle croît avec le niveau sur le TL). L'unidimensionnalité de la sous-échelle SP a été vérifiée, à chaque temps de collecte, par l'analyse des valeurs propres et par l'étude de l'adéquation d'un modèle d'Analyse Factorielle Confirmatoire (AFC) à un facteur. Pour évaluer cette adéquation, les indices suivants ont été utilisés : Root Mean Square Error Approximation (RMSEA, adéquation acceptable si <0,06) ; Comparative Fit Index (adéquation acceptable si >0,95) ; Tucker Lewis Index (adéquation acceptable si >0,95) et Standardized Root Mean Square Residual (adéquation acceptable si <0,08) (Hu et Bentler, 1999). L'adéquation d'un modèle monotone homogène de Mokken, modèle IRT non-paramétrique, a aussi été étudiée car une bonne adéquation de ce modèle aux données, évaluée à l'aide des coefficients H de Loevinger, indique que les trois hypothèses fondamentales de l'IRT sont vérifiées (Sijtsma et Molenaar, 2002). Enfin, le calcul du coefficient alpha de Cronbach à chaque temps de collecte a permis d'évaluer consistance interne de la sous-échelle SP, un coefficient supérieur à 0,7 étant considéré comme acceptable (Cronbach, 1951).

Adéquation du modèle de crédit partiel (Partial Credit Model - PCM) et estimation des paramètres d'item

Le PCM est modèle IRT pour données polytomiques de la famille de Rasch (cf. Matériel supplémentaire) couramment utilisé dans le champ de la santé (Anthoine *et al.*, 2012). L'adéquation de ce modèle a été évaluée séparément sur les données du premier et du deuxième temps de collecte à l'aide d'un test du Chi-2 d'adéquation. Ce test étant très sensible à la taille de l'échantillon, il a été effectué sur un échantillon de 400 sujets tirés au sort parmi les sujets de la cohorte. Cette taille d'échantillon a été évaluée comme suffisante pour l'estimation des paramètres d'un PCM (Smith *et al.*, 2008).

L'invariance de la mesure obtenue par la sous-échelle SP à T1 et à T2 a été vérifiée à l'aide des intervalles de confiance à 95% des estimations des paramètres d'item. En ajustant un PCM sur l'ensemble des données des deux temps de mesure sans distinction, il a été possible d'obtenir des estimations « moyennées » des paramètres d'items, comme recommandé lorsque la mesure est invariante au cours du temps (Norquist *et al.*, 2004; Wright, 1996).

Calcul de la DMCP-TL

Un modèle IRT de régression sur variable latente a été utilisé pour estimer la variation moyenne du niveau sur le TL au cours du temps dans les groupes amélioration et détérioration (cf. Matériel supplémentaire). La DMCP-TL a donc été définie comme l'effet temps sur l'échelle du TL (changement du niveau moyen sur le TL entre T1 et T2) dans le groupe amélioration pour la DMCP-TL d'amélioration et dans le groupe détérioration pour la DMCP-TL de détérioration. Pour ces analyses, les sujets ayant répondu « Bien meilleur » ou « Bien moins bon » à la QT ont été exclus et deux variables indicatrices ont été construites pour identifier les sujets ayant répondu « Plutôt meilleur » ou « Plutôt moins bon ».

Le niveau de SP observé étant mesuré sur l'échelle du score, le classement des sujets dans les catégories « ayant vécu un changement cliniquement pertinent » ou « n'ayant pas vécu de changement » à l'aide de la DMCP-TL obtenue a nécessité de traduire cette dernière en points de score. Un PCM a donc été utilisé pour obtenir la relation entre le niveau sur le TL et le score attendu pour la sous-échelle SP (cf. Matériel Supplémentaire). En utilisant cette relation pour la traduction, la différence de score équivalente à la DMCP-TL a été déterminée pour chaque SI possible, variant de 0,5 à 99,5 avec un incrément de 0,5. Ainsi, en connaissant le SI du patient, il était possible de déterminer si l'évolution de son score au cours des 6 mois était supérieure ou non à la DMCP-TL. La forme logistique du PCM n'a pas permis la traduction de la DMCP-TL pour les SI extrêmes (0 ou 100) ; une approximation à la valeur obtenue pour le SI le plus proche (0,5 et 99,5 respectivement) a été réalisée.

iii. Calcul des Se, Sp, VPP et VPN

Chaque patient de l'échantillon a été classé dans les catégories « ayant vécu un changement cliniquement pertinent » ou « n'ayant pas vécu de changement » au cours des six mois, à l'aide de la DMCP-Sc, de la DMCP-Sc_{SI} et de la DMCP-TL. Les Se, Sp, VPP et VPN de chaque type de DMCP ont ensuite été calculées en prenant la réponse à la QT comme classement de référence.

Par exemple, dans le cas de l'amélioration, les patients ayant répondu « Plutôt moins bon » et « Bien moins bon » étaient exclus, les patients ayant répondu « A peu près pareil »

étaient classés dans la catégorie « pas de changement » et les patients ayant répondu « Plutôt meilleur » et « Bien meilleur » étaient classés dans la catégorie « changement cliniquement pertinent ». Par exemple, la Se de la DMCP-Sc était calculée comme la proportion de patients classés dans la catégorie « changement cliniquement pertinent » par la DMCP-Sc parmi les patients classés dans cette même catégorie par la QT, classement de référence.

iv. Logiciels utilisés

Le logiciel Stata version 12.1 et le tableur Microsoft Office Excel 2007 ont été utilisés pour les analyses descriptives, les graphes, l'analyse factorielle et l'analyse par le modèle de Mokken (Hardouin *et al.*, 2011; StataCorp, 2012). Les paramètres d'item et l'adéquation du PCM ont été estimés à l'aide du logiciel RUMM 2030 (Andrich *et al.*, 2010). Enfin, la procédure NLMIXED du logiciel SAS version 9.3 a été utilisée pour estimer les DMCP-TL par le modèle IRT de régression sur variable latente (SAS Institute Inc, 2010).

3. Résultats

Au temps initial, 1709 patients (877 hommes – 56,1% et 686 femmes – 43,9%, le sexe n'était pas renseigné pour 146 patients) ont été inclus. La moyenne d'âge des participants était 55,7 ans (Ecart-Type – ET = 14,0) avec un minimum de 18 ans et un maximum de 80 ans. A six mois de suivi, le taux de réponse était égal à 89,4%, soit 1528 patients. Parmi eux, 58 n'avaient pas répondu à la QT à T2 et 300 avaient plus de deux réponses manquantes à la sous-échelle SP à T1 ou T2. Les données analysées dans cette étude concernaient donc 1170 patients. Le score moyen à la sous-échelle SP était égal à 52,1 (ET=22,4) à T1 et 51,7 (ET=23,3) à T2. Un histogramme du SI des sujets de l'échantillon est représenté dans la figure 10. Le SI était inférieur ou égal à 33 pour 269 (23,0%) patients et supérieur ou égal à 67 pour 372 (31,8%) patients.

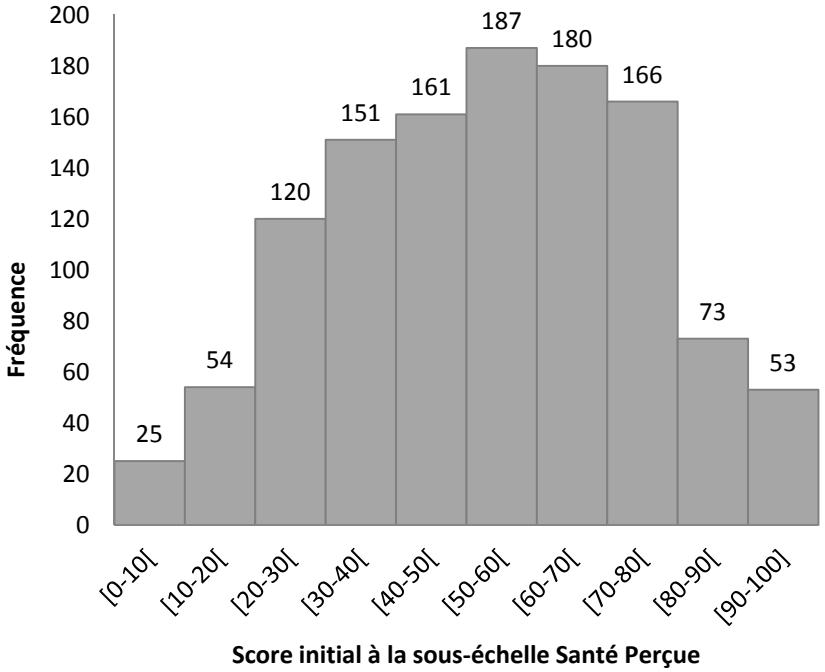


Figure 10 : Histogramme du score initial à la sous-échelle Santé Perçue

a. Détermination de la DMCD-Sc

La réponse à la QT était « Bien meilleur » pour 266 (22,7%) patients, « Un peu meilleur » pour 360 (30,8%), « A peu près pareil » pour 401 (34,3%), « Un peu moins bon » pour 112 (9,6%) et « Bien moins bon » pour 31 (2,6%). La DMCP-Sc de la sous-échelle SP était donc égale à 1,58 (ET = 0,76) points dans le cas de l'amélioration et à -7,91 (ET = 1,26) points dans le cas de la détérioration. Il est à noter que l'évolution moyenne du score entre T1 et T2 dans le groupe de patients considérés comme stables (ayant évalué leur état de santé à T2 comme « A peu près pareil » par rapport à six mois auparavant) était égale à -3,16 (ET=0,68). Un coefficient de corrélation polychorique égal à -0,29 était retrouvé entre l'évolution du score et les réponses à la QT.

Dans le groupe amélioration, le coefficient de corrélation de Pearson entre l'évolution du score et le SI était égal à -0,35 alors qu'il était égal à -0,62 dans le groupe détérioration. La figure 11 représente les boîtes à moustaches de la variation du score entre T1 et T2 en fonction du SI dans les groupes amélioration et détérioration. Globalement, dans le cas de

l'amélioration, plus le SI était élevé, plus l'évolution du score était faible. Inversement, dans le cas de la détérioration, plus le SI était élevé, plus l'évolution du score était importante.

Pour déterminer la DMCP-Sc_{SI}, les moyennes de l'évolution du score dans chaque sous-groupe de patients défini par le SI, indiquées dans la figure 11, ont été utilisées. Par exemple, la DMCP-Sc_{SI} d'amélioration était égale à 8,4 (ET=1,4) si le SI était compris dans l'intervalle [0 – 33] et à 2,5 (ET = 1,0) si le SI était dans l'intervalle]33 – 67[. Si le SI était dans l'intervalle [67 – 100], la DMCP-Sc_{SI} d'amélioration était fixée à zéro étant donnée la moyenne négative de l'évolution du score dans ce sous-groupe.

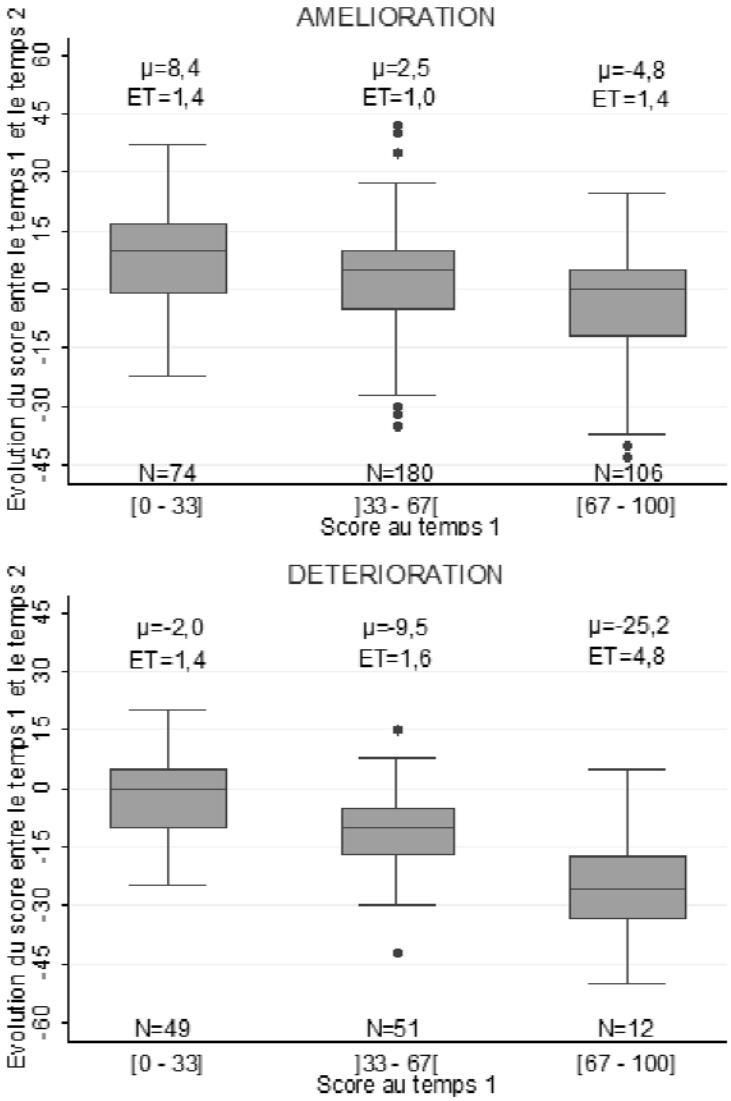


Figure 11 : Boites à moustaches de l'évolution du score à la sous-échelle Santé Perçue entre le temps 1 et le temps 2 en fonction du score au temps 1 dans les groupes de patients ayant répondu « Un peu meilleur » (amélioration) ou « Un peu moins bon » (détérioration) à la question de transition (μ : moyenne, ET : Ecart-Type, N : effectif)

b. Détermination de la DMCP-TL

Aux deux temps, une seule valeur propre était supérieure à un et le rapport de la première sur la seconde valeur propre était supérieur à quatre. L'ensemble des critères indiquait une adéquation acceptable du modèle d'AFC à un facteur sauf le RMSEA qui était égal à 0,088 à T1 et à 0,102 à T2. Cependant, aucun coefficient H de Loevinger n'a détecté de

violation aux hypothèses fondamentales de l'IRT. Enfin, un coefficient alpha de Cronbach égal à 0,81 à T1 et à 0,84 à T2 indiquait une bonne consistance interne de la sous-échelle SP.

L'hypothèse de la bonne adéquation d'un PCM aux données ne pouvait être rejetée ni à T1 ni à T2 en prenant un risque de première espèce à 5% ($p = 0,19$ à T1 et $p = 0,32$ à T2). La mesure fournie par la sous-échelle SP a été considérée comme invariante au cours du temps car les intervalles de confiance des 20 paramètres d'item estimés à T1 et T2 se recoupaient. La DMCP-TL était estimée à 0,0839 (ET = 0,0443) unités sur l'échelle du TL dans le cas de l'amélioration et à -0,4806 (ET = 0,0833) dans le cas de la détérioration. Il est à noter que la moyenne de l'évolution du niveau sur l'échelle du TL dans le groupe de patients considérés comme stables était égale à -0,1919 (ET = 0,0426).

La figure 12 décrit la relation entre le score attendu à la sous-échelle SP et le niveau sur le TL. Sa forme logistique est typique des modèles IRT de la famille de Rasch. En utilisant cette relation comme outil de traduction de la DMCP-TL en points de score, il a été possible de la représenter comme sur la figure 13 avec, en abscisse, la DMCP-TL en points de score et en ordonnées, le SI. Par exemple, le score d'un patient avec un SI de 20 devrait augmenter de 1,5 point de score à T2 pour classer ce patient dans la catégorie « changement cliniquement pertinent » à l'aide de la DMCP-TL alors qu'il faudrait une augmentation de seulement 0,5 point si le patient avait un SI de 80.

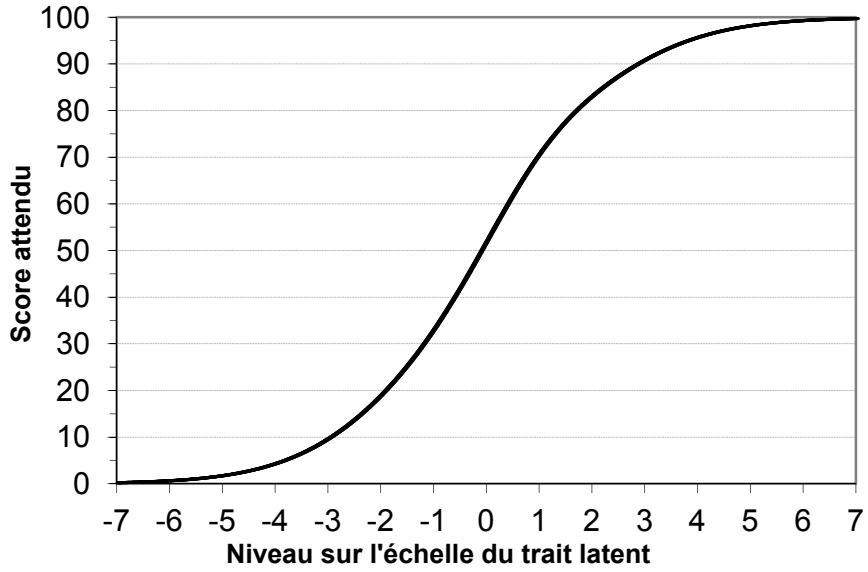


Figure 12 : Score attendu à la sous-échelle SP en fonction du niveau sur le trait latent

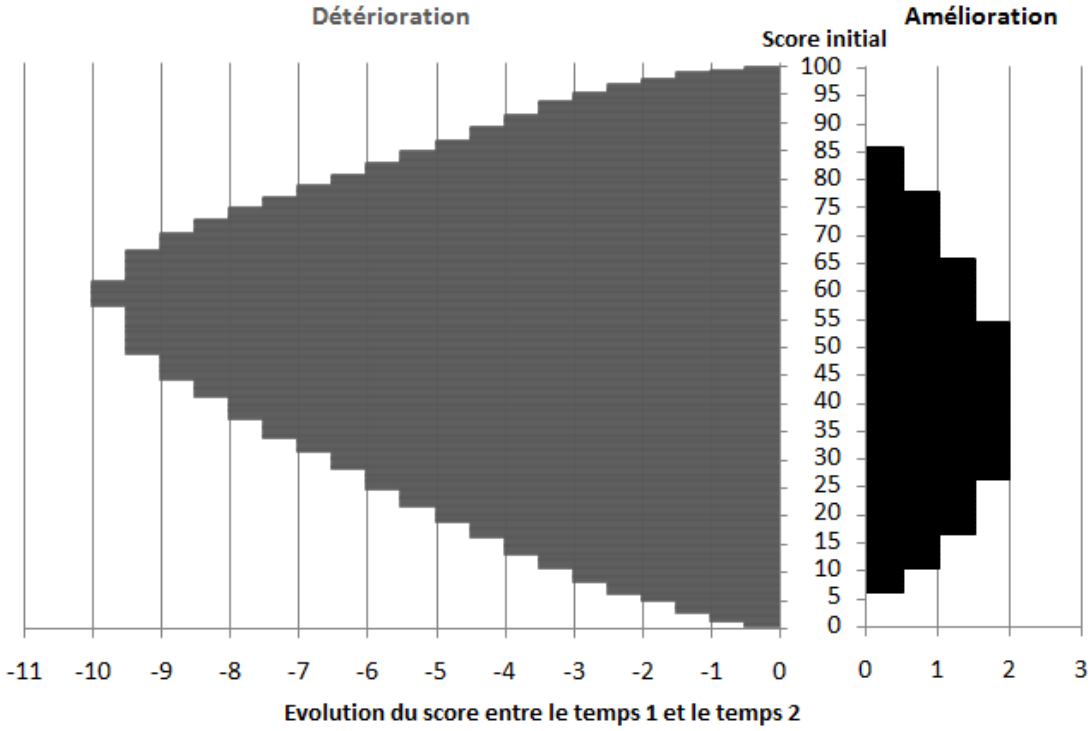


Figure 13 : Différence Minimale Cliniquement Pertinente déterminée sur l'échelle du Trait Latent dans le cas d'une amélioration (DMCP-TL = 0,0839) et d'une détérioration (DMCP-TL = -0,4806) traduite en points de score et en fonction du score initial

c. Calcul des Se, Sp, VPP et VPN

Les Se, Sp et valeurs prédictives des DMCP-Sc, DMCP-Sc_{SI} et DMCP-TL sont indiquées, dans le cas de l'amélioration et de la détérioration, dans le tableau 7. Toutes ces valeurs sauf une sont inférieures à 80%.

Tableau 7 : Performances de la Différence Minimale Cliniquement Pertinente déterminée sur le score (DMCP-Sc), sur l'échelle du trait latent (DMCP-TL) ou composée par plusieurs valeurs fonction du Score Initial (DMCP-Sc_{SI}) dans le cas de l'amélioration et de la détérioration de la santé perçue

		Se [$IC_{95\%}$]	Sp [$IC_{95\%}$]	VPP [$IC_{95\%}$]	VPN [$IC_{95\%}$]
Amélioration	DMCP-Sc	54,6% [50,7 – 58,5]	65,6% [60,9 – 70,2]	71,3% [67,2 – 75,3]	48,1% [43,9 – 52,3]
	DMCP-Sc_{SI}	56,6% [52,7 – 60,4]	68,6% [64,0 – 73,1]	73,8% [69,8 – 77,7]	50,3% [46,1 – 54,5]
	DMCP-TL	54,6% [50,7 – 58,5]	65,8% [61,2 – 70,5]	71,4% [67,4 – 75,5]	48,2% [44,0 – 52,4]
Détérioration	DMCP-Sc	44,1% [35,9 – 52,2]	65,8% [61,2 – 70,5]	31,5% [25,1 – 37,9]	76,7% [72,3 – 81,2]
	DMCP-Sc_{SI}	53,2% [45,0 – 61,3]	75,8% [71,6 – 80,0]	43,9% [36,5 – 51,3]	81,9% [78,0 – 85,9]
	DMCP-TL	51,1% [42,9 – 59,2]	63,8% [59,1 – 68,5]	33,5% [27,2 – 39,8]	78,5% [74,1 – 83,0]

$IC_{95\%}$: Intervalle de Confiance à 95%, Se : Sensibilité, Sp : Spécificité, VPP : Valeur Prédictive Positive, VPN : Valeur Prédictive Négative

4. Discussion

Cette étude a été menée dans le but d'évaluer les avantages des modèles IRT pour la détermination de la DMCP de la sous-échelle SP du questionnaire MOS-SF36 dans un échantillon de patients hospitalisés pour intervention thérapeutique dans le cadre d'une maladie chronique. Dans notre étude, l'utilisation d'un modèle IRT n'a pas permis d'obtenir des Se, Sp et valeurs prédictives supérieures pour la DMCP-TL par rapport à la DMCP-Sc, sauf dans le cas de la détérioration où sa Se et ses valeurs prédictives semblent légèrement augmentées. Pour la DMCP-Sc_{SI}, les Se, Sp et valeurs prédictives sont toutes supérieures à celles des DMCP-Sc et DMCP-TL.

L'absence d'avantage observée de l'utilisation des modèles IRT pour la détermination de la DMCP peut être expliquée par les figures 10 et 12. En effet, la relation entre le score attendu à la sous-échelle SP et le niveau sur le TL illustrée dans la figure 12 est quasi-linéaire pour un score allant de 20 à 80 environ. Cela signifie que, dans cet intervalle, l'échelle du score a quasiment les propriétés d'une échelle d'intervalle. Dans notre échantillon, comme l'indique la figure 10, 965 (82,5%) des patients avaient un SI entre 20 et 80. Peu d'erreurs de classifications entre les catégories « changement cliniquement pertinent » et « pas de changement » par la DMCP-Sc peuvent donc être expliquées par l'absence de propriétés de mesure d'intervalle de l'échelle du score. Par ailleurs, l'amplitude de la DMCP est aussi un facteur important à prendre en compte. En effet, cette amplitude est à peu près cinq fois supérieure dans le cas de la détérioration que dans celui de l'amélioration. Les performances légèrement meilleures observées dans notre étude pour la DMCP-TL dans le cas de la détérioration pourraient être dues à son amplitude. En effet, le défaut de propriétés de mesure d'intervalle de l'échelle du score pourrait être responsable de déformations plus importantes dans l'évaluation d'une grande différence que dans celle d'une petite différence : plus la quantité mesurée est grande, plus sa mesure sur l'échelle du score divergera de sa mesure sur celle du TL.

Un autre résultat important de cette étude concerne les résultats supérieurs de la DMCP-Sc_{SI}. Des recherches plus approfondies devraient être menées pour éclaircir l'origine de ce phénomène et pour déterminer s'il pourrait être dû à une perception différente du changement en fonction de la sévérité initiale. En effet, par exemple, la diminution de la DMCP avec l'augmentation du SI observée dans le cas de l'amélioration pourrait être aussi bien le résultat de l'effet plafond que du phénomène de régression vers la moyenne. De manière plus concrète, l'effet plafond découle de l'absence d'item capable de mesurer une amélioration minimale cliniquement pertinente pour les patients ayant un score déjà très haut initialement. L'évolution du score observée pour ces patients est donc inférieure à celle qui aurait été observée si aucun effet plafond n'était présent. Bien que cet effet soit plus faible que sur l'échelle du score, l'échelle du TL est aussi concernée par les effets plafond et plancher (Revicki et Cella, 1997). Ceci pourrait être une autre raison de l'absence de supériorité de la

DMCP-TL par rapport à la DMCP-Sc. Le phénomène de régression vers la moyenne est responsable d'une probabilité supérieure d'une diminution du score pour les sujets dont le SI est dans la partie supérieure de sa distribution dans l'échantillon (tendance statistique des scores extrêmes à devenir moins extrêmes au cours du suivi). Dans notre étude, la régression vers la moyenne pourrait expliquer la diminution du score (-4,8) observée en moyenne dans le sous-groupe de patients ayant un SI compris dans l'intervalle [67-100] dans le groupe amélioration (c'est-à-dire une diminution moyenne du score à la sous-échelle SP entre T1 et T2 alors que ces patients ont évalué leur état de santé à T2 comme étant meilleur qu'à T1).

Une des limites les plus largement citées dans la littérature concernant les méthodes « anchor-based » concerne la validité du critère externe utilisé (Crosby *et al.*, 2003; de Vet *et al.*, 2007; Kemmler *et al.*, 2011; Terwee *et al.*, 2010). Dans cette étude, les valeurs faibles observées pour les Se, Sp, valeurs prédictives et corrélations entre l'évolution du score de T1 à T2 et la QT posent la question de la validité de l'item du questionnaire MOS-SF36 évaluant l'évolution de la santé perçue en tant que critère externe. La validité de face de cet item pour l'évaluation de l'évolution du construit mesuré par la sous-échelle SP est pourtant bonne de manière évidente. Il existe cependant, chez les patients considérés comme stables (ayant répondu « à peu près pareil » à la QT), une diminution moyenne de leur SP qu'elle soit mesurée sur l'échelle du score (-3,16) ou sur celle du TL (-0,19). Plusieurs questions se posent à la vue de ces résultats (Terwee *et al.*, 2010) : le construit mesuré par la sous-échelle SP est-il le même que l'« état de santé » dont il est question dans la QT ? Cette QT a-t-elle encore une signification lorsqu'elle est posée pour évaluer l'état de santé six mois plus tôt (biais de rappel) ? Enfin, un phénomène de response-shift dans un ou plusieurs items de la sous-échelle SP survient-il entre T1 et T2 ? D'autres études devraient être menées pour répondre à ces questions. Une autre limite doit être discutée, celle de l'hétérogénéité des pathologies représentées dans la cohorte de sujets utilisée dans cette étude. En effet, l'utilisation d'une cohorte plus homogène concernant ces pathologies de même que l'utilisation d'un critère externe de meilleure validité auraient pu améliorer les valeurs de Se, Sp, VPP et VPN de chaque DMCP calculée dans cette étude sans, cependant, favoriser l'une des DMCP par rapport aux autres.

Ce travail est, à notre connaissance, le premier dans lequel un modèle IRT a été utilisé pour déterminer la DMCP sur l'échelle du TL. Ces modèles sont des outils puissants pour mesurer les phénomènes subjectifs sur une échelle d'intervalle. Cependant, notre étude montre que, pour la sous-échelle SP du questionnaire MOS-SF36, la capacité d'une valeur unique de DMCP à classer les sujets dans les catégories « Pas de changement » versus « Changement cliniquement pertinent » n'est pas amélioré lorsqu'elle est déterminée sur l'échelle du TL par rapport à sa détermination sur l'échelle du score. De plus, cette étude confirme les recommandations faites récemment par plusieurs auteurs sur l'utilisation de différentes valeurs de DMCP en fonction du SI (DMCP-Sc_{SI}) (Beaton *et al.*, 2011; Cook, 2008; Crosby *et al.*, 2004; Purcell *et al.*, 2010; Tubach *et al.*, 2005). Des méthodes pour déterminer le nombre de valeurs définissant la DMCP-Sc_{SI} et permettant d'obtenir les meilleures Se et Sp pour une échelle doivent être développées. Le choix de ce nombre devra évidement tenir compte du défi logistique rencontré en pratique d'avoir un trop grand nombre de valeurs, d'autant plus qu'elles sont le plus souvent différentes dans le cas de l'amélioration et de la détérioration.

5. Matériel supplémentaire

a. Le modèle de crédit partiel

Le PCM est un modèle IRT pour données polytomiques appartenant à la famille de Rasch dans lequel la probabilité que la réponse du sujet i ($i = 1, \dots N$) à l'item j ($j = 1, \dots, J$) ayant l catégories de réponse ($l = 1, \dots, m_j$) soit la catégorie y_l est modélisée comme suit :

$$P(Y_{ij} = y_l|\theta_i, \delta_{jl}) = \frac{exp\left(y_l\theta_i - \sum_{l=1}^{y}\delta_{jl}\right)}{\sum_{c=0}^{m_j} exp\left(c\theta_i - \sum_{l=1}^{c}\delta_{jl}\right)}$$

Où θ_i est le niveau sur le TL du sujet i et δ_{jl} est le paramètre d'item associé à la catégorie de réponse l de l'item j (Masters, 1982). La relation entre le niveau sur le TL et le score attendu $E(S)$ peut être calculée à l'aide de l'équation suivante (Embretson et Reise, 2000) :

$$E(S) = \sum_{j=1}^{J}\sum_{l=1}^{m_j} y_l \times P(Y_j = y_l/\theta, \delta_{jl})$$

b. Le PCM mixte longitudinal

Si θ est considéré comme une variable aléatoire ayant, par exemple, une distribution normale $N(\mu, \sigma^2)$, le PCM devient un modèle logistique à effets mixtes dans lequel les paramètres à estimer sont μ (la moyenne du niveau sur le TL dans l'échantillon), σ (l'erreur standard du niveau sur le TL dans l'échantillon) et δ_{jl} (les paramètres d'item) (Rijmen *et al.*, 2003). Pour des données répétées, une forme longitudinale du PCM à effets mixtes a été développée dans cette étude comme cela avait déjà été fait pour le modèle de Rasch par Blanchin *et al* (Blanchin *et al.*, 2010; Embretson, 1991). La probabilité d'une réponse dans la catégorie y de l'item j au temps t ($t = 1, \dots T$) est :

$$P(Y_{ij}^{(t)} = y^{(t)} | \theta_i^{(t)}, \delta_{jl}) = \frac{exp\left(y^{(t)}\theta_i^{(t)} - \sum_{l=1}^{y}\delta_{jl}\right)}{\sum_{c=0}^{m_j} exp\left(c\theta_i^{(t)} - \sum_{l=1}^{c}\delta_{jl}\right)}$$

Les paramètres d'item δ_{jl} sont supposés constants au cours du temps (i.e. l'invariance de la mesure est supposée) et peuvent être estimés ou considérés comme connus et fixes dans le modèle. Les autres paramètres à estimer sont les $\mu^{(t)}$ (la moyenne du niveau sur le TL dans l'échantillon au temps t), $\sigma^{(t)}$ (l'erreur standard du niveau sur le TL dans l'échantillon au temps t) et $\sigma^{(tt\prime)}$ (la covariance entre $\theta^{(t)}$ et $\theta^{(t\prime)}$) avec $t \neq t'$. La distribution du niveau sur le TL est supposée être une distribution multinormale de dimension T. La moyenne de l'évolution du construit mesuré par le questionnaire au cours du temps dans l'échantillon peut être évaluée par l'évolution de $\mu^{(t)}$ entre chaque temps t.

Des variables dichotomiques de groupe (G ayant pour réalisation g_i pour le sujet i) peuvent être introduites dans ce modèle afin d'évaluer l'effet temps dans différents groupes de l'échantillon:

$$P(Y_{ij}^{(t)} = y^{(t)} | \theta_i^{(t)}, \delta_{jl}, \alpha, \beta^{(t)}) = \frac{exp\left(y^{(t)}\theta_i^{(t)} + \boldsymbol{\alpha g_i} + \boldsymbol{\beta^{(t)} g_i} - \sum_{l=1}^{y}\delta_{jl}\right)}{\sum_{c=0}^{m_j} exp\left(c\theta_i^{(t)} + \boldsymbol{\alpha g_i} + \boldsymbol{\beta^{(t)} g_i} - \sum_{l=1}^{c}\delta_{jl}\right)}$$

Avec la contrainte d'identification $\beta^{(1)} = 0$ dans ce modèle, α représente la différence moyenne entre les deux groupes au temps initial ($t = 1$) et $\alpha + \beta^{(t)}$ représente la différence

moyenne entre les deux groupes au temps t. Les moyennes du niveau sur le TL à chaque temps t et t' et dans chaque groupe sont indiquées dans le tableau 8. Si, par exemple, la variable G indique la réponse à la QT (0 indiquant " A peu près pareil " et 1 indiquant " Un peu meilleur "), la DMCP-TL pour l'amélioration est égale à l'évolution moyenne du niveau sur le TL entre le temps t et le temps t' dan le groupe $G = 1$, c'est à dire $(\mu^{(t')} - \mu^{(t)}) + (\beta^{(t')} - \beta^{(t)})$.

Tableau 8 : Moyenne du niveau sur le Trait Latent (TL) à chaque temps t et t' et dans chaque groupe G lorsqu'une variable dichotomique de groupe est introduite dans le modèle de crédit partiel mixte longitudinal (par exemple, G indique la réponse à la question de transition : 0 indiquant "A peu près pareil " et 1 indiquant "Un peu meilleur ")

	Groupe	
	$G = 0$	$G = 1$
Temps t	$\mu^{(t)}$	$\mu^{(t)} + \alpha + \beta^{(t)}$
Temps t'	$\mu^{(t\prime)}$	$\mu^{(t\prime)} + \alpha + \beta^{(t\prime)}$
Evolution moyenne du niveau sur le TL (effet temps)	$\mu^{(t\prime)} - \mu^{(t)}$	$\left(\mu^{(t\prime)} - \mu^{(t)}\right) + \left(\beta^{(t')} - \beta^{(t)}\right)$

CHAPITRE 4 : INTERETS DES MODELES ISSUS DE LA THEORIE DE REPONSE A L'ITEM DANS LES ANALYSES SUR DONNEES LONGITUDINALES : UNE ETUDE PAR SIMULATION

Article en cours de finalisation, présentation orale lors du congrès suivant :

Rouquette A, Côté S, Hardouin JB, Falissard B. Item Response Theory (IRT) used to enhance accuracy of data analyses in longitudinal studies of child development: a simulation study. Society for Research in Child Development, 2014 special topic meeting: Developmental Methodology. September 11-13, 2014, San Diego, CA, USA.

Contribution des co-auteurs :

Alexandra Rouquette : Conception de l'étude, revue de la littérature, programmation des analyses statistiques, interprétation des résultats, rédaction, soumission et révision de l'article.

Sylvana Côté: Contribution à l'interprétation des résultats, à la révision de l'article.

Jean-benoit Hardouin : Contribution à la programmation des analyses statistiques, à l'interprétation des résultats

Bruno Falissard : Conception de l'étude, contribution à la programmation des analyses statistiques, à l'interprétation des résultats, à la rédaction et révision de l'article

Résumé

Lors des études de cohorte, un problème rencontré fréquemment est celui du choix des items à utiliser pour étudier l'évolution d'un phénomène subjectif au cours du temps. En effet, la plupart des analyses pour données longitudinales, par exemple l'analyse en classe latente de trajectoires (LCGA), sont appliquées sur le score, somme des réponses aux items à chaque temps. Dans ce cas, les items utilisés doivent être identiques à tous les temps et les items absents à certains temps ne sont pas utilisés entrainant ainsi une perte d'information. Dans le modèle de Rasch, modèle issu de la théorie de réponse à l'item (IRT), la propriété d'objectivité spécifique permet d'évaluer un même construit sur l'échelle du trait latent (TL) à partir d'ensembles différents d'items. L'utilisation du modèle de Rasch pourrait donc améliorer la précision de la mesure du phénomène en utilisant l'ensemble des items disponibles même s'ils sont absents à certains temps. L'objectif de cette étude était de comparer les performances de la LCGA appliquée sur le score (LCGA-Sc) ou sur l'estimation du TL par un modèle de Rasch (LCGA-θ_{est}) en fonction du nombre d'items absents à certains temps. Cent cinquante bases de données ont été simulées correspondant à celles d'une cohorte composée de trois sous-groupes ayant différentes trajectoires moyennes d'évolution du construit évalué par 10 items à quatre temps de collecte. Quatre configurations concernant le nombre et la difficulté des items absents à certains temps ont été étudiées. L'efficacité de la LCGA était évaluée par le pourcentage de sujets bien classés en prenant comme référence les sous-groupes simulés. Une meilleure efficacité de la LCGA-Sc était retrouvée lorsqu'aucun item n'était absent mais, alors que l'efficacité de la LCGA-θ_{est} restait stable, celle de la LCGA-Sc diminuait avec le nombre et la difficulté des items disponibles pour le calcul du score.

Mots-clefs : Trajectoires, Cohorte, Longitudinal, Questionnaire, Modèle de Rasch, Simulations

1. Introduction

Lors de la mise en place d'une étude de cohorte, le choix des questionnaires validés les plus adaptés aux phénomènes subjectifs à mesurer doit être fait dès la phase de planification (Rothman *et al.*, 2008). Cependant, au cours du suivi, en particulier lorsqu'il dure plusieurs dizaines d'années, certains questionnaires peuvent subir des améliorations, voire être substitués par un autre questionnaire considéré comme plus adapté que celui choisi initialement. Le phénomène subjectif étudié est alors mesuré à l'aide d'un instrument différent en fonction du temps de collecte. Classiquement, la mesure du phénomène est obtenue en additionnant les réponses (cotées, pondérées ou non) aux items du questionnaire et est appelée le score. Tant que l'instrument reste identique, ce score est comparable d'un temps à l'autre et il est donc possible d'étudier l'évolution du niveau des sujets pour le phénomène subjectif estimé par ce score au cours du temps. En revanche, si l'instrument change à certains temps de collecte, les scores obtenus à l'aide de ces différents instruments ne sont plus comparables.

Cette situation est particulièrement fréquente dans le domaine de l'épidémiologie développementale. Cette approche intègre les principes, théories et méthodes de la psychologie développementale dans la recherche épidémiologique afin d'éclaircir les mécanismes selon lesquels les processus développementaux affectent les risques de survenue de problèmes de santé (Costello *et al.*, 2006; Pillemer et White, 2005). L'étude de ces processus nécessite l'analyse de données issues de cohortes suivies au cours de stades du développement tels que la petite enfance, l'enfance, l'adolescence, etc. Cependant, l'expression d'un même phénomène subjectif est rarement identique au cours de ces différents stades. Par exemple, les symptômes dépressifs chez l'enfant prennent souvent la forme de symptômes somatiques ou d'irritabilité alors que chez l'adolescent l'abus de substance ou l'hypersomnie peuvent être des symptômes de dépression (Ryan ND *et al.*, 1987). L'évaluation des phénomènes subjectifs doit donc être adaptée au cours du suivi et prend la forme de questionnaires dont certains items sont différents, absents ou ajoutés lors de certaines périodes développementales.

Dans de tels cas, la pratique habituelle pour calculer un score comparable d'un temps à un autre est de n'utiliser que les items présents à tous les temps de collecte. L'information apportée par les items spécifiques de certains stades développementaux n'est donc pas prise en compte. Par exemple, dans la figure 14, est schématisée une étude hypothétique à trois temps de collecte où le même phénomène subjectif est mesuré à l'aide d'un ensemble d'items différent à chaque temps. Le calcul d'un score comparable au cours du temps ne pourra donc concerner que les items 2, 3 et 5 présents à tous les temps. L'absence de prise en compte de l'information apportée par les réponses aux items 1, 4, 6, 7 et 8, posés à certains temps seulement, pourrait être responsable d'une perte de précision à la fois de l'estimation du phénomène étudié et de l'ensemble des paramètres estimés dans les modèles statistiques l'utilisant. La validité des inférences issues de ce type d'études pourrait donc en être affaiblie.

Temps de collecte		T1	T2	T3
Mesure du phénomène subjectif Items binaires 0=Non, 1=Oui		Item 1 **Item 2** **Item 3** Item 4 **Item 5**	Item 1 **Item 2** **Item 3** **Item 5** Item 6 Item 7	 **Item 2** **Item 3** **Item 5** Item 7 Item 8
Calcul du SCORE	Somme des réponses aux items 2, 3 et 5	**[0 ; 3]**	**[0 ; 3]**	**[0 ; 3]**
Estimation du niveau sur le TRAIT LATENT	Modèle de Rasch tenant compte de tous les items disponibles	]-∞ ; +∞[	]-∞ ; +∞[	]-∞ ; +∞[

Figure 14 : Schématisation d'une étude à trois temps de collecte où le même phénomène subjectif est mesuré à l'aide d'ensembles d'items dichotomiques différents à chaque temps. Le score est calculé avec les items en gras et son score varie entre 0 et 3. L'estimation du niveau sur le trait latent par un modèle de Rasch utilise l'information apportée par l'ensemble des items

Les modèles issus de la théorie de réponse à l'item (IRT : Item Response Theory) sont des modèles permettant d'évaluer un phénomène subjectif (appelé le Trait Latent, TL) dont la mesure est exprimée sur une échelle d'intervalle, non bornée et dont le zéro est arbitrairement fixé. Le plus simple et le plus connu de ces modèles est le modèle de Rasch qui modélise la probabilité qu'un individu i $(i = 1, \dots N)$ réponde positivement à l'item j $(j = 1, \dots J)$, en fonction de son niveau sur le TL (θ_i) et d'un paramètre spécifique de l'item appelé difficulté d'item (δ_j) :

$$P(Y_{ij} = 1|\theta_i, \delta_j) = \frac{\exp(\theta_i - \delta_j)}{1 + \exp(\theta_i - \delta_j)}$$

Le modèle contient autant d'équations qu'il y a d'items dans le questionnaire et c'est grâce à l'ensemble des réponses de l'individu i à tous ces items que son niveau sur le TL, θ_i, pourra être estimé. La difficulté d'item représente le niveau sur le TL qu'un sujet doit atteindre pour avoir 50% de chance de répondre positivement à cet item et c'est grâce à l'ensemble des réponses des N individus à l'item j que δ_j va pouvoir être estimé (Embretson et Reise, 2000).

Dans cette théorie, l'échelle du TL est identique quelles que soient la population étudiée ou les conditions de la mesure (Embretson et Reise, 2000; Rupp et Zumbo, 2006). Les modèles IRT de la famille de Rasch ont la propriété d'objectivité spécifique. Cette propriété implique que, l'erreur d'échantillonnage mise à part, les estimations des paramètres d'items obtenues dans différents échantillons ou à différentes occasions de mesure sont identiques. De même, l'erreur de mesure mise à part, les estimations des niveaux sur le TL des individus, $\widehat{\theta_\iota}$, obtenues à partir de groupes d'items différents sont identiques (Hambleton *et al.*, 1991). Ainsi, dans l'exemple schématisé dans la figure 14, l'échelle du TL est identique à chaque temps, la présence d'items communs entre les différents temps permettant l'étalonnage de l'échelle. Les $\widehat{\theta_\iota}$ estimés par un modèle de Rasch à partir des différents ensembles d'items disponibles à chaque temps seront donc comparables d'un temps à l'autre.

L'hypothèse faite dans cette étude est que l'utilisation d'un modèle de Rasch permettra un gain de précision sur l'évaluation longitudinale d'un phénomène subjectif par rapport à l'utilisation du score lorsque certains items disponibles pour cette évaluation diffèrent d'un

temps à l'autre. Si cette hypothèse est vraie, les analyses statistiques pour données longitudinales utilisant cette évaluation devraient être plus performantes lorsqu'appliquées sur les $\widehat{\theta_i}$, estimés par un modèle de Rasch plutôt que sur le score, en particulier lorsque le nombre d'items disponibles pour le calcul du score est faible comparativement au nombre d'items disponibles pour l'estimation des $\widehat{\theta_i}$.

Ce travail a donc pour objectif de comparer, à partir de données simulées, les performances d'une méthode d'analyse pour données longitudinales appliquée sur les $\widehat{\theta_i}$ estimées par un modèle de Rasch ou sur le score, en fonction du nombre d'items disponibles pour le calcul du score. L'analyse en classe latente de trajectoires (Latent Class Growth Analysis, LCGA) est la méthode d'analyse pour données longitudinales qui a été choisie pour tester cette hypothèse (Berlin *et al.*, 2014; Jung et Wickrama, 2008; Muthén et Muthén, 2000; Nagin, 2005). Ce type d'analyse a pour but de détecter, parmi l'ensemble des trajectoires individuelles présentes dans la cohorte, des groupes de sujets ayant des trajectoires de forme homogène (croissantes, stables, décroissantes, etc.). Très utilisée en épidémiologie développementale, la LCGA est de plus en plus couramment rencontrée dans d'autres branches de l'épidémiologie car elle permet de mettre en évidence des trajectoires-types d'évolution du phénomène (ex. niveau d'anxiété, douleur chronique) et de les mettre en relation avec des facteurs environnementaux, biologiques, sociodémographique, etc. (Galera *et al.*, 2012; Jones *et al.*, 2012; Pryor *et al.*, 2011; Seegers *et al.*, 2011; Tu *et al.*, 2013; Walton *et al.*, 2014; Yeates *et al.*, 2009).

2. Méthodes

Afin d'évaluer les performances de la LCGA utilisant les $\widehat{\theta_i}$ estimés par un modèle de Rasch (LCGA-θ_{est}) ou utilisant le score (LCGA-Sc), une étude par simulation de Monte-Carlo a été utilisée. Le principe de ce type d'étude est de simuler des données selon un scénario choisi, c'est-à-dire dont les paramètres sont connus, puis d'appliquer la méthode statistique à évaluer sur ces données afin de déterminer ses capacités à retrouver le scénario ayant servi à les simuler. Dans cette étude, le scénario choisi correspondait à une situation type rencontrée en épidémiologie développementale : une étude de cohorte dans laquelle un même

phénomène subjectif unidimensionnel est évalué à quatre temps de collecte à l'aide d'un questionnaire comportant au maximum 10 items dichotomiques. Trois groupes de même taille constituent cette cohorte : un groupe dont la trajectoire moyenne de l'évolution du phénomène subjectif au cours du temps est de forme stable et de niveau bas (groupe « bas »), un autre dont la trajectoire est de forme stable et de niveau haut (groupe « haut ») et un dernier groupe où la trajectoire est croissante au cours du temps (groupe « croissant »). Les performances de la LCGA seront jugées sur sa capacité à retrouver ces trois groupes de trajectoire dans les données simulées.

a. Le modèle de simulation

Pour simuler des données longitudinales correspondant à cette situation type, un modèle de Rasch longitudinal a été utilisé. Dans ce modèle, la probabilité d'une réponse positive du sujet i ($i = 1$ à N) à l'item j ($j = 1$ à J), au temps t ($t = 1$ à T) s'écrit :

$$P\left(Y_{ij}^{(t)} = 1 | \theta_i^{(t)}, \delta_j\right) = \frac{\exp\left(\theta_i^{(t)} - \delta_j\right)}{1 + \exp\left(\theta_i^{(t)} - \delta_j\right)}$$

Avec $\theta_i^{(t)}$, le niveau sur le TL de l'individu i au temps t dont la distribution est supposée être multinormale de dimension T avec $\mu^{(t)}$ et $\sigma^{(t)}$ la moyenne et l'écart-type du niveau sur le TL dans l'échantillon au temps t respectivement et $\sigma^{(tt')}$, la covariance entre $\theta^{(t)}$ et $\theta^{(t')}$ avec $t \neq t'$. La difficulté de l'item j, δ_j est supposée constante au cours du temps dans ce modèle (i.e. l'invariance temporelle de la mesure est supposée) (Blanchin *et al.*, 2010).

b. Le choix des paramètres de simulation

Plusieurs paramètres peuvent être contrôlés dans ce modèle afin de simuler le scénario souhaité. Dans cette étude, le nombre de temps de collecte T a été fixé à 4 et le nombre d'items J à 10. Pour simuler trois groupes de trajectoire différents, un modèle de simulation ayant des valeurs différentes pour $\mu^{(t)}$ a été utilisé dans chacun de ces trois groupes dont la taille a été fixée à 1000 sujets (donc $N = 3000$). Dans les groupes « bas » et « haut », les $\mu^{(t)}$ étaient fixés à -1 et 1 respectivement quel que soit t et dans le groupe « croissant », les moyennes du TL à chaque temps étaient : $\mu^{(1)} = -0{,}5$; $\mu^{(2)} = 0$; $\mu^{(3)} = 0{,}5$ et $\mu^{(4)} = 1$. La

trajectoire moyenne de chacun des trois groupes sur l'échelle du TL ainsi obtenue est représentée dans la figure 15.

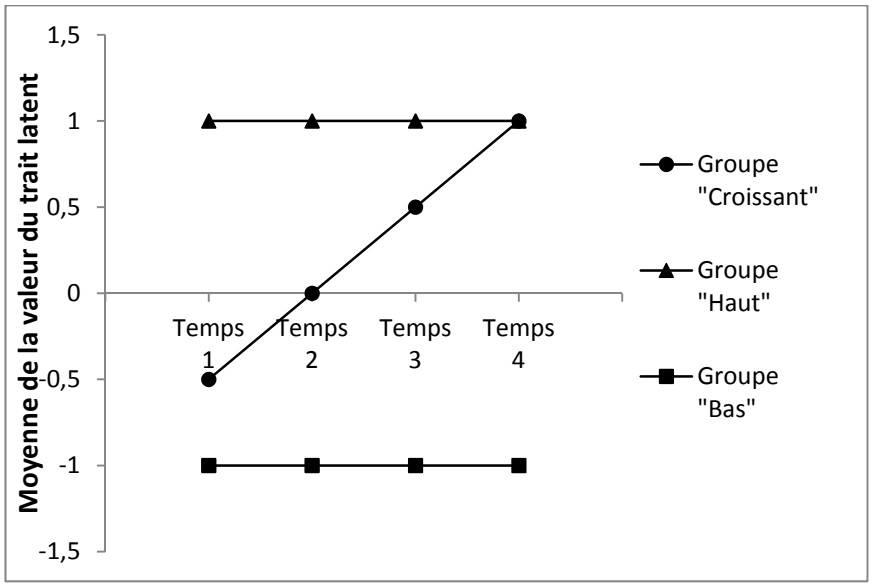


Figure 15 : Trajectoires moyennes sur l'échelle du trait latent dans chacun des trois groupes composant les échantillons simulés

Les variances $\sigma^{2(t)}$ et covariances $\sigma^{(tt\prime)}$ du TL ont été supposées identiques dans les trois groupes. Une LCGA sur les $\widehat{\theta_i}$ estimés par un modèle de Rasch a été appliquée sur des données issues de l'Etude Longitudinale des Enfants de Maternelle au Québec (ELEMQ) afin d'évaluer la valeur de ces paramètres dans des données réelles (Rouquette *et al.*, 2014). Les données utilisées concernaient l'évolution de la dimension des troubles internalisés (items d'anxiété et de dépression principalement) mesurée par neuf items administrés aux parents de 2000 enfants représentatifs de la population Québécoise aux âges de 6, 8, 10 et 12 ans. Sur ces données, la moyenne de la variance du TL sur les quatre temps de collecte était de 0,304. Concernant les corrélations entre les $\theta^{(t)}$, la moyenne des trois corrélations entre deux temps adjacents était de 0,903, la moyenne des deux corrélations corr($\theta^{(6\,ans)}$, $\theta^{(10\,ans)}$) et corr($\theta^{(8\,ans)}$, $\theta^{(12\,ans)}$) était de 0,740, enfin, corr($\theta^{(6\,ans)}$, $\theta^{(12\,ans)}$) était égale à 0,583. Ainsi, dans le modèle de simulation, $\sigma^{2(t)}$ a été fixé à 0,3 quel que soit t ; les corrélations entre les TL de deux temps adjacents ont été fixées à 0,8 ; la corrélation entre le 1er et le 3ème temps a

été fixée à 0,7 de même que la corrélation entre le 2ème et le 4ème temps ; enfin, la corrélation entre le 1er et le 4ème temps a été fixée à 0,6.

Concernant les difficultés d'items, étant données les $\mu^{(t)}$ et $\sigma^{(t)}$ choisies pour chaque groupe de trajectoire, il a été décidé de prendre les valeurs des déciles de la fonction de distribution d'une loi normale $\mathcal{N}(0,1)$ comme valeurs pour les δ_j. Ainsi, comprises entre -1,34 et 1,34, les δ_j couvraient l'étendue des valeurs prises par les $\theta_i^{(t)}$ dans les échantillons simulés. Les difficultés des items formant le questionnaire étaient de cette manière adaptées au niveau du phénomène évalué dans la population étudiée.

c. Les analyses des bases de données simulées et les critères de jugement de leur performance

Une fois les paramètres de ce modèle fixés, 150 bases de données ont été simulées. Les valeurs contenues dans ces bases étaient pour chacun des 3000 sujets : les réponses aux dix items codées 0 ou 1 à chaque temps, le groupe de trajectoire (« bas », « haut », « croissant ») et la valeur du TL simulée par le programme à chaque temps ($\theta_{sim}^{(t)}$) afin de prédire les réponses de chaque sujet aux 10 items. Quatre variables ont été ajoutées dans chacune des bases : le score à chaque temps ($S^{(t)}$), simple somme des réponses aux items.

Une LCGA, avec un nombre de classes fixé à 3, a été appliquée sur les variables $S^{(t)}$ de chacune de ces 150 bases. Le diagramme de chemin correspondant à cette LCGA-Sc est représenté dans la partie gauche de la figure 16. La classe latente C a trois catégories indiquant les groupes de trajectoire latente homogène (d'ordonnée à l'origine O et de pente P) détectés dans l'échantillon par la LCGA-Sc. La forme de la trajectoire latente a été imposée linéaire dans ce modèle comme indiqué par les charges factorielles reliant la pente P aux scores $S^{(t)}$. Sur ces bases, une LCGA, avec un nombre de classes fixé à 3, a aussi été appliquée sur les estimations des TL ($\widehat{\theta_i^{(t)}}$) par un modèle de Rasch à partir des réponses aux 10 items observées à chaque temps t. Son diagramme de chemin est représenté dans la partie droite de la figure 16 (LCGA-θ_{est}). Enfin, une LCGA, avec un nombre de classes fixé à 3, a été appliquée sur les valeurs du TL simulées à chaque temps ($\theta_i^{(t)}$) dans chacune de ces bases (LCGA-θ_{sim}). Le but

de cette dernière analyse, possible car dans le cadre d'une étude par simulation, était d'évaluer les performances de la LCGA appliquée sur le « vrai » niveau des sujets sur le TL, c'est-à-dire sans l'erreur de mesure inévitable lorsque ce niveau est estimé, par le calcul du score ou par un modèle IRT, à partir des items observés.

Les résultats d'une LCGA se composent d'une variable par classe indiquant, pour chaque sujet, sa probabilité d'appartenir à la classe en question. La classe d'appartenance la plus probable pour chacun des sujets a été récupérée dans chaque base de données pour la LCGA-Sc, la LCGA-θ_{est} et la LCGA-θ_{sim}. L'identification de la classe correspondant à chaque groupe de trajectoire simulé a été effectuée en repérant la classe la plus représentée dans chacun de ces groupes. Ainsi, le critère de jugement principal de la performance des trois LCGA a été calculé comme la moyenne du pourcentage de sujets bien classés sur les 150 simulations en prenant comme référence le groupe de trajectoire simulé.

En prenant la classe la plus probable comme classement des sujets suite à la LCGA, l'incertitude inhérente à cette méthode de classification est ignorée. Un indice mesurant l'incertitude globale de la classification par la LCGA (i.e. le degré de séparation des différentes classes), l'entropie relative, a été relevé lors de chaque analyse (Celeux et Soromenho, 1996; Jung et Wickrama, 2008). Cet indice varie de 0 à 1 et plus il se rapproche de 1, plus il indique une faible incertitude sur le classement. Un critère de jugement secondaire de la performance des LCGA était la moyenne de l'entropie relative sur les 150 bases simulées.

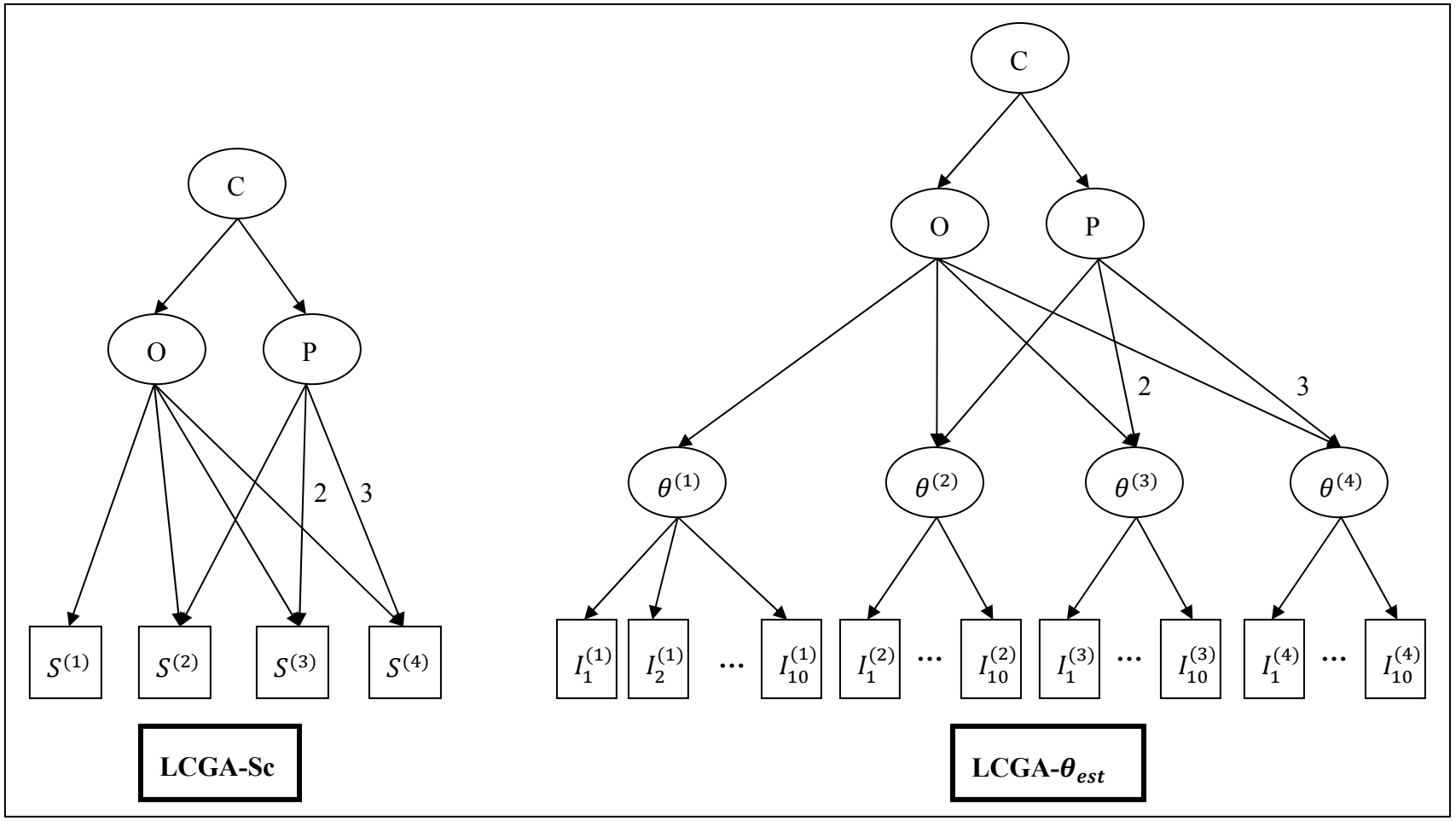


Figure 16 : Diagramme de chemin représentant une analyse en classe latente de trajectoires (Latent Class Growth Analysis, LCGA) utilisant 1/ LCGA-Sc : le score observé aux quatre temps t ($S^{(t)}$) ; 2/ LCGA-θ_{est} : le trait latent ($\theta^{(t)}$) estimé par un modèle de Rasch sur les 10 items j ($I_j^{(t)}$) observés à chaque temps t. O représente l'ordonnée à l'origine et P la pente de la trajectoire latente dans l'échantillon. C est la classe latente à trois catégories indiquant les trois groupes de trajectoire latente homogène. Les charges factorielles sont fixées à 1 sauf lorsque indiquées autrement.

d. Les différents scénarios d'items absents étudiés

Les critères de jugement de la performance de la LCGA-Sc et de la LCGA-θ_{est} ont été étudiés dans quatre configurations différentes concernant les items absents à certains temps. Ces configurations sont schématisées dans la figure 17. La configuration « Complet » était celle où l'ensemble des 10 items étaient disponibles pour le calcul du score et pour l'estimation du niveau sur le TL. Dans la configuration « 7 items difficiles », trois des items faciles n'étaient pas disponibles à tous les temps de collecte ; le score ne pouvait donc être calculé que sur les sept items difficiles présents à tous les temps alors que l'estimation du niveau sur le TL par un modèle de Rasch tenait compte de l'information supplémentaire fournie par les items 1, 2 et 4 disponibles à certains temps seulement. La configuration « 7 items faciles » était la même que la précédente sauf que les items absents à certains temps étaient des items difficiles. Enfin, dans la dernière configuration, « 4 items », seuls quatre items étaient disponibles pour le calcul du score.

Pour chacune des 150 bases de données simulées, après avoir appliqué les LCGA sur les trois mesures du phénomène subjectif disponibles (LCGA-Sc, LCGA-θ_{est} et LCGA-θ_{sim}) dans le cas « Complet », une copie des bases a été effectuée dans laquelle les réponses des sujets à l'item 1 au temps 2, l'item 2 aux temps 3 et 4 et l'item 4 au temps 1 ont été effacées. Les quatre variables $S^{(t)}$ ont été recalculées avec les sept items de la configuration « 7 items difficiles » puis la LCGA-Sc et la LCGA-θ_{est}ont été appliquées. Le pourcentage moyen de sujets bien classés et l'entropie moyenne ont ainsi pu être calculés pour cette configuration. La même démarche a été suivie pour les configurations « 7 items faciles » et « 4 items ».

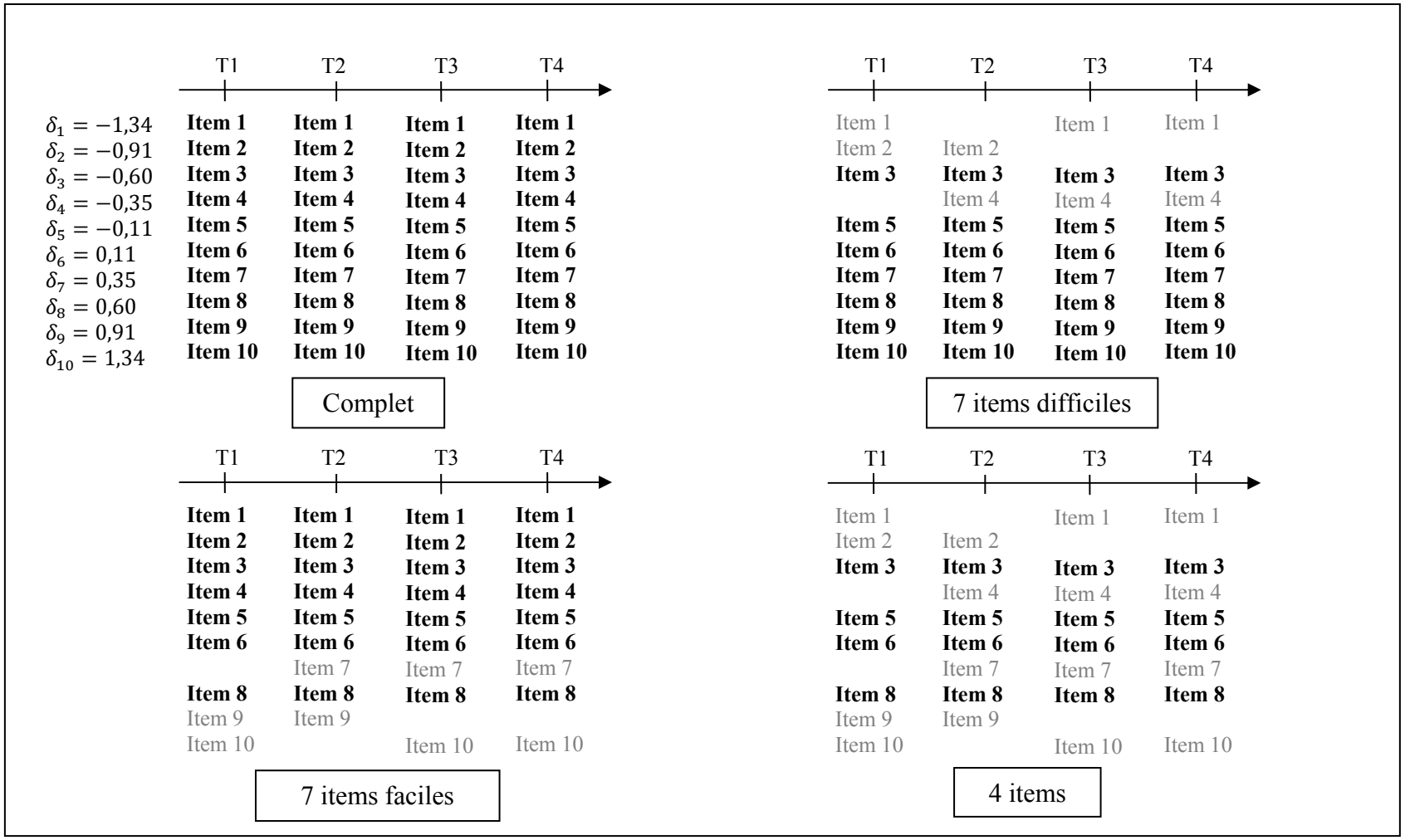


Figure 17 : Les quatre scenarios d'items absents étudiés. Les items disponibles pour calculer le score à chaque temps T de collecte sont en gras. Les items supplémentaires disponibles pour évaluer le trait latent sont en gris. La difficulté δ_j de l'item j était fixée dans le modèle de simulation.

e. Logiciels utilisés

Le programme simirt du logiciel Stata© v.12 a été utilisé pour la simulation des bases de données (Hardouin, 2005; StataCorp, 2012). Les LCGA-θ_{sim}, LCGA-Sc et LCGA-θ_{est} ont été appliquées sur ces bases de données grâce au logiciel Mplus© v.7 (Muthén et Muthén, 2012). Enfin, le package MplusAutomation du logiciel R v.3.1.0 a été utilisé pour automatiser l'application des LCGA et l'extraction des résultats via le logiciel Mplus sur les 600 (4 fois 150) bases de données analysées (Hallquist et Wiley, 2013; R Development Core Team, 2008).

3. Résultats

Le pourcentage moyen de sujets bien classés par LCGA appliquée sur chacune des mesures du phénomène subjectif disponibles et dans les quatre configurations étudiées est indiqué dans le tableau 9. Lorsqu'elle est appliquée sur θ_{sim}, la LCGA permet de retrouver le bon groupe de trajectoire pour 82,2% [82,1 – 82,4] des sujets de l'échantillon en moyenne. Le θ_{sim}est le niveau sur le TL sans erreur de mesure car c'est la valeur qui a été utilisée par le modèle de simulation pour générer les réponses aux items ; ce pourcentage est donc une estimation de la performance maximale de la LCGA sur ce type de données.

Quelle que soit la configuration, la LCGA appliquée sur les mesures du phénomène subjectif estimées à partir des réponses aux items, les scores ou les $\widehat{\theta_i}$, ne permet pas d'atteindre un tel pourcentage. Lorsque les 10 items sont disponibles pour le calcul du score et l'estimation des $\widehat{\theta_i}$, le pourcentage moyen de sujets bien classés est supérieur (77,4% [77,2 – 77,6]) pour la LCGA-Sc que pour la LCGA-θ_{est} (75,6% [75,4 – 75,8]). En revanche, lorsque certains items sont absents pour le calcul du score, ce résultat peut s'inverser. Par exemple, dans la configuration « 4 items », 72,8% [72,6 – 72,9] des sujets étaient bien classés par la LCGA-Sc en moyenne alors que ce pourcentage était de 74,2% [74,0 – 74,5] pour la LCGA-θ_{est}. La difficulté des items présents pour le calcul du score avait aussi une influence sur la performance des LCGA. En effet, lorsqu'il restait sept items difficiles pour calculer le score,

la LCGA-Sc était plus performante que la LCGA-θ_{est} alors que si les sept items disponibles pour calculer le score étaient faciles, la LCGA-θ_{est}était plus performante que la LCGA-Sc.

Tableau 9 : Moyenne du pourcentage de sujets bien classés (et son intervalle de confiance à 95%) par analyse en classe latente de trajectoires appliquée sur le score (LCGA-Sc), sur le niveau sur le trait latent estimé par modèle de Rasch (LCGA-θ_{est}) et sur le niveau sur le trait latent utilisé par le programme de simulation (LCGA-θ_{sim}) sur les 150 bases simulées, dans chacune des quatre configurations étudiées

Configuration	LCGA-Sc	LCGA-θ_{est}	LCGA-θ_{sim}
Complet	77,42 [77,23 – 77,62]	75,60 [75,38 – 75,81]	82,24 [82,06 – 82,43]
7 items difficiles	75,46 [75,27 – 75,65]	74,19 [73,97 – 74,40]	
7 items faciles	72,86 [72,60 – 73,12]	74,14 [73,91 – 74,37]	
4 items	72,75 [72,57 – 72,92]	74,23 [74,01 – 74,45]	

La moyenne de l'entropie sur les 150 simulations dans chacune des configurations étudiées est indiquée pour la LCGA-Sc, la LCGA-θ_{est}et la LCGA-θ_{sim}dans le tableau 10. Les résultats concordaient avec ceux obtenus lorsque le critère de jugement principal était le pourcentage moyen de sujets bien classés. Lorsque la LCGA était appliquée sur le score ou sur les $\widehat{\theta_i}$, l'entropie moyenne n'atteignait jamais celle obtenue avec la LCGA-θ_{sim}. Si les 10 items étaient disponibles pour le calcul du score, l'entropie moyenne était meilleure pour la LCGA-Sc par rapport à celle de la LCGA-θ_{est}. En revanche, lorsque le nombre d'items diminuait, en particulier lorsqu'il ne restait que des items faciles pour calculer le score, l'entropie moyenne de la LCGA-θ_{est} était plus importante.

Tableau 10 : Moyenne de l'entropie (et son intervalle de confiance à 95%) des analyses en classe latente des trajectoire appliquées sur le score (LCGA-Sc), sur le niveau sur le trait latent estimé par modèle de Rasch (LCGA-θ_{est}) et sur le niveau sur le trait latent utilisé par le programme de simulation (LCGA-θ_{sim}) sur les 150 bases simulées, dans chacune des quatre configurations étudiées

Configuration	LCGA-Sc	LCGA-θ_{est}	LCGA-θ_{sim}
Complet	0,775 [0,773 – 0,776]	0,755 [0,753 – 0,756]	0,909 [0,908 – 0,910]
7 items difficiles	0,734 [0,732 – 0,735]	0,733 [0,731 – 0,734]	
7 items faciles	0,723 [0,721 – 0,724]	0,732 [0,730 – 0,733]	
4 items	0,647 [0,645 – 0,650]	0,715 [0,713 – 0,716]	

4. Discussion

En prenant l'exemple d'une situation type rencontrée en épidémiologie développementale, cette étude de simulation a comparé les performances d'une même méthode d'analyse de données longitudinales, la LCGA, appliquée sur deux échelles différentes de mesure d'un même phénomène subjectif : le score et le TL. L'hypothèse était que la LCGA-θ_{est} a de meilleures performances que la LCGA-Sc, en particulier lorsqu'il existe des items supplémentaires pour estimer le TL, impossibles à utiliser pour le calcul du score car absents à certains temps. Les résultats de cette étude confortent cette hypothèse. En effet, dans la configuration « 4 items » où le score n'était calculé qu'à partir de quatre items alors que l'estimation du TL par le modèle de Rasch tenait compte de l'information apportée par des items supplémentaires présents à certains temps, la LCGA-θ_{est} avait, comme attendu, un pourcentage moyen de sujets bien classés et une entropie moyenne supérieurs à ceux de la LCGA-Sc. Cette meilleure performance de la LCGA-θ_{est} était aussi observée lorsque sept items étaient disponibles pour le calcul du score mais seulement si les difficultés des items supplémentaires pour l'estimation du TL étaient supérieures à celles des items utilisés pour le

score. En revanche, dans les configurations « Complet » ou « 7 items difficiles » les performances de la LCGA-Sc étaient meilleures.

Dans les conditions de simulation des données de cette étude, les résultats obtenus dans la configuration « Complet » ne sont pas surprenants. En effet, le modèle de simulation de données était basé sur le modèle de Rasch qui repose sur des hypothèses strictes mais conférant au score des qualités psychométriques équivalentes à celles du TL. L'exhaustivité du score sur le TL est l'une des propriétés du modèle de Rasch et signifie qu'à chaque valeur du score, quelle que soit la manière dont il est calculé, correspond une et une seule valeur du TL (Embretson, 1991). Cette propriété n'est pas respectée par les modèles IRT n'appartenant pas à la famille des modèles de Rasch sauf si le calcul du score tient compte de certains paramètres d'item à l'aide de pondérations. En pratique, de nombreux questionnaires ne répondent pas aux hypothèses strictes du modèle de Rasch et il est possible qu'en simulant des données à l'aide d'un autre modèle que celui de Rasch, les meilleures performances de la LCGA-Sc ne soient pas observées dans cette configuration. Par ailleurs, la propriété d'intervalle de l'échelle du TL est retrouvée dans l'échelle du score lorsque les difficultés des items répondant à un modèle de Rasch sont régulièrement espacées sur l'échelle du TL. Les valeurs des difficultés des dix items fixées dans le modèle de simulation étaient les dixièmes percentiles d'une loi normale, ce choix permettant de couvrir l'ensemble des niveaux sur le TL retrouvés dans l'échantillon étudié. Cependant, ces difficultés n'étant pas régulièrement espacées sur le TL, le défaut de propriété d'intervalle de l'échelle du score engendré est probablement une des raisons pour laquelle les performances de la LCGA-Sc ne sont pas plus proches des performances maximales possibles de la LCGA sur ces données (celles de la LCGA-θ_{sim}) dans cette configuration.

Ces résultats indiquent donc que, dans les conditions simulées, lorsque les données sont en adéquation avec un modèle de Rasch et qu'il y a autant d'items disponibles pour calculer le score ou estimer le TL, la LCGA-Sc est plus performante que la LCGA-θ_{est} probablement parce que le processus d'estimation du TL par le modèle de Rasch, génère une erreur de mesure sur θ_{est} plus importante que celle qui entache le score dans ce cas. Ceci est sans doute aussi la raison des meilleures performances de la LCGA-Sc dans la configuration

« 7 items difficiles ». En revanche, lorsque le nombre d'items disponibles pour le calcul du score diminuait encore (configuration « 4 items »), la précision du score comme mesure du phénomène subjectif n'était plus suffisante et la LCGA-θ_{est} devenait plus performante car bénéficiant de l'information des items supplémentaires présents à certains temps. En plus de la quantité d'items, les résultats de la configuration « 7 items faciles » indiquaient que la qualité des items influence aussi les performances des LCGA. Dans les conditions du scénario simulé dans cette étude, à nombre d'items constant, les performances de la LCGA-Sc sont meilleures lorsque la difficulté des items disponibles pour calculer le score est supérieure.

Une des limites régulièrement citées des études par simulation est que les scénarios simulés ne recouvrent pas l'ensemble des scénarios possibles et que, par conséquent, leurs résultats ne sont pas toujours transposables aux situations rencontrées en pratique. Cette étude est, à notre connaissance, la première étude de simulations à s'intéresser aux conséquences de la non prise en compte des items présents à certains temps seulement sur l'évaluation longitudinale d'un phénomène subjectif. Un seul scénario a été étudié afin d'offrir un premier aperçu des paramètres influents. D'autres études sont nécessaires pour étudier l'influence des autres caractéristiques de ce type d'études telles que le nombre de temps de collecte, les corrélations entre les $\theta^{(t)}$, la taille de l'échantillon, etc. L'influence de certaines hypothèses faites dans le modèle de simulations est aussi à évaluer comme celles du modèle de Rasch car, comme déjà dit ci-dessus, certain questionnaires ne sont pas en adéquation avec le modèle de Rasch. L'hypothèse d'invariance temporelle de la mesure (δ_j constantes au cours du temps) est aussi très forte, en particulier dans le champ de l'épidémiologie développementale où un phénomène de response-shift est observé au cours des stades de développement (Boylan *et al.*, 2011; Millsap, 2011). Dans le contexte de la méthode d'analyse statistique choisie, la LCGA, l'influence du nombre de groupes, de la répartition des sujets dans chacun des groupes et de la forme de la trajectoire, devrait aussi être étudiée. Enfin, les performances d'autres types d'analyses sur données longitudinales, telles que les modèles mixtes appliqués sur le score ou sur l'estimation du niveau sur le TL par exemple, pourraient être évaluées.

Cette étude a donc permis de mettre en évidence l'influence de la non prise en compte des items présents à certains temps seulement dans une étude longitudinale sur les

performances de la LCGA appliquée sur le score ou sur l'estimation du niveau sur le TL à l'aide d'un modèle de Rasch. Dans le scénario simulé, l'impact de cette non prise en compte reste faible : le gain, dans la configuration « 4 items » la plus favorable à LCGA-θ_{est}, est de 45 sujets bien classés par LCGA-θ_{est}(2227 sujets bien classés) en plus par rapport à la LCGA-Sc (2182 sujets bien classés) sur 3000 sujets. Cependant, comme indiqué dans le paragraphe ci-dessus, ce scénario ne représente pas l'ensemble des scénarios possibles en pratique et cet impact pourrait être plus important dans d'autres conditions. Ainsi, dans la pratique, il serait intéressant d'envisager une analyse de sensibilité vérifiant les résultats obtenus sur le score en appliquant l'analyse pour données longitudinale utilisées sur l'estimation du trait latent, lorsque certains items sont présents à certains temps seulement, en particulier si leur difficulté est élevée.

CHAPITRE 5 : DISCUSSION GENERALE

Le but de ce travail de thèse était d'explorer les problèmes méthodologiques que peut soulever l'utilisation d'instruments de mesure subjective en épidémiologie. Trois études empiriques ont été menées afin d'aborder trois points essentiels de l'évaluation et de l'utilisation d'une mesure en épidémiologie : l'étude de sa validité, l'évaluation de ses propriétés mathématiques et son utilisation dans un schéma longitudinal. La question sous-jacente à l'ensemble de ce travail est celle des éventuels biais d'information pouvant être introduits par l'utilisation des mesures subjectives par rapport aux mesures plus classiques telles que le poids, le nombre de cigarettes fumées par jour ou n'importe quel dosage biologique. En pratique, quelles précautions méthodologiques doivent prendre les épidémiologistes dans un tel cas ? Cette dernière partie reprendra les résultats des trois études présentées dans cette thèse afin de discuter les biais possiblement engendrés par le caractère subjectif de ces mesures et de repérer les éléments de bonnes pratiques pour les éviter. Les forces et les limites des travaux présentés dans cette thèse seront présentées avant d'évoquer certaines des perspectives de recherche sur le vaste thème abordé dans ce travail puis de conclure sur la place des mesures subjectives et l'utilisation des techniques psychométriques en épidémiologie.

1. La validité de la mesure

La première étude présentée dans ce travail a abordé la question de la taille d'échantillon nécessaire pour la validation interne d'une échelle de mesure. Lorsque le nombre d'items dans l'échelle variait entre 10 et 45, une précision de ±0,05 était obtenue avec un échantillon de 300 sujets pour le coefficient α de Cronbach lorsque sa valeur attendue était fixée à 0,7, valeur minimale acceptable pour ce coefficient. Pour une valeur attendue supérieure, un nombre moins important de sujets était nécessaire afin d'obtenir la même précision. C'est donc finalement l'étude de la structure dimensionnelle de l'instrument qui est déterminante pour la taille de l'échantillon dans ce genre d'étude. En effet, un minimum de 300 sujets était en général nécessaire pour obtenir une solution stable et fiable de l'analyse factorielle dans les conditions rencontrées en psychiatrie.

L'emploi de l'expression « structure dimensionnelle » sous-entend que l'existence de variables latentes est postulée. Autrement dit, si la validité structurelle d'une échelle n'est pas étudiée et que seul le coefficient α de Cronbach est calculé, le cadre d'utilisation de cette échelle est la CTT. Dans ce cas, un échantillon moins important est donc suffisant et c'est peut-être une des raisons pour laquelle la CTT, malgré les critiques qui lui sont faites, est encore très largement utilisée dans la littérature (Borsboom, 2005). Si la structure dimensionnelle est étudiée, ce premier travail a montré que le choix de l'ACP, plutôt que celui de l'AFE, nécessitait un échantillon moins grand et c'est, là aussi, probablement une des raisons pour laquelle l'utilisation de l'ACP est encore prédominante dans la littérature alors qu'elle n'appartient pas formellement à la famille des modèles à variables latentes (Fabrigar *et al.*, 1999; Ford *et al.*, 2006; Widaman, 1993). Les résultats obtenus par ACP et AFE sont cependant comparables dans bien des cas. Néanmoins, ce premier travail montrait que les charges factorielles étaient surestimées par ACP et ce, d'autant plus que le nombre d'items dans l'échelle était faible, sans influence de la taille de l'échantillon sur ce biais.

Lors de l'utilisation d'un instrument de mesure subjective dans une étude épidémiologique, un des premiers points à vérifier est que sa structure dimensionnelle a été étudiée lors de sa validation. Si ce n'est pas le cas, le cadre théorique choisi est la CTT. Si l'échelle présente de bonnes qualités par ailleurs (validités inter-concept, de face, de contenu, une bonne sensibilité au changement, etc.), son utilisation permettra l'obtention de résultats fiables mais les conclusions ne pourront concerner que le phénomène tel que défini par les items de l'échelle. La comparaison avec les résultats d'une autre étude sur ce même phénomène mesuré avec une autre échelle sera impossible. Par exemple, le BDI et l'échelle auto-administrée de dépression de Zung (SDS : Zung Self-Rating Depression Scale) sont deux échelles de mesure de la dépression (Shafer, 2006). A la différence du BDI, la SDS contient des items de symptomatologie positive. Dans la figure 18, un exemple hypothétique d'étude de la relation entre un facteur d'exposition et la dépression a été schématisé à l'aide d'un DAG. Dans cet exemple, si la dépression est mesurée par le BDI, aucune association significative n'est retrouvée alors que si la SDS est utilisée, une association significative est

retrouvée. Ceci est possible dans le cadre de la CTT car le construit dépression est défini par les items de l'échelle utilisée pour le mesurer.

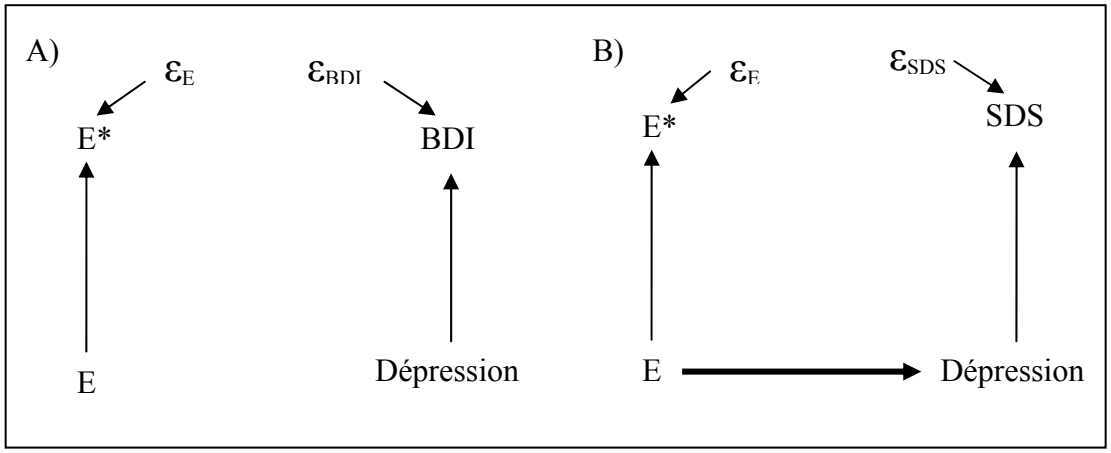


Figure 18 : Graphe orienté acyclique représentant l'étude de l'association entre un facteur d'exposition E (mesuré par E*) et la dépression mesurée, à l'aide de la théorie classique des tests, par A) l'inventaire de dépression de Beck (BDI) et par B) l'échelle auto-administrée de dépression de Zung (SDS) (ε est l'erreur de mesure, la flèche en gras représente l'association significative étudiée)

Lorsque la structure dimensionnelle de l'échelle a été étudiée et qu'il est retrouvé plusieurs dimensions, l'utilisation d'un seul score pour l'ensemble des dimensions équivaut à choisir le cadre théorique de la CTT. En revanche, si un score par dimension est calculé, comme recommandé en psychométrie, il sera possible d'étudier l'association entre le facteur d'exposition et chacune des dimensions du phénomène mesuré par l'échelle (Brown, 2006; Falissard, 2008). En reprenant l'exemple du BDI, deux dimensions sont habituellement retrouvées : les attitudes négatives envers soi et les symptômes somatiques. Pour la SDS, trois dimensions sont habituellement retrouvées : les symptômes négatifs, les symptômes somatiques et les symptômes positifs (Shafer, 2006). Une des explications possible de la situation observée dans le cadre de la CTT et représentée dans la figure 18 peut être schématisée comme dans la figure 19 où le facteur d'exposition est retrouvé indépendant du phénomène dépression mesuré par les deux sous-dimensions du BDI mais y est associé lorsque mesuré par la SDS car associé à la sous-dimension « symptômes positifs » de cette échelle.

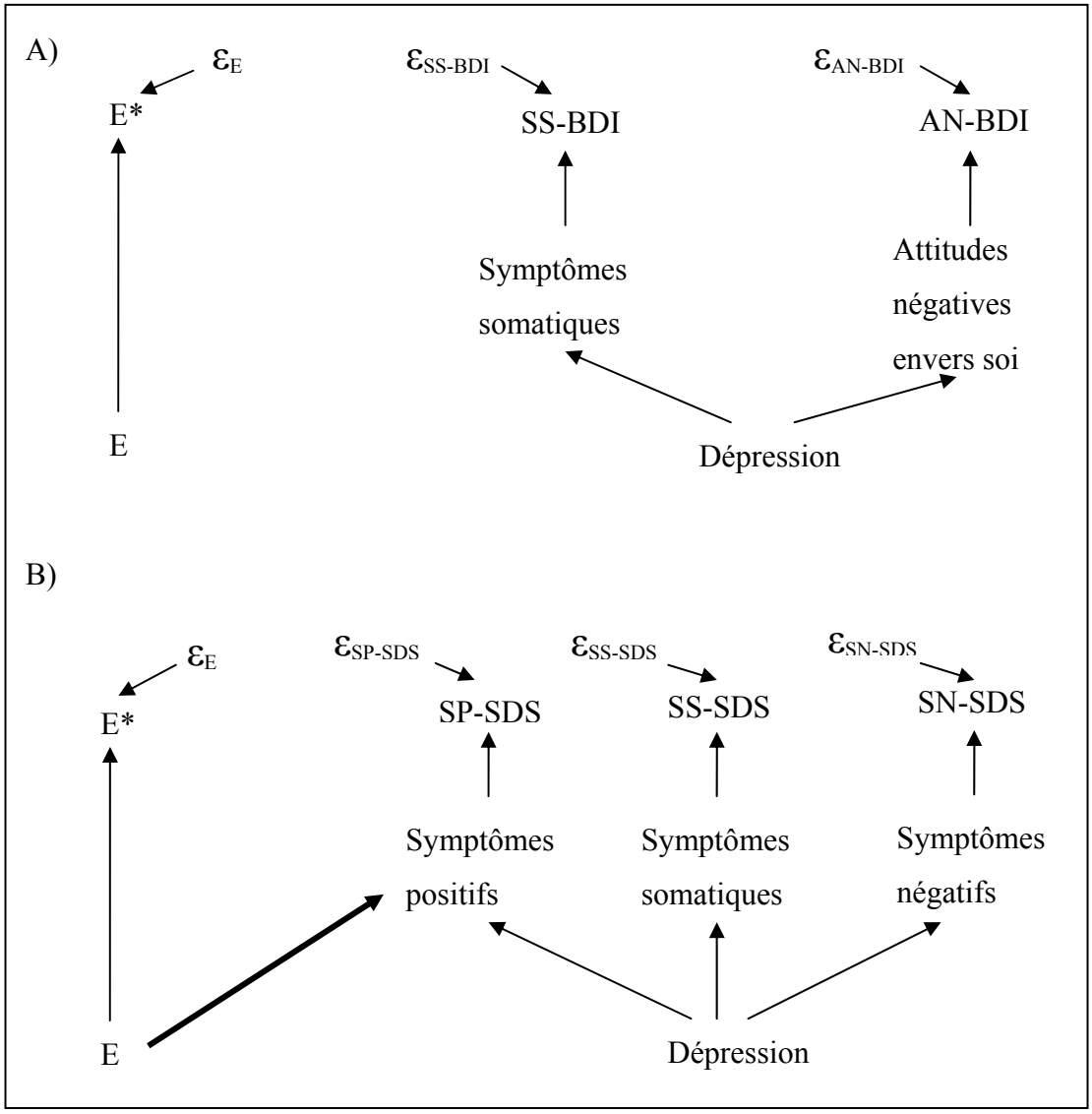


Figure 19 : Graphe orienté acyclique représentant l'étude de l'association entre un facteur d'exposition E (mesuré par E*) et les sous dimensions du phénomène « dépression » tel que mesuré par A) l'inventaire de dépression de Beck (BDI) ayant deux sous-dimensions : les symptômes somatiques (SS-BDI) et les attitudes négatives envers soi (AN-BDI), B) l'échelle auto-administrée de dépression de Zung (SDS) ayant trois sous-dimensions : les symptômes positifs (SP-SDS), les symptômes somatiques (SS-SDS) et les symptômes négatifs (SN-SDS) (ε est l'erreur de mesure, la flèche en gras représente l'association significative étudiée)

Dans cet exemple hypothétique, est illustrée la difficulté de définir la notion de biais d'information dans le cas des mesures subjectives multidimensionnelles. La distinction des trois sous-dimensions de la SDS à l'aide des modèles à variables latentes permet un degré d'analyse plus important. La présence d'une association significative avec la dépression telle

que mesurée par la SDS observée en CTT est alors comprise différemment. Cependant, comment déterminer celle des deux échelles mesurant au mieux le phénomène « dépression » ? Les symptômes positifs de la SDS font-ils partie, de manière consensuelle, de la théorie définitoire de la dépression ? Si oui, l'absence de lien entre le facteur d'exposition et la dépression telle que mesurée par le BDI est due à un biais de mesure, i.e. le BDI ne couvre pas l'ensemble des domaines définissant le construit « dépression ». Si non, c'est l'intégration de ces symptômes positifs dans la SDS qui crée le biais d'information (ils font, dans ce cas, partie de l'erreur de mesure et le biais d'information est différentiel, cf. figure 4 dans le chapitre 1). Le rationnel concernant la définition du construit à mesurer est donc d'une importance capitale lors de la validation et du choix d'un instrument de mesure subjective pour une étude épidémiologique.

L'intérêt de l'étude de la structure dimensionnelle lors de la validation d'un instrument de mesure subjective est donc, ici, mis en valeur. La taille de l'échantillon utilisé pour cette étude est aussi importante à vérifier, comme l'a montré ce premier travail. En effet, une taille insuffisante peut mener à une conclusion erronée quant à la structure dimensionnelle de l'échelle. Par exemple, pour une échelle à trois facteurs et 10 items, deux des items (14,9%) seraient en moyenne retrouvés chargés par un mauvais facteur si l'échantillon est de 50 sujets. Une telle erreur serait à l'origine d'un mauvais calcul de score pour chacune des sous-dimensions de l'échelle car calculés sur les mauvais items. Les biais d'information en découlant lors d'une utilisation dans une étude épidémiologique seraient au minimum non-différentiels.

2. La propriété d'intervalle de l'échelle de mesure

La deuxième étude de ce travail de thèse s'est intéressée à l'interprétation de la mesure obtenue lors de l'utilisation d'un instrument de mesure subjective et en particulier, à l'interprétation de son évolution au cours du temps à l'aide de la DMCP. Lorsque la CTT est utilisée, il existe un phénomène de dépendance de la DMCP du questionnaire au score observé initial des sujets. L'hypothèse de ce deuxième travail était que ce phénomène était dû au défaut de propriété d'intervalle de l'échelle du score observé en CTT. Ainsi, si l'échelle du

vrai score est une échelle d'intervalle où le vrai score du sujet est noté S, son score observé S* peut être considéré comme entaché d'une erreur de mesure dont l'importance dépend du vrai score S. Le biais d'information engendré lors de l'utilisation d'une telle échelle dans une étude épidémiologique est donc au minimum non-différentiel.

Si, dans cette étude épidémiologique, c'est l'évolution du score d'un même sujet au cours du temps qui est évaluée, la mesure est donc répétée au cours du temps. Cette situation est schématisée à l'aide d'un DAG dans la partie A de la figure 20. Les erreurs de mesure, ε_1* et ε_2*, des scores observés S_1* et S_2* aux temps 1 et 2 respectivement sont corrélées car c'est le même instrument de mesure qui est utilisé à chaque temps. Par ailleurs, leur dépendance au vrai score correspondant à chaque temps, S_1 et S_2, due au défaut de propriété d'intervalle de l'échelle de mesure est représentée ($S_1 \rightarrow \varepsilon_1$* et $S_2 \rightarrow \varepsilon_2$*). La DMCP est calculée comme la moyenne de la différence de score observée entre le temps 1 et le temps 2, DMCP*, dans la strate des sujets ayant répondu « un peu meilleur » (ou « un peu moins bon ») à la QT. A l'aide du DAG représenté dans la partie A de la figure 20, il est possible de repérer quatre chemins pouvant expliquer la dépendance de la DMCP* au vrai score initial S_1 :

- $S_1 \rightarrow S_1^* \rightarrow$DMCP*
- $S_1 \rightarrow \varepsilon_1^* \rightarrow S_1^* \rightarrow$DMCP*
- $S_1 \rightarrow S_2 \rightarrow S_2^* \rightarrow$DMCP*
- $S_1 \rightarrow S_2 \rightarrow \varepsilon_2^* \rightarrow S_2^* \rightarrow$DMCP*

Dans la partie B de cette figure, la même situation est schématisée dans le cas où l'échelle utilisée pour la mesure de S_1 et S_2 est celle du TL. Cette échelle étant supposée d'intervalle dans l'IRT, aucune dépendance des erreurs de mesure ε_1** et ε_2** aux vrais niveaux S_1 et S_2 des sujets pour le phénomène subjectif n'est représentée. Ainsi, il n'existe dans ce cas plus que deux chemins reliant S_1 à la DMCP** observée sur l'échelle du TL :

- $S_1 \rightarrow S_1^{**} \rightarrow$DMCP**
- $S_1 \rightarrow S_2 \rightarrow S_2^{**} \rightarrow$DMCP**

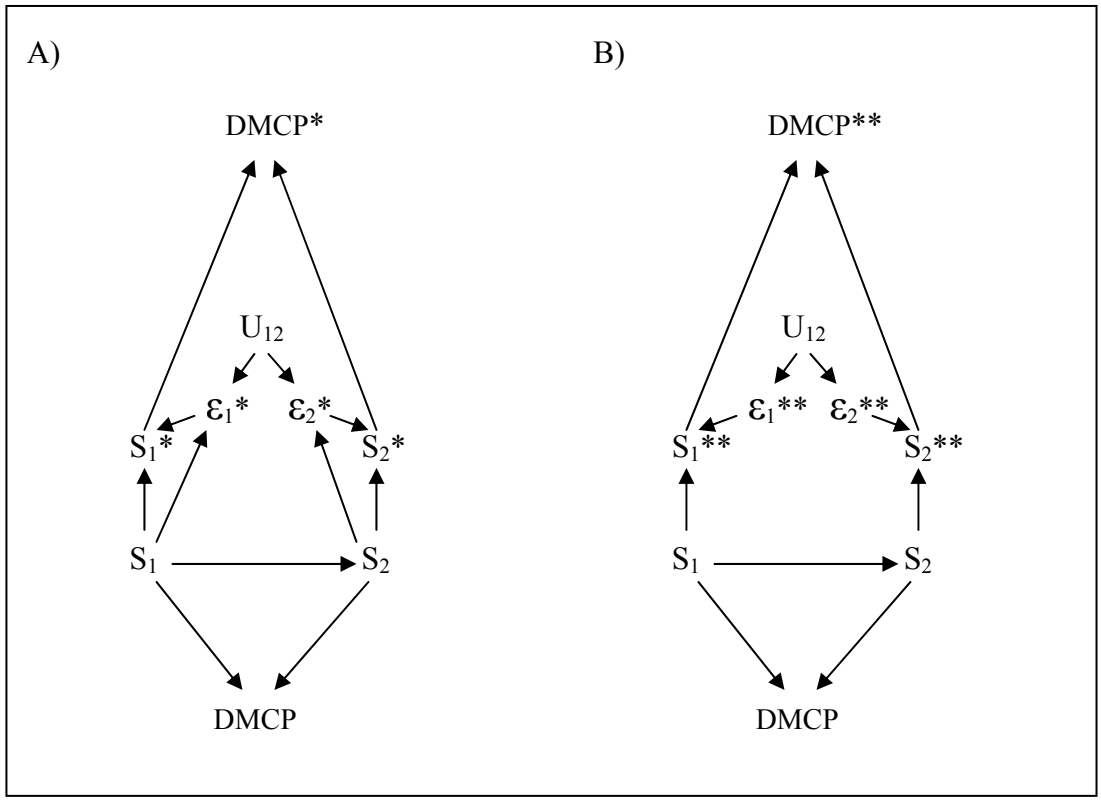


Figure 20 : Graphe orienté acyclique représentant le calcul de la Différence Minimale Cliniquement Pertinente (DMCP) à l'aide A) d'une échelle n'ayant pas les propriétés d'intervalle (*représente les estimations observées), B) d'une échelle d'intervalle ; (**représente les estimations observées) ; ε est l'erreur de mesure, U_{12} représente la cause commune – méthode d'évaluation identique - responsable de la corrélation entre $\varepsilon_1$et ε_2)

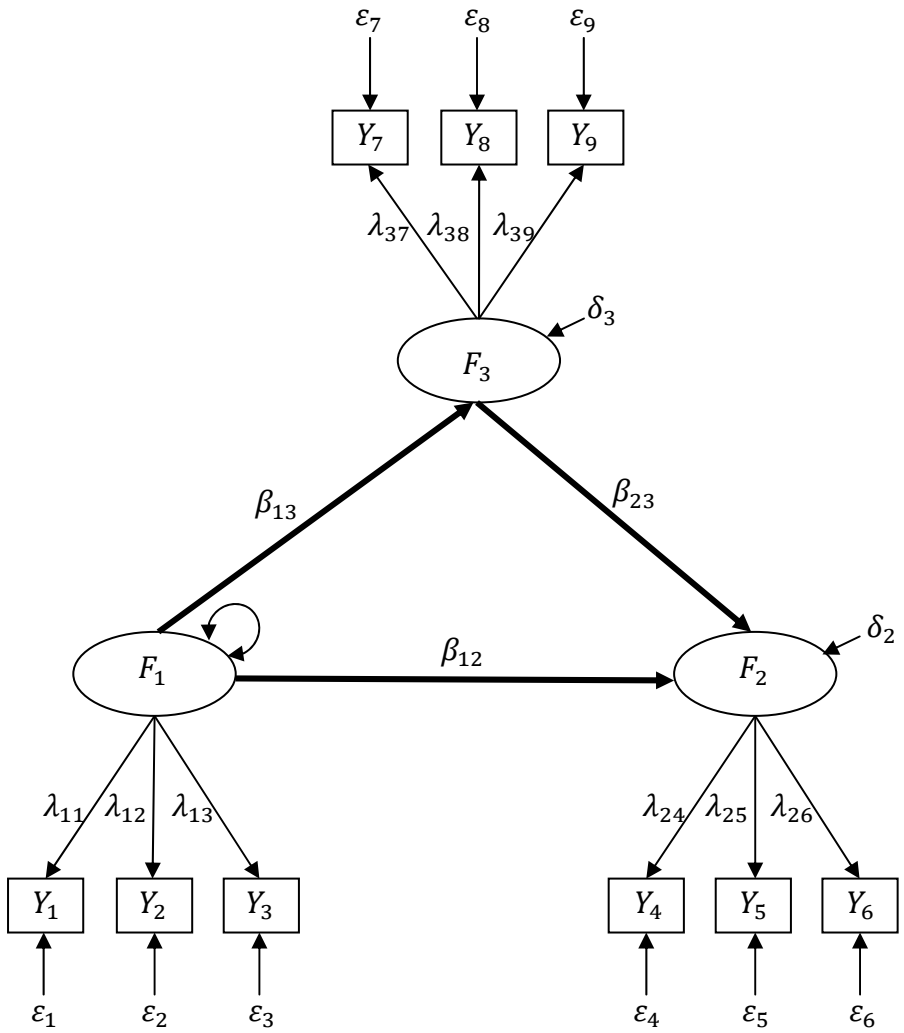
L'utilisation de DAG dans cette situation permet d'apporter des éléments de discussion supplémentaires concernant les résultats obtenus dans la deuxième étude de ce travail de thèse. La dépendance de la DMCP au score initial engendrée par le défaut de propriété d'intervalle de l'échelle du score est visible dans la partie A de la figure 20, cependant, il faut noter que la situation schématisée dans la partie B est idéalisée. En effet, aucune flèche n'a été représentée entre S_1 et ε_1** ou S_2 et ε_2** pour souligner la propriété d'intervalle de l'échelle du TL mais la précision des mesures sur cette échelle par un modèle IRT est plus faible pour des niveaux extrêmes (Embretson, 1991). Formellement, ces flèches représentant la dépendance des erreurs de mesure au vrai niveau sur le phénomène mesuré devraient donc aussi figurer dans la partie B mais pour une raison différente que dans la partie A, l'imprécision existant sur les valeurs extrêmes. Les faibles différences de résultats observées dans cette deuxième étude sont

probablement en partie dues à l'intensité des liens représentés par ces différentes flèches et dont les forces peuvent se compenser d'un cas à l'autre.

Enfin, un des points les plus intéressants mis en évidence par ces DAG nécessite de ne raisonner qu'à partir des vraies valeurs des niveaux sur le phénomène mesuré S_1, S_2 et DMCP. En examinant le triangle situé dans la partie basse de chacun de ces deux DAG concernant ces vraies valeurs, il est possible de montrer que la dépendance de la différence de score au score initial observée empiriquement ne découle pas seulement de la méthode de mesure. En effet, lorsque qu'une mesure est répétée chez un même individu, il est très souvent (quasi systématiquement dans le domaine de la psychométrie) possible de faire l'hypothèse d'un lien causal entre la valeur initiale et les valeurs ultérieures. Ce lien est responsable d'une dépendance de la vraie DMCP à la vraie valeur initiale S_1. Dans ce triangle, la DMCP est définie à l'aide de S_1 et S_2 or S_2 est défini à l'aide S_1 grâce au lien causal. Il est donc possible en connaissant la valeur de S_1 de connaître la valeur de la DMCP. Ainsi, ce raisonnement causal met en évidence, en accord avec les résultats de cette deuxième étude, le fait qu'il n'existe pas d'unique valeur de DMCP par questionnaire. Comme certains auteurs l'ont recommandé dernièrement, la DMCP d'un questionnaire devrait être définie comme un ensemble de valeurs en fonction du score initial du sujet (Copay *et al.*, 2007; Crosby *et al.*, 2004, 2003; Revicki *et al.*, 2008; Tubach *et al.*, 2005).

Dans une étude épidémiologique, le défaut de propriété d'intervalle d'une échelle de mesure entrainera systématiquement au moins un biais d'information non-différentiel (cf. figure 4, partie A dans le chapitre 1). Comme discuté ci-dessus, en cas de schéma longitudinal, un biais sur l'estimation de la relation entre deux mesures répétées peut en résulter et ainsi conduire à des conclusions erronées concernant l'évolution au cours du temps du phénomène subjectif. Cependant, la vérification de la propriété d'intervalle d'une échelle ayant été validée à l'aide de la CTT est peu envisageable en pratique. En revanche, si un modèle à variables latentes est utilisé pour étudier la structure de l'échelle, il sera possible de pondérer chaque item afin de permettre le calcul d'un score respectant la propriété d'intervalle de l'échelle de la variable latente. Ces poids seront fonction des charges factorielles dans le modèle en facteurs communs et spécifique ou des paramètres d'items des modèles IRT.

Une autre solution s'offre aux épidémiologistes et est encore peu utilisée en pratique dans ce domaine : les modèles d'équations structurelles (SEM : Structural Equation Modeling) (Arlinghaus *et al.*, 2012; Beran et Violato, 2010; Tu, 2009). Ces modèles font partie des modèles d'analyse causale mais, à la différence des DAG, ce sont des modèles d'analyse statistique où des hypothèses sont faites sur la distribution des variables et sur la forme des liens entre variables (Dumas *et al.*, 2014). Dans un SEM, il existe deux parties : 1/ le modèle de mesure définissant les liens entre les variables observées et les variables latentes (selon le modèle en facteurs communs ou spécifique ou selon un modèle IRT) et 2/ le modèle structurel représentant les hypothèses sur les liens causaux directs ou indirects entre les variables latentes (ou entre variables latentes et certaines variables observées) (R. B. Kline, 2005). La figure 21 représente le diagramme de chemin et la transcription mathématique d'un SEM avec trois variables latentes. Dans le modèle structurel, représenté par les flèches en gras, les hypothèses sont qu'un lien causal direct de la variable latente F_1 vers la variable latente F_2 existe mais que l'effet de F_1sur F_2est aussi issu d'un lien causal indirect passant par la variable latente intermédiaire F_3. A l'aide des SEM, il est donc possible de travailler directement sur la variable latente à partir des variables observées sans passer par le calcul d'un score risquant de compromettre la propriété d'intervalle de l'échelle de mesure utilisée. Le défaut de cette propriété ne peut donc pas être une source de biais lors de l'estimation des coefficients structurels dans ce type de modèle.



Modèle de mesure (issu du modèle en facteurs communs et spécifique)

$$Y_j = \alpha_j + \lambda_{mj}F_m + \varepsilon_j \quad \text{avec } \varepsilon_j \sim \mathcal{N}(0, \sigma_j),\ j = 1 \text{ à } 9,\ m = 1 \text{ à } 3$$

$$\lambda_{11} = \lambda_{24} = \lambda_{37} = 1 \text{ et } F_1 \sim \mathcal{N}(0, \psi_1)$$

Modèle structurel

$$\begin{cases} F_2 = \alpha_{12} + \beta_{12}F_1 + \beta_{23}F_3 + \delta_2 \\ F_3 = \alpha_{13} + \beta_{13}F_1 + \delta_3 \end{cases}$$

avec $\delta_m \sim \mathcal{N}(0, \xi_m)$ pour $m \neq 1$

Figure 21 : Exemple de diagramme de chemin et de sa transcription mathématique pour un modèle d'équations structurelles à trois variables latentes (F) et neuf items (Y). Les résidus sont notés ε, les charges factorielles λ, les constantes α, les coefficients structurels β et les erreurs structurelles δ. Les flèches en gras représentent le modèle structurel, i.e. les hypothèses sur les liens causaux existant entre les variables latentes

3. L'utilisation dans un schéma longitudinal

Après la validation et l'interprétation de la mesure, c'est son utilisation dans une étude de cohorte par une méthode d'analyse statistique pour données longitudinales, la LCGA, qui a fait l'objet de la dernière étude de ce travail de thèse. Dans cette étude, c'est finalement l'influence du niveau de précision de l'estimation du phénomène subjectif (le score ou l'estimation du niveau sur le TL par un modèle de Rasch) sur les performances de la LCGA qui a été étudiée. La conclusion est que si le score ne permet pas d'estimer le phénomène avec une précision suffisante par manque d'items présents à tous les temps, l'utilisation d'un modèle de Rasch pour estimer le TL peut permettre d'améliorer cette précision et donc d'améliorer les performances de la LCGA, en prenant en compte certains items non disponibles pour le calcul du score.

La figure 22 résume la situation rencontrée dans cette étude à l'aide d'un DAG (seuls trois temps sont représentés) où les estimations TL_t* des niveaux sur le TL à chaque temps t sont associées à une erreur de mesure ε_t. Cette erreur est habituellement liée à la vraie valeur du TL au même temps et, dans le cadre de cette étude, corrélée aux erreurs des mesures faites aux autres temps car utilisant les mêmes items. Le défaut de précision ainsi représenté n'entraine que des biais non-différentiels sur l'estimation des liens entre, par exemple, l'ordonnée à l'origine de la trajectoire latente O et les TL_t, ou entre la pente P et les TL_t, etc (cf. figure 4, partie B dans le chapitre 1). Dans la dernière étude de ce travail, le principal facteur qui varie entre les différentes configurations et entre les deux types d'analyses LCGA-Sc et LCGA-θ_{est}est le nombre d'items utilisés pour estimer le phénomène subjectif. Moins il y a d'items pris en compte, plus l'effet des ε_t est important et plus les performances de la LCGA sont altérées.

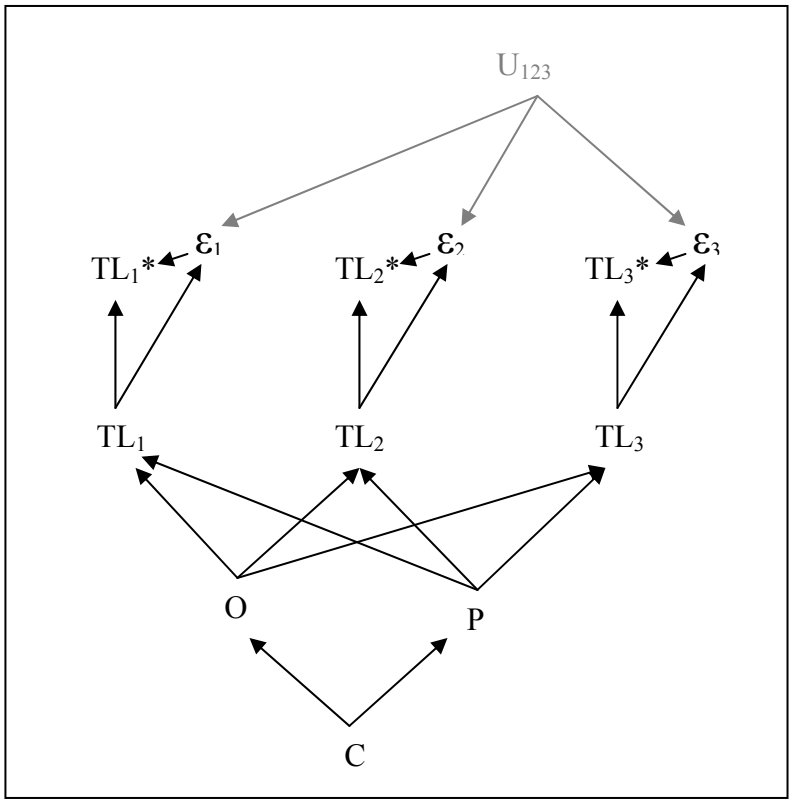


Figure 22 : Graphe orienté acyclique représentant une analyse en classe latente (C) de trajectoires (de pente P et d'ordonnées à l'origine O) (LCGA). Les TL_t sont les vrais niveaux sur le trait latent à chaque temps t, les TL_t* représentent les estimations des TL_t ; les ε_t sont les erreurs de mesure à chaque temps t et U_{123} représente la cause commune (méthode d'évaluation) responsable de la corrélation entre les ε_t

Dans cette étude, une distinction a été faite entre les deux configurations comportant sept items concernant la difficulté des items disponibles pour le calcul du score. Dans un questionnaire, les difficultés des items doivent être réparties sur l'ensemble de l'étendue des valeurs du niveau sur le TL représentées dans l'échantillon auquel il va être appliqué. Ceci permet d'assurer une précision suffisante aux estimations du phénomène subjectif mesuré pour l'ensemble des sujets de l'échantillon, qu'ils aient un niveau haut, bas ou moyen. Le scénario simulé dans cette dernière étude comporte trois groupes : un premier ayant une trajectoire basse et stable au cours du temps, un autre une trajectoire stable et haute et le troisième une trajectoire croissante au cours du temps. Dans la configuration « 7 item faciles » où certains items difficiles manquent pour le calcul du score, les estimations de la mesure du phénomène subjectif des sujets ayant un haut niveau seront donc moins précises que celles des sujets ayant

un niveau bas. Ainsi, dans la figure 22, une flèche pourrait être ajoutée entre l'ordonnée à l'origine O, représentant le niveau de base pour le phénomène subjectif dans les trois groupes, et chaque ε_t. De même, une flèche pourrait être ajoutée entre la pente P, représentant le temps, et les ε_t étant donnée la présence d'un groupe de trajectoire croissante au cours du temps. Cette situation est donc celle entrainant un biais différentiel pour l'estimation des différentes associations entre O, P, C et les TL_t (c.f. figure 4, partie D dans le chapitre 1) et il est probable que les erreurs de classement fourni par la LCGA dans ce type de configuration soient donc des erreurs différentielles.

Ce type de biais est aussi celui qui pourrait théoriquement survenir lorsque l'hypothèse d'invariance temporelle de la mesure est faite à tort. Dans les modèles IRT, l'invariance de la mesure peut être supposée lorsque les paramètres d'items, difficultés et discriminations, ont une valeur stable au cours du temps. Dans le modèle en facteurs communs et spécifique, ceci correspond à la stabilité temporelle des charges factorielles et des constantes. Si cette hypothèse est faite à tort lors de l'estimation de la mesure du phénomène subjectif au cours du temps, l'erreur de mesure associée à cette estimation sera dépendante du temps et, là aussi, une flèche pourrait être ajoutée entre la pente P, représentant le temps, et les ε_t dans la figure 22. L'influence du non respect de cette hypothèse sur les performances de la LCGA n'a pas été étudiée dans cette dernière étude car le modèle de simulation supposait l'invariance temporelle de la mesure.

L'impact du biais évalué dans la dernière étude restait faible dans les conditions du scénario simulé. Il est difficile de l'évaluer dans des scénarios différents sans la mise en place d'autres études par simulation. Les recommandations qui peuvent donc être faites à l'heure actuelle lors de l'utilisation d'instruments de mesure subjective dans une étude épidémiologique ayant pour but d'étudier l'évolution du phénomène subjectif au cours du temps sont donc de deux types. Le premier concerne les caractéristiques de l'instrument utilisé avec la vérification de l'adéquation entre les difficultés des items le composant et les niveaux du phénomène subjectif dans l'échantillon lors de la période étudiée. L'invariance de la mesure au cours de la période étudiée doit aussi être évaluée et des méthodes statistiques permettant la prise en compte d'un phénomène de « response-shift » doivent être utilisées le

cas échéant. Le deuxième type de recommandations concerne la méthode statistique à utiliser pour l'estimation de la mesure du phénomène. Si la méthode de calcul du score est basée sur un modèle de mesure issu d'un modèle à variable latente et que l'ensemble des items disponibles à chaque temps sont identiques, les résultats de cette troisième étude indiquent qu'il n'y a aucune raison de ne pas s'en servir. Voire même, il est préférable de s'en servir car les estimations du niveau sur le TL obtenues à l'aide des modèles IRT pourraient être moins précises. Si ces deux conditions ne sont pas réunies, l'utilisation de techniques plus sophistiquées pour l'estimation de la mesure du phénomène subjectif peut être avantageuse dans certains cas et est donc à discuter. Enfin, des méthodes statistiques non paramétriques sont aussi disponibles lorsque les propriétés mathématiques du score (hypothèse d'invariance temporelle non respectée, difficultés des items non-régulièrement espacées sur l'étendue des niveaux sur le TL, etc.) ne répondent pas aux exigences de l'application des méthodes paramétriques. Par exemple, dans le contexte des groupes de trajectoires homogènes, le package KML du logiciel R permet d'appliquer un algorithme non-paramétrique ayant le même but que la LGCA : la clusterisation de données longitudinales (Genolini et Falissard, 2010).

4. Forces et limites

La principale préoccupation de l'épidémiologiste depuis la planification de son étude jusqu'à l'interprétation des résultats est d'éviter ou, au minimum, d'identifier les biais qui pourraient fausser les résultats de son étude. Une des grandes forces de ce travail de thèse est qu'en s'intéressant à différentes phases de l'évaluation et de l'utilisation des instruments de mesure subjective, les trois études qui y sont présentées ont permis de mettre en évidence la multiplicité des sources potentielles de biais d'information pouvant résulter de leur utilisation. Par ailleurs, ces trois études ont permis d'apporter des éléments de correction à certaines croyances ancrées dans la littérature : la règle du ratio nombre de sujets sur nombre d'items contenus dans l'échelle, par exemple, pour le calcul du nombre de sujets à inclure dans les études de validation d'échelle ; la responsabilité du défaut de propriété d'intervalle de l'échelle du score dans la dépendance de la DMCP au score initial ; l'existence d'une valeur

unique de DMCP par questionnaire ; etc. Enfin, l'utilisation d'études par simulation a eu deux avantages dans ce travail de thèse. Dans la première étude, la simulation de différents scénarios a permis d'établir un outil de détermination du nombre de sujets nécessaire dans différentes conditions pouvant être rencontrées en pratique lors d'une étude de validation de questionnaire en psychiatrie. Dans la deuxième étude, la comparaison des résultats obtenus lors de l'analyse des données simulées aux valeurs des paramètres entrés dans le modèle de simulation a permis de quantifier l'impact du biais engendré par la non prise en compte de certains items dans le calcul du score.

Ce travail de thèse a cependant certaines limites qui sont finalement le pendant de ses forces. Par exemple, les études par simulation utilisées, si elles permettent la génération de données correspondant à différents scénarios, ne peuvent pas simuler l'ensemble des scénarios possibles en pratique. Il est donc bien souvent nécessaire d'extrapoler les résultats des études de simulation lorsque le temps est venu de les appliquer en pratique. Par exemple, pour une étude de validation d'une échelle à cinq facteurs et 63 items, la taille de l'échantillon nécessaire ne pourra être qu'extrapolée à partir de résultats obtenus dans la première étude de ce travail. De même, l'impact de l'impossible prise en compte de certains items lors du calcul du score sur les performances de la LCGA-Sc dans la troisième étude est faible dans le scénario simulé, mais que serait-il avec une répartition différente des sujets dans les groupes de trajectoire ou avec un nombre supérieur de groupes de trajectoire, par exemple ? S'il est difficile d'obtenir une généralisation universelle lors d'études par simulation, leur flexibilité dans la simulation de différents scénarios est toutefois un atout par rapport aux études sur données réelles où les résultats ne sont applicables qu'à la situation rencontrée dans ces données en particulier. Dans la deuxième étude par exemple, la différence attendue entre les performances de la DMCP-Sc et de la DMCP-TL n'a pas été observée dans cet échantillon mais sur un échantillon différent (ayant un score moyen de santé perçue bien plus haut par exemple) peut-être qu'une différence apparaitrait. D'autres projets de recherche sont nécessaires pour tenter de répertorier l'ensemble des différents facteurs influents sur la présence et sur l'impact des biais d'information liés à l'utilisation des instruments de mesure subjective en épidémiologie.

5. Perspectives de recherche

Dans cette optique plusieurs projets de recherche sont en cours ou en préparation en collaboration avec les différentes unités de recherche déjà impliquées dans les études présentées dans ce travail ou de nouvelles collaborations. Une première thématique concerne l'invariance temporelle de la mesure et la présence de « response-shift » (Brown, 2006; Millsap, 2011; Sprangers et Schwartz, 1999). Dans la deuxième étude présentée dans ce manuscrit, il existait chez les patients considérés comme ayant un état de santé stable entre les deux temps de collecte (ayant répondu « à peu près pareil » à la QT) une diminution moyenne de leur santé perçue qu'elle soit mesurée sur l'échelle du score ou sur celle du TL à l'aide de la sous-échelle SP du questionnaire MOS-SF36. Comme discuté dans cette étude, ce phénomène posait plusieurs questions dont celle de la présence de response-shift dans un ou plusieurs items de la sous-échelle SP entre les deux temps de collecte. Le phénomène de response-shift apparaît lorsque des changements de référence vis-à-vis du phénomène subjectif mesuré surviennent au cours du temps. Par exemple, dans le cadre de la douleur chronique, une adaptation des sujets à leur douleur peut avoir lieu au cours du temps et entrainer une modification de leur conception du phénomène « douleur ». Dans un tel cas, à douleur identique aux deux temps de collecte, les sujets pourront répondre différemment aux items d'un questionnaire sur la douleur chronique. Une demande de financement par l'EA 4275 à Nantes a été déposée à l'Agence Nationale pour la Recherche en vue d'étudier la problématique de la définition et de la détermination de la DMCP d'un questionnaire en présence de response-shift.

Une autre thématique est la présence d'un phénomène appelé fonctionnement différentiel d'item (DIF : Differential Item Functioning) lié à une caractéristique des sujets (le sexe, l'âge, la région géographique d'habitation, etc.). A l'image du response-shift (qui n'est autre qu'un DIF lié non pas à une caractéristique du sujet mais au temps), le DIF associé à une caractéristique du sujet survient lorsque, dans les modèles IRT, les paramètres d'items, difficultés et/ou discriminations, ont une valeur différente en fonction d'une caractéristique. Dans le modèle en facteurs communs et spécifique, le DIF correspond à des charges factorielles et/ou des constantes différentes en fonction d'une caractéristique des sujets

(Millsap, 2011). Par exemple, un résultat fréquemment retrouvé dans la littérature est le niveau de la qualité de vie perçue inférieur chez les femmes. Il est possible que si le questionnaire de qualité de vie utilisé contient des items affectés de DIF sur le sexe, cette différence observée soit la résultante d'un biais de mesure lié à la présence de ce DIF-sexe. Cette question a été étudiée sur un échantillon représentatif de la population française et il semble que même en prenant en compte dans les analyses le DIF-sexe existant dans le questionnaire MOS-SF36, cette différence entre hommes et femmes persiste (Hardouin *et al.*, 2012). Ce résultat a entrainé une réflexion sur l'impact possible de la présence de DIF en termes de biais de mesure. Une étude par simulation a donc été programmée et est en cours d'analyse dans le but d'étudier cette problématique. Le critère de jugement principal en est le biais sur l'estimation d'une différence concernant le phénomène mesuré par un questionnaire entre deux groupes concernés par du DIF. Les différents paramètres contrôlés dans le modèle de simulation sont : l'amplitude de la vraie différence entre les deux groupes, le pourcentage d'items du questionnaire touchés par le DIF, l'ampleur du DIF sur chaque item, etc.

Une autre thématique, ne portant pas sur le biais de mesure, est l'application en pratique et l'apport des techniques psychométriques, en particulier des SEM, en épidémiologie. Comme décrit précédemment, un des intérêts de ces modèles est qu'ils permettent de travailler directement sur les variables latentes et donc d'éviter les différentes sources de biais potentiellement engendrées par le calcul du score. Un autre avantage de ces modèles est la possibilité qu'ils offrent d'étudier différentes hypothèses concernant les liens de causalité existant entre les phénomènes subjectifs étudiés. Par exemple, un nouveau modèle théorique du fonctionnement, du handicap et de la santé (CIF : Classification Internationale du Fonctionnement) a été publié en 2001. Aucune étude empirique n'a, à ce jour, encore testé ce modèle dans sa globalité (World Health Organization, 2001). Ce modèle a plusieurs composantes distinguant les conséquences d'une condition de santé (déficience des fonctions organiques et des structures anatomiques, limitations des activités, restriction de la participation sociale) et les facteurs contextuels (personnels et environnementaux). Aucune hypothèse n'est faite dans ce modèle théorique sur les liens de causalité existant entre ces composantes. Les liens représentés dans le document officiel de l'OMS le sont par des flèches

bidirectionnelles reliant toutes les composantes entre elles. L'application des SEM sur les données d'une étude de cohorte française multicentrique de sujets souffrant de gonarthrose et suivis pendant trois ans a permis d'étudier en longitudinal les relations causales et les interactions existant entre les différentes composantes de la CIF mesurées à l'aide de différents questionnaires de santé perçue (Guillemin *et al.*, 2012). Cette étude est en cours de finalisation.

6. Conclusion

La place de l'étude des phénomènes subjectifs en épidémiologie n'est plus discutée à l'heure actuelle où les modèles complexes et le rôle intriqué des facteurs biologiques, psychologiques, environnementaux et sociaux sont de plus en plus envisagés dans le chemin causal vers la santé. En s'appropriant les techniques psychométriques développées depuis plus d'un siècle par les chercheurs en psychologie, en éducation et en sociologie, les épidémiologistes se doivent d'en connaitre les principes fondateurs afin d'être capables de repérer les sources potentielles de biais pouvant être introduites dans leur étude lors de l'utilisation des instruments de mesure subjective. Des modèles spécifiques (IRT, SEM, etc.) à ce type de mesure sont parfois nécessaires pour les analyser et leur diffusion devrait être plus large vue la fréquence actuelle de l'intégration de phénomènes subjectifs dans les études épidémiologiques. Ces modèles ont par ailleurs certains avantages, telle que l'analyse causale permise par les SEM, qui pourraient apporter de précieuses informations lors de l'analyse des déterminants et des conséquences d'un problème de santé.

L'insuffisance de diffusion des méthodes psychométriques dans les formations et livres de référence proposés en épidémiologie est effectivement un des freins à leur utilisation dans ce domaine mais il ne serait pas juste de ne citer que celui-ci. L'importance de l'aspect calculatoire de ces techniques et la nécessité de l'utilisation de logiciels spécialisés et payants (tels Mplus©, Lisrel©, Amos©, etc.) ont longtemps été des barrières à l'application de ces modèles et ce, quel que soit le domaine d'application. De nombreux développements de programmes sur des logiciels plus couramment utilisés en épidémiologie ont eu lieu ces dernières années avec, par exemple, l'introduction d'un module de SEM depuis la version 12

de Stata© sortie en 2011 ou encore la multiplication des packages dédiés à l'IRT, aux SEM, aux analyses en classe latentes, etc. dans le logiciel gratuit R ces dernières années.

Finalement, le plus grand frein à l'application de ces techniques est probablement l'hypothèse qu'elles nécessitent : l'existence de variables latentes. Un débat, dépassant les limites du travail exposé ici, existe effectivement sur la plausibilité de l'existence de telles quantités non-observables remettant ainsi en question l'ensemble de la théorie sous-tendant les modèles à variables latentes. Toute personne découvrant pour la première fois ces modèles se retrouve face à cette interrogation de la plausibilité d'une telle hypothèse et de la signification de telles variables. Si un niveau de dépression égal à zéro représente la moyenne dans la population, que représente-t-il lorsqu'il se dirige vers moins l'infini ? L'absence de borne et donc la propriété d'intervalle de l'échelle des variables latentes est elle concevable pour des quantités telles que la douleur, la qualité de vie, la satisfaction ? La formation sur le cadre théorique de la construction des questionnaires et échelles de mesure est nécessaire pour aider les épidémiologistes à se poser ce type de questions. Ils pourront ensuite décider d'adhérer ou non à ce cadre théorique, ce qui déterminera l'ensemble des opérations et modèles qu'ils pourront appliquer sur les données issues de ce type d'instrument.

BIBLIOGRAPHIE

Adachi N, Onuma T, Nishiwaki S, Murauchi S, Akanuma N, Ishida S, Takei N. (2000) Inter-ictal and post-ictal psychoses in frontal lobe epilepsy: a retrospective comparison with psychoses in temporal lobe epilepsy. *Seizure*, 9, 328-35.

Aleamoni LM. (1973) Effects of size of sample on eigenvalues, observed communalities, and factor loadings. *J Appl Psychol*, 58, 266-9.

Andrich D, Sheridan BS, Luo G. (2010) Rumm2030: Rasch Unidimensional Measurement Models. RUMM Laboratory, Perth, Western Australia.

Anthoine E, Moret L, Regnault A, Sébille V, Hardouin JB. (2012) How PRO measures are psychometrically validated? A review of publications on primary validation. Communication personnelle.

Arlinghaus A, Lombardi DA, Willetts JL, Folkard S, Christiani DC. (2012) A Structural Equation Modeling Approach to Fatigue-related Risk Factors for Occupational Injury. *Am. J. Epidemiol*, kws219.

Armitage P, Colton T. (1998) Encyclopedia of biostatistics: PRI-SPH. J. Wiley.

Baker DW, Hays RD, Brook RH. (1997) Understanding changes in health status. Is the floor phenomenon merely the last step of the staircase? *Med Care,* 35, 1-15.

Bartholomew, DJ, Knott M, Moustaki I. (2011) Latent Variable Models and Factor Analysis: A Unified Approach. Wiley Series in Probability and Statistics. Wiley.

Basker M, Moses PD, Russell S, Russell PS. (2007) The psychometric properties of Beck Depression Inventory for adolescent depression in a primary-care paediatric setting in India. *Child Adolesc Psychiatry Ment Health*, 1, 8-14.

Beaton DE. (2003) Simple as possible? Or too simple? Possible limits to the universality of the one half standard deviation. *Med Care*, 41, 593-6.

Beaton DE, Boers M, Wells GA. (2002) Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. *Curr Opin Rheumatol,* 14, 109-14.

Beaton DE, Bombardier C, Katz JN, Wright JG, Wells G, Boers M, Strand V, Shea B. (2001) Looking for important change/differences in studies of responsiveness. OMERACT MCID Working Group. Outcome Measures in Rheumatology. Minimal Clinically Important Difference. *J Rheumatol,* 28, 400-5.

Beaton DE, van Eerd D, Smith P, van der Velde G, Cullen K, Kennedy CA, Hogg-Johnson S. (2011) Minimal change is sensitive, less specific to recovery: a diagnostic testing approach to interpretability. *J Clin Epidemiol,* 64, 487-96.

Beck AT. (1991) Relationship between the Beck Anxiety Inventory and the Hamilton Anxiety Rating Scale with anxious outpatients. *J Anxiety Disord*, 5, 213-23.

Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. (1961) An inventory for measuring depression. *Arch Gen Psychiatry*, 4, 561-71.

Bell MD, Lysaker PH, Beam-Goulet JL, Milstein RM, Lindenmayer JP. (1994) Five-component model of schizophrenia: assessing the factorial invariance of the positive and negative syndrome scale. *Psychiatry Res,* 52, 295-303.

Beran TN, Violato C. (2010) Structural equation modeling in medical research: a primer. *BMC Research Notes*, 3, 267-87.

Berlin KS, Parra GR, Williams NA. (2014) An Introduction to Latent Variable Mixture Modeling (Part 2): Longitudinal Latent Class Growth Analysis and Growth Mixture Models. *J Pediatr Psychol,* 39, 188-203.

Bird SB, Dickson EW. (2001) Clinically significant changes in pain along the visual analog scale. *Ann Emerg Med,* 38, 639-43.

Blanchin M, Hardouin JB, Le Neel T, Kubis G, Blanchard C, Mirallie E, Sebille V, (2010) Comparison of CTT and Rasch-based approaches for the analysis of longitudinal Patient Reported Outcomes. *Stat Med*, 30, 825-38.

Boini S, Erpelding ML, Fagot-Campagna A, Mesbah M, Chwalow J, Penfornis A, Coliche V, Mollet E, Meadows K, Briançon S. (2010) Factors associated with psychological and behavioral functioning in people with type 2 diabetes living in France. *Health Qual Life Out*, 8, 124-32.

Bollen KA. (2002) Latent variables in psychology and the social sciences. *Annu rev psycho*, 53, 605-34.

Bonicatto S, Dew AM, Soria JJ. (1998) Analysis of the psychometric properties of the Spanish version of the Beck Depression Inventory in Argentina. *Psychiatry Res,* 79, 277-85.

Bonilla J, Bernal G, Santos A, Santos D. (2004) A revised Spanish version of the Beck Depression Inventory: psychometric properties with a Puerto Rican sample of college students. *J Clin Psychol,* 60, 119-30.

Borsboom D. (2005) Measuring the Mind: Conceptual Issues in Contemporary Psychometrics. Cambridge University Press.

Borsboom D. (2008) Latent Variable Theory. *Measurement: Interdisciplinary Research and Perspectives,* 6, 25-53.

Bouyer J. (2009) Épidémiologie: principes et méthodes quantitatives. Éd. Tec & doc.

Boylan KR, Miller JL, Vaillancourt T, Szatmari P. (2011) Confirmatory factor structure of anxiety and depression: evidence of item variance across childhood. *Int J Methods Psychiatr Res,* 20, 194-202.

Briançon S, Boini S, Bertrais S, Guillemin F, Galan P, Hercberg S. (2011) Long-term antioxidant supplementation has no effect on health-related quality of life: The randomized, double-blind, placebo-controlled, primary prevention SU.VI.MAX trial. *Int J Epidemiol*, 40, 1605-16.

Brown TA. (2006) Confirmatory factor analysis for applied research. The Guilford Press, New York.

Burton A, Altman DG, Royston P, Holder RL. (2006) The design of simulation studies in medical statistics. *Stat Med*, 25, 4279-92.

Castro-Costa E, Uchoa E, Firmo JO, Lima-Costa MF, Prince M. (2008) Association of cognitive impairment, activity limitation with latent traits in the GHQ-12 in the older elderly. The Bambui Health and Aging Study (BHAS). *Aging Clin Exp Res,* 20, 562-8.

Cattell RB. 1978. The scientific use of factor analysis in behavioral and life sciences. Plenum press, New York.

Celeux G, Soromenho G. (1996) An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13, 195-212.

Chapman LK, Williams SR, Mast BT, Woodruff-Borden J. (2009) A confirmatory factor analysis of the Beck Anxiety Inventory in African American and European American young adults. *J Anxiety Disord*, 23, 387-92.

Clancy CM, Eisenberg JM. (1998) Outcomes research: measuring the end results of health care. *Science* 282, 245-6.

Cohen J. (1988) Statistical Power Analysis for the Behavioral Sciences. L. Erlbaum Associates.

Comrey AL. (1978) Common Methodological Problems in Factor Analytic Studies. *J Consult Clin Psych*, 46, 648-59.

Comrey AL, Lee HB. (1992) A first course in factor analysis. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Cook CE. (2008) Clinimetrics Corner: The Minimal Clinically Important Change Score (MCID): A Necessary Pretense. *J Man Manip Ther* 16, E82-3.

Copay AG, Subach BR, Glassman SD, Polly DW Jr, Schuler TC. (2007) Understanding the minimum clinically important difference: a review of concepts and methods. *Spine J*, 7, 541-6.

Costello AB, Osborn JW. (2005) Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis. *Practical Assessment, Research & Evaluation* 10, 1-9.

Costello EJ, Foley DL, Angold A. (2006) 10-year research update review: the epidemiology of child and adolescent psychiatric disorders: II. Developmental epidemiology. *J Am Acad Child Adolesc Psychiatry*, 45, 8-25.

Cronbach LJ. (1951) Coefficient alpha and the internal structure of a test. *Psychometrika,* 16, 297-334.

Crosby RD, Kolotkin RL, Williams GR. (2003) Defining clinically meaningful change in health-related quality of life. *J Clin Epidemiol,* 56, 395-407.

Crosby RD, Kolotkin RL, Williams GR. (2004) An integrated method to determine meaningful changes in health-related quality of life. *J Clin Epidemiol,* 57, 1153-60.

Dagnan D, Jahoda A, McDowell K, Masson J, Banks P, Hare D. (2008) The psychometric properties of the Hospital Anxiety and Depressions Scale adapted for use with people with intellectual disabilities. *J Intell Disabil Res,* 52, 942-9.

Dawkins N, Cloherty ME, Gracey F, Evans JJ. (2006) The factor structure of the Hospital Anxiety and Depression Scale in acquired brain injury. *Brain Injury,* 20, 1235-9.

De Vet HC, Ostelo RW, Terwee CB, van der Roer N, Knol DL, Beckerman H, Boers M, Bouter LM. (2007) Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. *Qual Life Res*, 16, 131-42.

De Vet HC, Terluin B, Knol DL, Roorda LD, Mokkink LB, Ostelo RW, Hendriks EJ, Bouter LM, Terwee CB. (2010) Three ways to quantify uncertainty in individually applied "minimally important change" values. *J Clin Epidemiol,* 63, 37-45.

De Vet HC, Terwee CB, Mokkink LB, Knol DL. (2011) Measurement in Medicine: A Practical Guide, Practical Guides to Biostatistics and Epidemiology. Cambridge University Press.

Droesbeke JJ, Lejeune M, Saporta G. (2005) Modèles statistiques pour données qualitatives. Technip, Paris.

Dumas O, Siroux V, Le Moual N, Varraso R. (2014) Causal analysis approaches in epidemiology. *Rev Epidemiol Sante Publique*, 62, 53-63.

Ellwood PM. (1988) Shattuck lecture--outcomes management. A technology of patient experience. *N Engl J Med*, 318, 1549-56.

Embretson SE. (1991) A multidimensional latent trait model for measuring learning and change. *Psychometrika,* 56, 495-515.

Embretson SE, Reise SP. (2000) Item Response Theory for Psychologists. L. Erlbaum Associates.

Evans RG, Marmor TR. (1996) Etre ou ne pas être en bonne santé: biologie et déterminants sociaux de la maladie. John Libbey Eurotext.

Evans T. (2001) Challenging Inequities in Health: From Ethics to Action. Oxford University Press.

Everitt BS. (1975) Multivariate analysis: the need for data, and other problems. *Brit J Psychiat,* 126, 237-40.

Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ. (1999) Evaluating the Use of Exploratory Factor Analysis in Psychological Research. *Psychol Methods* 4, 272-99.

Falissard B. (2008) Mesurer la subjectivité en santé: Perspective méthodologique et statistique. Masson.

Fan X, Thompson B. (2001) Confidence Intervals about Score Reliability Coefficients, Please: An EPM Guideline Editorial. *Educ Psychol Meas*, 61, 517-31.

Farrell GA. (1998) The mental health of hospital nurses in Tasmania as measured by the 12-item General Health Questionnaire. *J Adv Nurs,* 28, 707-12.

Fedorowicz VJ, Falissard B, Foulon C, Dardennes R, Divac SM, Guelfi JD, Rouillon F. (2007) Factors associated with suicidal behaviors in a large French sample of inpatients with eating disorders. *Int J Eat Disorder*, 40, 589-95.

Feldt LS. (1965) The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika*, 30, 357-70.

Ferreira ML, Herbert RD, Ferreira PH, Latimer J, Ostelo RW, Nascimento DP, Smeets RJ. (2011) A critical review of methods used to determine the smallest worthwhile effect of interventions for low back pain. *J Clin Epidemiol*, 65, 253-61.

Floyd FJ, Widaman KF. (1995) Factor analysis in the development and refinement of clinical assessment instruments. *Psychol Assessment*, 7, 286-99.

Ford JK, MacCallum RC, Tait M. (2006) The application of exploratory factor analysis in applied psychology: a critical review and analysis. *Personnel Psychology*, 39, 291-314.

Fresan A, De la Fuente-Sandoval C, Loyzaga C, Garcia-Anaya M, Meyenberg N, Nicolini H, Apiquian R. (2005) A forced five-dimensional factor analysis and concurrent validity of the Positive and Negative Syndrome Scale in Mexican schizophrenic patients. *Schizophr Res*, 72, 123-9.

Friedman S, Samuelian JC, Lancrenon S, Even C, Chiarelli P. (2001) Three-dimensional structure of the Hospital Anxiety and Depression Scale in a large French primary care population suffering from major depression. *Psychiatry Res*, 104, 247-57.

Gabryelewicz T, Styczynska M, Pfeffer A, Wasiak B, Barczak A, Luczywek E, Androsiuk W, Barcikowska M. (2004) Prevalence of major and minor depression in elderly persons with mild cognitive impairment--MADRS factor analysis. *Int J Geriatr Psych*, 19, 1168-72.

Gaito J. (1980) Measurement Scales and Statistics: Resurgence of an old Misconception. *Psychol Bull*, 87, 564-67.

Galera C, Cote SM, Bouvard MP, Pingault JB, Melchior M, Michel G, Boivin M, Tremblay RE. (2012) Early risk factors for hyperactivity-impulsivity and inattention trajectories from age 17 months to 8 years. *Arch Gen Psychiatry*, 68, 1267-75.

Galinowski A, Lehert P. (1995) Structural validity of MADRS during antidepressant treatment. *Int Clin Psychopharm*, 10, 157-61.

Genolini C, Falissard B. (2010) KmL: k-means for longitudinal data. *Comput Stat,* 25, 317-28.

Goodman LA. (1974) Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models. *Biometrika*, 61, 215-31.

Gorenstein C, Andrade L, Vieira Filho AH, Tung TC, Artes R. (1999) Psychometric properties of the Portuguese version of the Beck Depression Inventory on Brazilian college students. *J Clin Psychol,* 55, 553-62.

Gorsuch RL. (1983) Factor Analysis, 2nd ed. Lawrence Erlbaum Associates, London.

Green SA, Yang Y. (2009) Reliability of summed item scores using structural equation modeling: an alternative to coefficient alpha. *Psychometrika,* 74, 155-67.

Greenland S, Pearl J, Robins J. (1999) Causal Diagrams for Epidemiologic Research. *Epidemiology,* 10, 37-48.

Grunebaum MF, Keilp J, Li S, Ellis SP, Burke AK, Oquendo MA, Mann JJ. (2005) Symptom components of standard depression scales and past suicidal behavior. *J Affect Disorders,* 87, 73-82.

Guadagnoli E, Velicer WF. (1988) Relation of sample size to the stability of component patterns. *Psychol Bull*, 103, 265-75.

Guillemin F, Rat AC, Roux CH, Fautrel B, Mazieres B, Chevalier X, Euller-Ziegler L, Fardellone P, Verrouil E, Morvan J, Pouchot J, Coste J, Saraux A, KHOALA cohort study. (2012) The KHOALA cohort of knee and hip osteoarthritis in France. *Joint Bone Spine,* 79, 597-603.

Guyatt GH, Cook DJ. (1994) Health status, quality of life, and the individual. *JAMA*, 272, 630.

Hallquist M, Wiley J. (2013) Package "MplusAutomation": Automating Mplus Model Estimation and Interpretation. CRAN.

Hambleton RK, Swaminathan H, Rogers HJ. (1991) Fundamentals of Item Response Theory. Sage Publications.

Hamilton M. (1960) A rating scale for depression. *J Neurol Neurosurg Psychiatry,* 23, 56-62.

Hankins M. (2008) The factor structure of the twelve item General Health Questionnaire (GHQ-12): the result of negative phrasing? *Clinical Practice and Epidemiology in Mental Health*, 4, 10-8.

Hardouin JB. (2005) SIMIRT: Stata module to process data generated by IRT models. Boston College Department of Economics.

Hardouin JB, Audureau E, Leplège A, Coste J. (2012) Spatio-temporal Rasch analysis of quality of life outcomes in the French general population. Measurement invariance and group comparisons. *BMC Medical Research Methodology,* 12, 182-93.

Hardouin JB, Bonnaud-Antignac A, Sébille V. (2011) Nonparametric item response theory using Stata. *Stata Journal,* 11, 30-51.

Harvey PD, Davidson M, White L, Keefe RS, Hirschowitz J, Mohs RC, Davis KL. (1996) Empirical evaluation of the factorial structure of clinical symptoms in schizophrenia: effects of typical neuroleptics on the brief psychiatric rating scale. *Biol Psychiatry*, 40, 755-60.

Hays RD, Lipscomb J. (2007) Next steps for use of item response theory in the assessment of health outcomes. *Qual Life Res*, 16 Suppl 1, 195-9.

Hays RD, Morales LS, Reise SP. (2000) Item response theory and health outcomes measurement in the 21st century. *Med Care,* 38, II28-42.

Hays RD, Woolley JM. (2000) The concept of clinically meaningful difference in health-related quality-of-life research. How meaningful is it? *Pharmacoeconomics,* 18, 419-23.

Helm HW Jr, Boward MD. (2003). Factor structure of the Beck Depression Inventory in a university sample. *Psychol Rep*, 92, 53-61.

Hernán MA, Cole SR. (2009) Invited Commentary: Causal Diagrams and Measurement Bias. *Am J Epidemiol*, kwp293, 4p.

Hernán MA, Hernández-Díaz S, Robins JM. (2004). A Structural Approach to Selection Bias. *Epidemiology*, 15, 615-25.

Hernán MA, Hernández-Díaz S, Werler MM, Mitchell AA. (2002) Causal Knowledge as a Prerequisite for Confounding Evaluation: An Application to Birth Defects Epidemiology. *Am J Epidemiol*, 155, 176-84.

Hill AB. (1965) The Environment and Disease: Association or Causation? *Proc R Soc Med*, 58, 295-300.

Hogarty KY, Hines CV, Kromrey JD, Ferron JM, Mumford KR. (2005) The Quality of Factor Solutions in Exploratory Factor Analysis: The Influence of Sample Size, Communality, and Overdetermination. *Educ Psychol Meas*, 65, 202-6.

Honey GD, Sharma T, Suckling J, Giampietro V, Soni W, Williams SC, Bullmore ET. (2003) The functional neuroanatomy of schizophrenic subsyndromes. *Psychol Med*, 33, 1007-18.

Horton M, Tennant A. (2011). Applying Rasch analysis to the SF-36 physical function scale: effect of dependent items. *Trials*, 12, A75.

Hu L, Bentler PM. (1999) Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct Equ Modeling*, 6, 1-55.

Hu LT, Bentler PM, Kano Y. (1992) Can test statistics in covariance structure analysis be trusted? *Psychol Med*, 112, 351-62.

Iwata N, Mishima N, Okabe K, Kobayashi N, Hashiguchi E, Egashira K. (2000) Psychometric properties of the State-Trait Anxiety Inventory among Japanese clinical outpatients. *J Clin Psychol*, 56, 793-806.

Iwata N, Mishima N, Shimizu T, Mizoue T, Fukuhara M, Hidano T, Spielberger CD. (1998) Positive and negative affect in the factor structure of the State-Trait Anxiety Inventory for Japanese workers. *Psychol Rep*, 82, 651-6.

Jackson JE. (1991) A user's guide to principal components. John Wiley & sons, New York.

Jaeschke R, Singer J, Guyatt GH. (1989) Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials*, 10, 407-15.

Jensen MP, Chen C, Brugger AM. (2003) Interpretation of visual analog scale ratings and change scores: a reanalysis of two clinical trials of postoperative pain. *J Pain*, 4, 407-14.

Jo SA, Park MH, Jo I, Ryu SH, Han C. (2007) Usefulness of Beck Depression Inventory (BDI) in the Korean elderly population. *Int J Geriatr Psych*, 22, 218-23.

Jones DJ, Runyan DK, Lewis T, Litrownik AJ, Black MM, Wiley T, English DE, Proctor LJ, Jones BL, Nagin DS. (2012) Trajectories of childhood sexual abuse and early adolescent HIV/AIDS risk behaviors: the role of other maltreatment, witnessed violence, and child gender. *J Clin Child Adolesc Psychol*, 39, 667-80.

Jöreskog KG. (1969) A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34, 183-202.

Jung T, Wickrama KS. (2008). An Introduction to Latent Class Growth Analysis and Growth Mixture Modeling. *Social and Personality Psychology Compass,* 2, 302-17.

Kabacoff RI, Segal DL, Hersen M, Van Hasselt VB. (1997) Psychometric properties and diagnostic utility of the Beck Anxiety Inventory and the State-Trait Anxiety Inventory with older adult psychiatric outpatients. *J Anxiety Disord,* 11, 33-47.

Kay SR, Sevy S. (1990) Pyramidical model of schizophrenia. *Schizophr Bull*, 16, 537-45.

Kemmler G, Giesinger J, Holzner B. (2011) Clinically relevant, statistically significant, or both? Minimal important change in the individual subject revisited. *J Clin Epidemiol*, 64, 1467-8.

Kilic C, Rezaki M, Rezaki B, Kaplan I, Ozgen G, Sagduyu A, Ozturk MO. (1997) General Health Questionnaire (GHQ12 & GHQ28): psychometric properties and factor structure of the scales in a Turkish primary care sample. *Soc Psychiatry Psychiatr Epidemiol*, 32, 327-31.

Killgore WD. (1999) Empirically derived factor indices for the Beck Depression Inventory. *Psychol Rep*, 84, 1005-13.

Kline RB. (2005) Principles and Practice of Structural Equation Modeling, Methodology in the social sciences. Guilford Press.

Kline T. (2005) Psychological Testing: A Practical Approach to Design and Evaluation. SAGE Publications.

Lachar D, Bailley SE, Rhoades HM, Espadas A, Aponte M, Cowan KA, Gummattira P, Kopecky CR, Wassef A. (2001) New subscales for an anchored version of the Brief

Psychiatric Rating Scale: construction, reliability, and validity in acute psychiatric admissions. *Psychol Assessment*, 13, 384-95.

Lalonde M. (1974) Nouvelle perspective de la santé des canadiens. Un document de travail. Ministère de la Santé Nationale et du Bien-être Social, Ottawa.

Lancon C, Reine G, Llorca PM, Auquier P. (1999) Validity and reliability of the French-language version of the Positive and Negative Syndrome Scale (PANSS). *Acta Psychiatr Scand,* 100, 237-43.

Lauridsen HH, Hartvigsen J, Manniche C, Korsholm L, Grunnet-Nilsson N. (2006) Responsiveness and minimal clinically important difference for pain and disability instruments in low back pain patients. *BMC Musculoskelet Disord,* 7, 82-98.

Lee KH, Harris AW, Loughland CM, Williams LM. (2003) The five symptom dimensions and depression in schizophrenia. *Psychopathology*, 36, 226-33.

Leplège A, Ecosse E, Coste J, Pouchot J, Perneger T. (2001) Le questionnaire MOS SF-36: Manuel de l'utilisateur et guide d'interprétation des scores. Estem.

Liang MH. (2000) Longitudinal construct validity: establishment of clinical meaning in patient evaluative instruments. *Med Care*, 38, II84-90.

Lindenmayer JP, Czobor P, Volavka J, Lieberman JA, Citrome L, Sheitman B, McEvoy JP, Cooper TB, Chakos M. (2004) Effects of atypical antipsychotics on the syndromal profile in treatment-resistant schizophrenia. *J Clin Psychiat*, 65, 551-6.

Loo R. (1983) Caveat on Sample Sizes in Factor Analysis. *Percept Mot Skills,* 56, 371-4.

Lopez-Castedo A, Fernandez L. (2005) Psychometric properties of the Spanish version of the 12-item General Health Questionnaire in adolescents. *Percept Mot Skills*, 100, 676-80.

Loza B, Kucharska-Pietura K, Kopacz G, Debowska G. (2003) Factor structure of paranoid schizophrenia: a prospective study. *Psychopathology,* 36, 132-41.

Lykouras L, Oulis P, Psarros K, Daskalopoulou E, Botsis A, Christodoulou GN, Stefanis C. (2000) Five-factor model of schizophrenic psychopathology: how valid is it? *Eur Arch Psy Clin N,* 250, 93-100.

MacCallum RC, Widaman KF, Zhang S, Hong S. (1999) Sample Size in Factor Analysis. *Psychol Methods,* 4, 84-99.

MacMahon B, Pugh TF. (1970) Epidemiology: Principles and Methods. Little, Brown Book Group Limited.

Masters GN. (1982) A Rasch model for partial credit scoring. *Psychometrika,* 47, 149-174.

McHorney CA. (1997) Generic health measurement: past accomplishments and a measurement paradigm for the 21st century. *Ann Intern Med,* 127, 743-50.

Mellenbergh GJ. (1996) Measurement precision in test score and item response models. *Psychological Methods,* 1, 293-299.

Michell J. (1986) Measurement Scales and Statistics: A Clash of Paradigms. *Psychol Bull* 100, 398-407.

Millsap RE. (2011) Statistical Approaches to Measurement Invariance. Routledge.

Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC. (2010) The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol*, 63, 737-745.

Moustaki I, Knott M. (2000) Generalized latent trait models. *Psychometrika,* 65, 391-411.

Mundfrom DJ, Shaw DG, Ke TL. (2005) Minimum Sample Size Recommendations for Conducting Factor Analyses. *International Journal of Testing*, 5, 159-68.

Munoz DJ, Chen E, Fischer S, Roehrig M, Sanchez-Johnson L, Alverdy J, Dymek-Valentine M, le Grange D. (2007) Considerations for the use of the Beck Depression Inventory in the assessment of weight-loss surgery seeking patients. *Obes Surg,* 17, 1097-101.

Muthén B, Muthén LK. (2000) Integrating Person-Centered and Variable-Centered Analyses: Growth Mixture Modeling With Latent Trajectory Classes. *Alcoholism: Clinical and Experimental Research*, 24, 882-891.

Muthén BO. (2002). Beyond SEM: general latent variable modeling. *Behaviormetrika*, 29, 81-117.

Muthén LK, Muthén BO. (1998-2012) MPlus. Statistical Analysis With Latent Variables. User's Guide. Seventh Edition.

Nagin D. (2005) Group-based Modeling Of Development. Harvard University Press.

National Institutes of Health - Office of Behavioral & Social Sciences Research. (2010) E-Source, Behavioral & Social Sciences Research, Patient Reported Outcomes. URL: http://www.esourceresearch.org/eSourceBook/PatientReportedOutcomes/3ReasonstoMeasurePROs/tabid/150/Default.aspx, visité le : 28/08/2014.

Nelder JA, Wedderburn RWM. (1972) Generalized Linear Models. *J R Stat Soc Ser A-G*, 135, 370-84.

Nguyen Thi PL, Briançon S, Empereur F, Guillemin F. (2002) Factors determining inpatient satisfaction with care. *Soc Sci Med,* 54, 493-504.

Norman GR, Stratford P, Regehr G. (1997) Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. *J Clin Epidemiol*, 50, 869-79.

Norquist JM, Fitzpatrick R, Dawson J, Jenkinson C. (2004) Comparing alternative Rasch-based methods vs raw scores in measuring change in health. *Med Care*, 42, I25-36.

Novick MR. (1966) The axioms and principal results of classical test theory. *J Math Psychol*, 3, 1-18.

Nunnaly JC. 1978. Psychometric theory, 2nd ed. McGraw-Hill, New York.

Olden M, Rosenfeld B, Pessin H, Breitbart W. (2009) Measuring depression at the end of life: is the Hamilton Depression Rating Scale a valid instrument? *Assessment*, 16, 43-54.

Organisation Mondiale de la Santé. (1986) Charte d'Ottawa pour la promotion de la santé. Organisation Mondiale de la Santé, Ottawa (Canada).

Organisation Mondiale de la Santé. (1946) Préambule à la Constitution de l'Organisation mondiale de la Santé, tel qu'adopté par la Conférence internationale sur la Santé, New York, 19-22 juin 1946 ; signé le 22 juillet 1946 par les représentants de 61 Etats et entré en vigueur le 7 avril 1948. Actes officiels de l'Organisation mondiale de la Santé 2, 100.

Organisation Mondiale de la Santé. (1996) Health Interviews Surveys. Toward the international harmonisation of methods and instruments. URL: http://www.euro.who.int/__data/assets/pdf_file/0017/111149/E72841.pdf. Visité le : 28/08/2014.

Pallant JF, Bailey CM. (2005) Assessment of the structure of the Hospital Anxiety and Depression Scale in musculoskeletal patients. *Health Qual Life Out*, 3, 82-91.

Paolaggi JB, Coste J. (2001) Le raisonnement médical: de la science à la pratique clinique. Editions Estem.

Parker RD, Flint EP, Bosworth HB, Pieper CF, Steffens DC. (2003). A three-factor analytic model of the MADRS in geriatric depression. *Int J Geriatr Psych*, 18, 73-7.

Peterson RA. (1994). A Meta-Analysis of Cronbach's Coefficient Alpha. *J Consum Res*, 21, 381-91.

Pillemer DB, White SH. (2005) Developmental Psychology and Social Change: Research, History, and Policy. Cambridge University Press.

Powell R. (2003) Psychometric properties of the Beck Depression Inventory and the Zung Self Rating Depression Scale in adults with mental retardation. *Ment Retard* 41, 88-95.

Pryor LE, Tremblay RE, Boivin M, Touchette E, Dubois L, Genolini C, Liu X, Falissard B, Cote SM. (2011) Developmental trajectories of body mass index in early childhood and their risk factors: an 8-year longitudinal study. *Arch Pediatr Adolesc Med*, 165, 906-12.

Purcell A, Fleming J, Bennett S, Burmeister B, Haines T. (2010) Determining the minimal clinically important difference criteria for the Multidimensional Fatigue Inventory in a radiotherapy population. *Support Care Cancer,* 18, 307-15.

R Development Core Team. (2008) R: A language and environment for statistical computing. R Foundation for statistical Computing, Vienna, Austria.

Revelle W. (2008) R Documentation: Procedures for Personality, Psychometric, and Psychological Research. Help pages for package "psych" version 1.0-58: Principal Axis Factor Analysis. CRAN.

Revicki D, Hays RD, Cella D, Sloan J. (2008) Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol*, 61, 102-9.

Revicki DA, Cella D, Hays RD, Sloan JA, Lenderkin WR, Aaronson NK. (2006) Responsiveness and minimal important differences for patient reported outcomes. *Health Qual Life Out,* 4, 70-5.

Revicki DA, Cella DF. (1997) Health status assessment for the twenty-first century: item response theory, item banking and computer adaptive testing. *Qual Life Res,* 6, 595-600.

Rijmen F, Tuerlinckx F, De Boeck P, Kuppens P. (2003). A nonlinear mixed model framework for item response theory. *Psychol Methods*, 8, 185-205.

Roger VL. (2011) Outcomes research and epidemiology: the synergy between public health and clinical practice. *Circ Cardiovasc Qual Outcomes*, 4, 257-9.

Rothman KJ, Greenland S, Lash TL. (2008). Modern Epidemiology. Wolters Kluwer Health/Lippincott Williams & Wilkins.

Rouquette A, Côté SM, Pryor LE, Carbonneau R, Vitaro F, Tremblay RE. (2014) Cohort Profile: The Quebec Longitudinal Study of Kindergarten Children (QLSKC). *Int. J. Epidemiol*, 43, 23-33.

Rubin HR, Ware JE Jr, Nelson EC, Meterko M. (1990) The Patient Judgments of Hospital Quality (PJHQ) Questionnaire. *Med Care*, 28, S17-8.

Rumeau-Rouquette C, Bréart G, Padieu R. (1986) Méthodes en épidémiologie, Médecine Sciences. Flammarion, Paris.

Rupp A, Zumbo D. (2006) Understanding Parameter Invariance in Unidimensional IRT Models. *Educ Psychol Meas*, 66, 63-84.

Ryan ND, Puig-Antich J, Ambrosini P, Rabinovich H, Robinson D, Nelson B, Iyengar S, Twomey J. (1987) The clinical picture of major depression in children and adolescents. *Arch Gen Psychiatry*, 44, 854-861.

Salamero M, Marcos T, Gutierrez F, Rebull E. (1994) Factorial study of the BDI in pregnant women. *Psychol Med*, 24, 1031-5.

Salokangas RK, Honkonen T, Stengard E, Koivisto AM. (2002) Symptom dimensions and their association with outcome and treatment setting in long-term schizophrenia. Results of the DSP project. *Nord J Psychiat*, 56, 319-27.

Samsa G, Edelman D, Rothman ML, Williams GR, Lipscomb J, Matchar D. (1999) Determining clinically important differences in health status measures: a general approach with illustration to the Health Utilities Index Mark II. *Pharmacoeconomics,* 15, 141-55.

Sanchez-Lopez Mdel P, Dresch V. (2008) The 12-Item General Health Questionnaire (GHQ-12): reliability, external validity and factor structure in the Spanish population. *Psicothema*, 20, 839-43.

SAS Institute Inc. (2010) Procedures Guides. SAS Institute Inc, Cary, NC.

Seegers V, Petit D, Falissard B, Vitaro F, Tremblay RE, Montplaisir J, Touchette E. (2011) Short sleep duration and body mass index: a prospective longitudinal study in preadolescence. *Am J Epidemiol*, 173, 621-9.

Serretti A, Jori MC, Casadei G, Ravizza L, Smeraldi E, Akiskal H. (1999) Delineating psychopathologic clusters within dysthymia: a study of 512 out-patients without major depression. *J Affect Disorders*, 56, 17-25.

Shafer AB. (2006) Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. *J Clin Psychol,* 62, 123-46.

Shek DT. (1990) Reliability and factorial structure of the Chinese version of the Beck Depression Inventory. *J Clin Psychol*, 46, 35-43.

Sijtsma K. (2008) On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika,* 74, 107-20.

Sijtsma K, Molenaar IW. (2002) Introduction to Nonparametric Item Response Theory. SAGE Publications.

Sloan JA. (2005) Assessing the minimally clinically significant difference: scientific considerations, challenges and solutions. *COPD*, 2, 57-62.

Smith AB, Rush R, Fallowfield LJ, Velikova G, Sharpe M. (2008) Rasch fit statistics and sample size considerations for polytomous data. *BMC Med Res Methodol*, 8, 33-40.

Smith AB, Selby PJ, Velikova G, Stark D, Wright EP, Gould A, Cull A. (2002) Factor analysis of the Hospital Anxiety and Depression Scale from a large cancer population. *Psychol Psychother*, 75, 165-76.

Spearman C. (1904) "General Intelligence" Objectively Determined and Measured. *Am J Psychol,* 15, 201-92.

Sprangers MAG, Schwartz CE. (1999) Integrating response shift into health-related quality of life research: a theoretical model. *Soc Sci Med,* 48, 1507-15.

StataCorp LP. (2012) Stata Statistical Software: Release 12.1. College Station, TX.

Steer RA, Kumar G, Ranieri WF, Beck AT. (1995) Use of the Beck Anxiety Inventory with adolescent psychiatric outpatients. *Psychol Rep*, 76, 459-65.

Steer RA, Rissmiller DJ, Ranieri WF, Beck AT. (1993) Structure of the computer-assisted Beck Anxiety Inventory with psychiatric inpatients. *J Pers Assess*, 60, 532-42.

Stevens SS. (1946) On the Theory of Scales of Measurement. *Science*, 103, 677-80.

Stratford PW, Binkley J, Solomon P, Finch E, Gill C, Moreland J. (1996) Defining the minimum level of detectable change for the Roland-Morris questionnaire. *Phys Ther*, 76, 359-68.

Stratford PW, Binkley JM, Riddle DL, Guyatt GH. (1998) Sensitivity to change of the Roland-Morris Back Pain Questionnaire: part 1. *Phys Ther,* 78, 1186-96.

Stucki G, Daltroy L, Katz JN, Johannesson M, Liang MH. (1996) Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. *J Clin Epidemiol,* 49, 711-17.

Sullivan M. (2003) The new subjective medicine: taking the patient's point of view on health care and health. *Soc Sci Med*, 56, 1595-604.

Szklo M, Nieto J. (2007) Epidemiology: Beyond the Basics. Jones & Bartlett Learning.

Ten Klooster PM, Drossaers-Bakker KW, Taal E, van de Laar MA. (2006) Patient-perceived satisfactory improvement (PPSI): interpreting meaningful change in pain from the patient's perspective. *Pain*, 121, 151-7.

Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, de Vet HC. (2007) Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol,* 60, 34-42.

Terwee CB, Roorda LD, Dekker J, Bierma-Zeinstra SM, Peat G, Jordan KP, Croft P, de Vet HC. (2010) Mind the MIC: large variation among populations and methods. *J Clin Epidemiol*, 63, 524-34.

Townsend JB, Ashby FG. (1984) Measurement Scales and Statistics: The Misconception Misconceived. *Psychol Bull*, 96, 394-401.

Tu YK. (2009). Commentary: Is structural equation modelling a step forward for epidemiologists? *Int. J. Epidemiol*, 38, 549-51.

Tu YK, Tilling K, Sterne JA, Gilthorpe MS. (2013) A critical evaluation of statistical approaches to examining the role of growth trajectories in the developmental origins of health and disease. *Int J Epidemiol*, 42, 1327-39.

Tubach F, Ravaud P, Baron G, Falissard B, Logeart I, Bellamy N, Bombardier C, Felson D, Hochberg M, van der Heijde D, Dougados M. (2005) Evaluation of clinically relevant changes in patient reported outcomes in knee and hip osteoarthritis: the minimal clinically important improvement. *Ann Rheum Dis*, 64, 29-33.

US Department of Health and Human Services. (2009) Guidance for industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. URL: http://www.fda.gov/downloads/Drugs/Guidances/UCM193282.pdf. Visité le : 28/08/2014.

Uslu RI, Kapci EG, Oncu B, Ugurlu M, Turkcapar H. (2008) Psychometric properties and cut-off scores of the Beck Depression Inventory-II in Turkish adolescents. *J Clin Psychol Med*, S 15, 225-33.

Van Walraven C, Mahon JL, Moher D, Bohm C, Laupacis A. (1999) Surveying physicians to determine the minimal important difference: implications for sample-size calculation. *J Clin Epidemiol*, 52, 717-23.

VanderWeele TJ, Hernán MA. (2012) Results on Differential and Dependent Measurement Error of the Exposure and the Outcome Using Signed Directed Acyclic Graphs. *Am J Epidemiol,* 175, 1303-10.

Velicer WF, Fava JL. (1998) Effects of variable and subject sampling on factor pattern recovery. *Psychol Methods,* 3, 231-51.

Velicer WF, Peacock AC, Jackson DN. (1982) A Comparison of Component and Factor Patterns: A Monte Carlo Approach. *Multivar Behav Res*, 17, 371-88.

Velleman PF, Wilkinson L. (1993) Nominal, Ordinal, Interval, and Ratio Typologies are Misleading. *Am Stat*, 47, 65-72.

Ventura J, Nuechterlein KH, Subotnik KL, Gutkind D, Gilbert EA. (2000) Symptom dimensions in recent-onset schizophrenia and mania: a principal components analysis of the 24-item Brief Psychiatric Rating Scale. *Psychiatry Res*, 97, 129-35.

Villalta-Gil V, Vilaplana M, Ochoa S, Dolz M, Usall J, Haro JM, Almenara J, Gonzalez JL, Lagares C. (2006) Four symptom dimensions in outpatients with schizophrenia. *Compr Psychiatry*, 47, 384-8.

Walton DM, Eilon-Avigdor Y, Wonderham M, Wilk P. (2014) Exploring the Clinical Course of Neck Pain in Physical Therapy: A Longitudinal Study. Arch Phys Med Rehab 95, 303-8.

Wang YP, Andrade LH, Gorenstein C. (2005). Validation of the Beck Depression Inventory for a Portuguese-speaking Chinese community in Brazil. *Braz J Med Biol Res*, 38, 399-408.

Ware JE Jr, Gandek B. (1998) Overview of the SF-36 Health Survey and the International Quality of Life Assessment (IQOLA) Project. *J Clin Epidemiol*, 51, 903-12.

Ware JE Jr, Sherbourne CD. (1992) The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care*, 30, 473-83.

Werneke U, Goldberg DP, Yalcin I, Ustun BT. (2000) The stability of the factor structure of the General Health Questionnaire. *Psychol Med*, 30, 823-9.

Widaman KF. (1993) Common Factor Analysis Versus Principal Component Analysis: Differential Bias in Representing Model Parameters? *Multivar Behav Res*, 28, 263-311.

Woolrich RA, Kennedy P, Tasiemski T. (2006) A preliminary psychometric evaluation of the Hospital Anxiety and Depression Scale (HADS) in 963 people living with a spinal cord injury. *Psychol Health Med*, 11, 80-90.

World Health Organization. (2001) International Classification of Functionning, Disability and Health. World Health Organization, Geneva.

Wright BD. (1996) Comparison requires stability. *Rash Measurement Trans*, 10, 506.

Yeates KO, Taylor HG, Rusin J, Bangert B, Dietrich A, Nuss K, Wright M, Nagin DS, Jones BL. (2009) Longitudinal trajectories of postconcussive symptoms in children with mild traumatic brain injuries and their relationship to acute clinical status. *Pediatrics*, 123, 735-43.

Yost KJ, Sorensen MV, Hahn EA, Glendenning GA, Gnanasakthy A, Cella D. (2005) Using multiple anchor- and distribution-based estimates to evaluate clinically meaningful change on the Functional Assessment of Cancer Therapy-Biologic Response Modifiers (FACT-BRM) instrument. *Value Health*, 8, 117-27.

ANNEXES

Annexe 1 : Sample Size Requirements for the Internal Validation of Psychiatric Scales (Article 1)

Annexe 2 : The Minimal Clinically Important Difference determined using Item Response Theory Models: an attempt to solve the issue of the association with baseline score (Article 2)

ANNEXE 1: SAMPLE SIZE REQUIREMENTS FOR THE INTERNAL VALIDATION OF PSYCHIATRIC SCALES (ARTICLE 1)

Alexandra ROUQUETTE, Bruno FALISSARD

Abstract:

The ratio of subjects to variables (*N/p*), as a rule to calculate the sample size required in internal validity studies on measurement scales, has been recommended without any strict theoretical or empirical basis being provided. The purpose of the present study was to develop a tool to determine sample size for these studies in the field of psychiatry. First, a literature review was carried out to identify the distinctive features of psychiatric scales. Then, two simulation methods were developed to generate data according to: 1/ the model for factor structure derived from the literature review and 2/ a real dataset. This enabled the study of the quality of solutions obtained from principal component analysis or Exploratory Factor Analysis (EFA) on various sample sizes. Lastly, the influence of sample size on the precision of Cronbach's alpha coefficient was examined. The *N/p* ratio rule is not upheld by this study: short scales do not allow smaller sample size. As a rule of thumb, if one's aim is to reveal the factor structure, a minimum of 300 subjects is generally acceptable but should be increased when the number of factors within the scale is large, when EFA is used and when the number of items is small.

Keywords: Sample size, validation studies, factor analysis, questionnaires, psychiatry.

Introduction

One of the most critical methodological issues when designing a study and planning the statistical analysis, is the number of subjects to include. Generally, the sample size is based on the power of a statistical test of hypothesis. In descriptive studies, this approach cannot be used, and it is usually the range of the confidence interval of a given parameter which determines sample size. This is likely to be the case in internal validity studies of measurement scales in which, traditionally, two types of parameters are of interest: Cronbach's alpha coefficient (α) which assesses reliability, and factor analysis loadings which explore the dimensional structure of the scale. In practice, these loadings are estimated either by Principal Component Analysis (PCA) or by Exploratory Factor Analysis (EFA). A formula for the confidence interval of Cronbach's alpha coefficient was developed by Feldt in the 1960s (Fan *et al.*, 2001; Feldt, 1965). The sample size required for a desired precision of this coefficient can, therefore, be easily assessed. In fact, the difficulty in establishing a simple rule for sample size calculation in internal validity studies arises from the use of factor analysis.

Many recommendations regarding sample size in factor analysis have been made, but none are founded on a strict theoretical or empirical basis. The most widely used rule uses the ratio of the number of subjects (N) to the number of items (p), and this varies from three to 10 depending on authors (Cattell, 1978; Everitt, 1975; Gorsuch, 1983; Nunnaly, 1978). Other authors have suggested an absolute minimum sample size of 50 to 500 to enable factor analysis (Aleamoni, 1973; Comrey, 1978; Comrey *et al.*, 1992; Loo, 1983). Given these various recommendations and their lack of documented explanation, some researchers have put them to the test by studying the consequences of using factor analysis on insufficient sample sizes. They all found that, in addition to N, two other parameters are important to obtain accurate and stable solutions: firstly the ratio of the number of variables to the number of factors (ratio p/M, which is an indicator of 'factor overdetermination', a concept defined by MacCallum in 1999 as the degree to which each factor is clearly represented by a sufficient number of variables, at least three or four); and secondly the level of factor loadings (which reflects the level of communalities, the communality of a variable being the portion of the variance that a variable shares with the common factors). The lower the p/M ratio and the

factor loading level, the larger the sample size required for a given accuracy and stability of solutions obtained from factor analysis (Guadagnoli *et al.*, 1988; Hogarty *et al.*, 2005; MacCallum *et al.*, 1999; Mundfrom *et al.*, 2005; Velicer *et al.*, 1998). All these studies have shown that sample size partly depends on the nature of the data: their 'strength'. Strong data in factor analysis means uniformly high communalities without cross-loadings, plus several variables loading strongly on each factor (Costello *et al.*, 2005; Fabrigar *et al.*, 1999). The stronger the data, the smaller the sample size required. It does not therefore seem possible to recommend a general rule for sample size calculation that is valid in all the fields to which psychometric procedures apply.

However, in each field, there are distinctive features. In psychiatry, factor loading values are usually close to 0.6, the *p/M* ratio can vary from three to 20 or more, depending on scales, and the number of items is often different for each factor within a scale (Dawkins *et al.*, 2006; Gabryelewicz *et al.*, 2004; Iwata *et al.*, 2000; Loza *et al.*, 2003). Another characteristic observed in psychiatric scales is the shape of the scree plot. Unidimensionality is rare, and usually there is a first dimension representing a large part of the variance contained in the data (30 to 35%), and then there are one or more other dimensions explaining smaller and decreasing proportions of variance (from 15 to 5%) (Chapman *et al.*, 2009; Sanchez-Lopez Mdel *et al.*, 2008; Uslu *et al.*, 2008; Villalta-Gil *et al.*, 2006). This factor structure can be explained by the presence of correlated factors or, likewise, by a two-order factor model in which a second order factor explains the pattern of correlations among the first order factors.

The unresolved methodological issue about sample size in validation studies of measurement scales can lead to erroneous conclusions being drawn if the sample is too small. Conversely, the inclusion of too many subjects in a study wastes time and resources for researchers. The main purpose of this study is therefore to use the distinctive features encountered in psychiatric scales to develop a tool for the determination of the sample size required in internal validity studies on such scales in order to guarantee an acceptable level of precision for Cronbach's alpha coefficient and, above all, accuracy and stability of the factor solution. A secondary aim is to determine the influence of the choice of PCA or EFA on the sample size required and on the accuracy of the factor solution.

Material and Methods

This study comprised three stages. The first consisted in a literature review to determine the shared characteristics of psychiatric scales. The second used simulations to study the influence of sample size on the stability and accuracy of the solutions obtained from PCA and EFA. These simulations were based, firstly, on artificial data generated according to the factor pattern observed in psychiatric scales from the literature review, and then on real data. Finally, the influence of sample size on the precision of Cronbach's alpha coefficient in the conditions encountered in psychiatry was studied.

Literature Review

10 psychiatric scales were selected taking account of the frequency of their use in clinical practice and their representativeness of different pathologies encountered in psychiatry:

- Positive And Negative Syndrome Scale (PANSS - 30 items)
- Brief Psychiatric Rating Scale (BPRS – 18 items)
- Beck Anxiety Inventory (BAI - 21 items)
- State-Trait Anxiety Inventory (STAI - 40 items)
- Hamilton Anxiety Rating Scale (HAMA - 14 items)
- Hamilton Rating Scale for Depression (HAMD - 17 items)
- Montgomery-Asberg Depression Rating Scale (MADRS - 10 items)
- Beck Depression Inventory (BDI - 21 items)
- Hospital Anxiety and Depression Scale (HADS - 14 items)
- General Health Questionnaire (GHQ – 12 items)

Articles including results of PCA or EFA concerning any of these ten scales were sought in the Medline database using the following keywords: for each scale, the "name of the scale" and/or "its abbreviation", the expressions "factor analysis" and/or "components analysis" and the article language "English" and/or "French". A pre-selection was carried out on the basis of the abstracts, and articles were then included if the following three criteria were met: the factor structure of one of the ten scales was studied using PCA or EFA; eigenvalues

or percentage of variance accounted for by each factor before rotation were specified; sample size was equal to or greater than 100.

In each article, the following data were collected: the method used for factor extraction (PCA or EFA), the rotation method used (orthogonal or oblique), the number of factors extracted, the eigenvalues or the percentage of variance accounted for by each factor before rotation, the number of items per factor and the values of the factor inter-correlations. When the loading matrix was reproduced, the mean of the salient loadings was calculated by considering only the higher value in case of cross-loadings. If several groups were studied, only the results from the largest group were considered. Likewise, if analyses were carried out on data collected at different times, only the results collected at the initial collection time were considered. All these data were recorded on the Microsoft® Office Excel 2007 spreadsheet program and descriptive statistical analyses for each of these variables were performed using R software 2.6.2 (R Development Core Team, 2008).

Simulation Studies

Simulations based on artificial data.

The simulation method developed here is based on the common factor model and is described in the appendix. To summarize, certain important points should be noted. In this simulation model, two hypotheses are set. The first is the existence of a simple structure, i.e. each item loads on a single factor and all the non-salient loadings are equal to zero. The second is that all salient loadings (λ) are equal. When a common factor model is used, responses have a normal distribution. To come closer to real-life instruments, these responses were categorised into four-class ordered variables as in a four-point Likert response pattern. The response distribution was different for each item in the scale and non-symmetrical so as to simulate floor and ceiling effects. Finally, parameters that can be controlled using this method are: the number of items (p), the number of factors (M), the number of items loading on each factor in the scale (p_m, m=1 to M), the value of salient loadings (λ), the level of the factor inter-correlations ($cor(F_m,F_{m'})$, $m \neq m'$) and the sample size (N).

For M and p, we decided to study the values usually encountered in psychiatry, i.e. scales with two, three or four factors and a number of items varying between 10 and 45 (p= 10, 15, 20, 25, 30, 35, 40 or 45). The results from the literature review then enabled the determination of the value of λ and p_m. Levels of factor inter-correlations were chosen amongst the values encountered in the literature review, and also in order to obtain the percentage of variance accounted for by each factor that was nearest to the mean of this percentage found in the review. Once all these parameter values were determined, two sets of 10 000 samples were generated for each sample size studied (N= 50, 100, 150, 200, 300, 500, 1 000) and for each condition defined by M and p. Then, PCA was performed on one set and EFA on the other. These two methods of factor extraction were followed by a promax rotation which is an oblique rotation method as recommended when factors are correlated with each other (Costello and Osborn, 2005; Fabrigar *et al.*, 1999; Floyd *et al.*, 1995). To determine the adequate sample size, three criteria were used as a threshold for good quality of the factor solution:

- standard deviation of the salient loadings obtained after rotation over the 10 000 simulations (σ_λ) below 0.05 (95% confidence interval of the salient loadings close to $\ddot{\lambda} \pm 0.1$)
- percentage of simulations in which all the items in the scale loaded on the right factor (i.e. that which is determined in the simulation model) after rotation ($R_{\%}$) greater than 90%
- the mean of percentages of items loading on the wrong factor in the scale after rotation over the 10 000 simulations ($W_{\%}$) below 1%

When EFA was performed, the percentage of simulations where Heywood cases occurred (i.e. loading estimates greater than 1.0, which occurs only with EFA) was also estimated. Finally, for either method (PCA and EFA), the mean of the salient loadings over the 10 000 simulations (μ_λ) was computed.

Simulations based on real data.

To offer a complementary perspective, a simulation study was also conducted by the aid of an important real data set of 1009 patients consecutively hospitalized between January 1988 and July 2004 in the Eating Disorder Unit of the Clinique des Maladies Mentales et de l'Encéphale at Sainte-Anne Hospital, Paris, France. Patient characteristics and procedures have been described previously in Fedorowicz *et al.*, 2007 (Fedorowicz *et al.*, 2007). We focused on two instruments, the 13-item version of the BDI (Beck *et al.*, 1961) and the 21-item version of the HAMD (Hamilton, 1960). For each of these scales, a parallel analysis was performed to determine the number of factors to extract. Next, two sets of 10 000 samples were repeatedly drawn from the entire sample (with replacement) for each sample size: 100, 200, 300, 400, 500, 600, 700 and 800. Then, PCA was performed on one set and EFA on the other, followed by a promax rotation in the case of a multidimensional instrument. The mean of the standard deviations of the loadings was then calculated over the 10 000 samples for each sample size.

These analyses were performed using R software 2.6.2. The function `princomp` was used for PCA and the loading matrix obtained was rotated using `promax` with a constant set at four (Costello and Osborn, 2005; Jackson, 1991). For EFA, the function `factanal` (with the argument `rotation=promax)`,which uses the maximum likelihood estimation procedure, was chosen for two reasons: it finds the solution with the optimal statistical properties and it is likely the most widely used method (Revelle, 2008). Finally, the draw was performed using the function `sample` and parallel analysis using the function `scree.plot` from the `psy` package.

Precision of Cronbach's alpha coefficient

The most widely cited minimum value considered as acceptable for the Cronbach's alpha coefficient is 0.7 (Fedorowicz *et al.*, 2007; Nunnaly, 1978; Peterson, 1994). We therefore chose to study the half-width of the confidence interval of this coefficient for three expected values (α=0.7, 0.8 and 0.9) in relation to p and N (same values as previously). Feldt's

formula for this confidence interval was used with type I error rate set at 0.05 (Fan and Thompson, 2001; Feldt, 1965).

Upper bound: $CI_{upper} = 1 - \left[(1-\alpha) \times \Im_{0.025, ddl_1, ddl_2}\right]$

Lower bound: $CI_{lower} = 1 - \left[(1-\alpha) \times \Im_{0.975, ddl_1, ddl_2}\right]$

where $ddl_1 = N - 1$, $ddl_2 = (N-1) \times (p-1)$ and $\Im$ represent the values of the F-distribution for percentiles 0.025 and 0.975 respectively.

Results

Psychiatric scale characteristics

The keywords used for the search in Medline database enabled the identification of 827 studies. Amongst these, 232 articles were pre-selected on the basis of the abstracts, and a total of 56 articles met the inclusion criteria. Five of these articles showed results from factor analysis on two of the scales selected for this review, which finally increased the total to 61 references. **Table 1** contains, for each scale, the total number of references included and the number of references extracting the same number of factors for each.

[Table 1 near here]

In order to estimate a pattern of factor structure encountered in psychiatric scales, the descriptive statistical analyses were carried out over all the references without considering the number of factors found in the scales. The means of percentages of variance accounted for by each factor before rotation are shown in **table 2** for each scale and a box-plot of these percentages over all the references is provided in **figure 1**.

[Table 2 and figure 1 near here]

The loadings matrix was present in 95.1% (58) of the references. The mean of the salient loadings was 0.626 with a median (med) of 0.636 and an interquartile range (IQR) of [0.587; 0.662]. This mean was 0.635 (med=0.642, IQR=[0.601; 0.671]) when the method of factor extraction was PCA (80.3% - 49 - of the references) and 0.593 (med=0.601, IQR=[0.545; 0.637]) in the case of EFA. The orthogonal rotation method was used in 63.9%

(39) of the references and the values of factor inter-correlations were reported in 34.4% (21) which represented 51 values (mean=0.356, med=0.33, IQR=[0.155; 0.535]). Concerning the *p/M* ratio, on average 7.1 items loaded on each factor in the scale (med=6, IQR=[5; 10.5]) but this number varied depending on the number and the rank of the factors present within the scale as is shown in **table 3**.

[Table 3 near here]

Sample size influence on the quality of solutions obtained using PCA or EFA

Results using artificial data

Choice of the parameter values for the simulation models. The determination of λ was based on the literature review so that λ was fixed at 0.6. Determination of the p_m values was based on the percentages shown in **table 3**. For example, in the three-factor model, the largest integer not greater than $p \times 0.45$ was chosen as the value for p_1, the largest integer not greater than $p \times 0.35$ as the value for p_2 and the remaining items loaded on the third factor. As regards the values of factor inter-correlations, they were set at 0.45 in the two-factor model, at 0.45 for $cor(F_1,F_2)$ and 0.35 for the two other inter-correlations in the three-factor model and finally, in the four-factor model, at 0.45 for $cor(F_1,F_2)$, $cor(F_2,F_4)$ and $cor(F_1,F_4)$ and 0.35 for the three other inter-correlations. **Figure 2** shows the path diagram for the three-factor simulation model with 10 items.

[Figure 2 near here]

Criteria of quality of the factor solutions. To reduce amounts of data presented in the results, only the details concerning the three criteria σ_λ, $R_{\%}$ and $W_{\%}$ in the case of a three-factor scale are shown. **Table 4** presents results when PCA was performed and **Table 5** when it was EFA. All three criteria, $\sigma_\lambda<0.05$, $R_{\%}>90\%$ and $W_{\%}<1\%$ were met when $N= 500$ if the scale contained less than 25 items, and when N=300 if the scale contained 25 items or more in the case of PCA. When EFA was performed, *N* needed to be larger to reach the thresholds: 1 000 if the scale contained less than 20 items, 500 if there were 25 items or more. For a two-factor scale, on the whole, *N* could be smaller to meet the thresholds: 300 unless the scale contained less than 30 items and EFA was used, in which case *N* needed to be 500. In contrast, with both

methods of factor analysis, a higher *N* value (500) was necessary when the scale contained four factors (and the criteria were not satisfied when *N*=1 000 in the case of EFA and *p* below 20). Concerning the percentage of simulations where Heywood cases occurred when EFA was performed, it was always under 2% whatever the number of factors in the scale with these values of *N*.

[Table 4 and table 5 near here]

In order to narrow the sample size required to meet the criteria, we interpolated values from the curves representing σ_λ in relation to *N* for the two methods of factor extraction, and each value of *p* and *M*. The junction between these curves and the line corresponding to σ_λ=0.05 allowed the determination of the sample sizes required with a precision of 50 subjects. Results are summarized in **table 6.** Numbers reported in this table were always overestimated and at these sample sizes, the two other criteria were always met.

[Table 6 near here]

Accuracy of factor solutions. **Figure 3** shows the relationship between μ_λ and *N* for each value of *p* and each method of factor extraction in the case of a three-factor scale. When PCA was used, the smaller the number of items, the greater the distance from the expected value (λ=0.6) μ_λ. There was little influence of *N*. Conversely, in the case of EFA, sample size had rather more influence and, whatever the number of items, all the curves tended towards the expected value as *N* increased. The shape of these curves was the same when there were two or four factors within the scale, but the overestimation of the value of the salient loadings was all the greater when *M* was greater in the case of PCA. Likewise, the sample size required to tend towards the expected value was also much greater when *M* was greater in the case of EFA.

[Figure 3 near here]

Standard deviation of the loadings using real data.

Due to missing data, analyses were performed on 960 (95.1%) subjects for the BDI and 817 (81.0%) subjects for the HAMD. Parallel analysis suggested extracting one factor for the BDI and three factors for the HAMD. **Figure 4** shows the mean of the standard deviations of

the loadings over the 10 000 samples in relation to sample size in the case of PCA or EFA followed by a promax rotation for each scale. For the BDI, this mean was lower than 0.05 when the sample size was equal to or greater than 100 in the case of PCA. When EFA was used, the sample size needed to be larger, i.e. around 250, to obtain a mean lower than 0.05. In the case of the HAMD, even with 800 subjects the mean of the standard deviations of the loadings was higher than 0.05.

[Figure 4 near here]

These rather unsatisfactory results found in the case of the HAMD, especially when EFA was performed, needed to be further investigated. We hypothesized that high standard deviations resulted from the possible presence of several underlying factor structures. To test this hypothesis, normal mixture modeling (function `Mclust` from the `mclust` package of the R software 2.6.2) was performed on the distribution of each salient loading of the HAMD for a sample size equal to 400 (10000 samplings). The hypothesis of a unique component was systematically rejected and the number of components which optimized the BIC, ranged from two to six with a mode equal to three (the simulation program ruled out the possibility of an artificial phenomenon of label switching).

Influence of sample size on the precision of Cronbach's alpha coefficient

The half-width of the 95% confidence interval of Cronbach's alpha coefficient in relation to N for the three expected values (α=0.7, 0.8 and 0.9) is shown in **figure 5**. Only the two extreme values for the number of items (p=10 and 45) are represented because, as can be seen from this figure, there was little influence of p on the precision of Cronbach's alpha coefficient in the conditions studied here. A half-width of 0.05 was reached when N=300 for α=0.7, 150 for α=0.8 and only 50 for α=0.9.

[Figure 5 near here]

Discussion

These simulation studies, approaching as closely as possible the conditions usually met in practice during an internal validity study on a psychiatric scale, provide an answer to

researchers facing the unavoidable issue of sample size in this field. When the factor structure underlying the instrument is clear, **Table 6** gives the estimates for the numbers of subjects required to obtain stable and accurate solution in factor analysis in various usual conditions, defined by the number of items and the number of factors present within a psychiatric scale. These estimates can then be adapted to the results set out in **figure 5** according to the desired precision of the Cronbach's alpha coefficient.

As shown by the simulation study using artificial data, a sample size of 300 is generally required, but it needs to be increased in three cases: when the number of factors within the scale is large, when EFA is chosen as the method for factor extraction and when the number of items is small. One of the most important results of this study is this last point. Indeed, it shows how the use of the *N/p* ratio rule can be deleterious, particularly for scales with a small number of items. This is consistent with the conclusions drawn by other recent simulation studies on sample size in factor analysis. These studies did not however provide a simple answer to the sample size issue because of the wide ranges of the parameter values (λ, *p, M*) studied (Guadagnoli and Velicer, 1988; Hogarty *et al.*, 2005; MacCallum *et al.*, 1999; Mundfrom *et al.*, 2005; Velicer and Fava, 1998). Another important result concerns the choice between the two different methods of factor extraction. Criticisms have been voiced in the literature against the use of the PCA. The common factor model rests on the assumption of the existence of latent variables that explain the inter-item correlations observed. It is often remarked that PCA is not fully compatible with this assumption (Costello and Osborn, 2005; Fabrigar *et al.*, 1999; Floyd and Widaman, 1995). Another criticism concerns the part of variance taken into account to estimate the loadings. In the common factor model, the shared variance of each item is partitioned from its unique variance and error variance whereas in PCA, this distinction is not made (Fabrigar *et al.*, 1999; Ford *et al.*, 1986; Widaman, 1993). Relationships between items are therefore overestimated and in the conditions occurring in psychiatry, loading estimates obtained by PCA are all the more overestimated when *p* is small and *M* large; and when *N* is large, this bias does not diminish (**figure 3**). The use of EFA is therefore recommended in this field to obtain factor solutions with a lesser bias.

Considering the difficulty in recommending a general rule for sample size calculation valid in all the fields to which psychometric procedures apply, the literature review made it possible to determine an "average" pattern of factor structure characteristic of psychiatric scales. While a review is not as accurate as a formal meta-analysis, it suggested that, in psychiatry, a particular factor structure is generally observed. Factors are correlated, salient loadings are close to 0.6 and there is a rather good factor overdetermination with an average *p/M* ratio greater than 7. The simulation of the categorical data was then performed on the basis of these characteristics and took into account different levels of floor and ceiling effects for each item. This was not the case in the previous simulation studies exploring sample size in factor analysis (Guadagnoli and Velicer, 1988; Hogarty *et al.*, 2005; MacCallum *et al.*, 1999; Mundfrom *et al.*, 2005; Velicer and Fava, 1998; Velicer *et al.*, 1982). The conditions encountered in psychiatry were therefore nearly reproduced in the artificial data. This helped to obtain results appropriate to this field that can be easily used in practice.

Concerning the limitations of the present results, two assumptions were made that could have artificially increased the strength of the artificial data as compared to real psychiatric data. One of these assumptions concerns the equality of the salient loadings. The absence of any significant influence of this on the quality of the factor solutions has been highlighted in a simulation study conducted by Velicer and Fava in 1998 (Velicer and Fava, 1998). The other assumption relates to simple structure (absence of cross-loadings and non-salient loadings set at zero). The simulation study based on real data suggests that the sample sizes recommended here could be underestimated. This is not sure. Different factor solutions were observed after resampling from the real data set. The standard deviations of loadings were thus high because of the melded fluctuations due to sampling and to the mixture of factor solutions. The interpretation of these standard deviations is not straightforward and, obviously, future studies are needed to further explore this area. At this point, we can conclude that sample sizes presented in the **table 6** represent minimal values determined from an idealized situation in which the common factor model is true. In practice, the stability of a solution obtained from real data can require a larger sample size. Of course, the present results are based on an “average” psychiatric scale and can vary according the properties of a given

instrument. However, certain elements of knowledge concerning *p* and *M* could help to obtain a clearer idea. For example, determination of the internal validity of a five-factor psychiatric scale requires at least 400 subjects if PCA is chosen as the method of factor extraction, and 450 in the case of EFA. Finally, we chose to study the influence of sample size on the precision of Cronbach's alpha coefficient, but recent developments suggest more appropriate methods for reliability estimation, such as those based on nonlinear structural equation modelling (Green *et al.*, 2009) or estimation of the greatest lower bound (Sijtsma, 2009a). However, debate is still open concerning which method should be used (Sijtsma, 2009b) and the Cronbach's alpha coefficient is by far the most used in practice.

Conclusion

The rule of the *N/p* ratio, which has already been criticised in previous studies on required sample sizes for factor analysis, is not upheld by the results of this simulation study, and researchers should refrain from using it. The validation of short scales (i.e. with a small number of items) does not warrant smaller sample size. If one's aim is to reveal the factor structure, under the hypothesis that the underlying common factor model is true, a minimum of 300 subjects is generally acceptable in the conditions encountered in the field of psychiatry. This sample size needs, however, to be larger when the expected number of factors within the scale is large. Furthermore, this study shows that, to obtain more accurate solutions, researchers should choose EFA as the method for factor extraction.

Acknowledgements

The authors wish to thank the two reviewers for their helpful suggestions especially concerning the addition of simulations based on real data.

Declaration of interest statement

The authors have no competing interests.

Appendix

The common factor model postulates that each observed variable is a linear function of one or more common factors and one unique factor. Its fundamental equation can be written:

$$y_j = \lambda_{j1}F_1 + \lambda_{j2}F_2 + ... + \lambda_{jm}F_m + ... + \lambda_{jM}F_M + \varepsilon_j$$

where y_j is the vector of the N subjects' answers to the item j (j=1 to p) and F_m the vector of the N subjects' non-observable scores on the common factor m (m=1 to M). Each item j loads on each common factor m with the factor loading λ_{jm}. The unique factor ε_j, for each item j is independent ($\perp$) from all the F_m and from the other $\varepsilon_{(j'\neq j)}$ (Brown, 2006). In our simulation model, two hypotheses are set out. The first is the existence of a simple structure, i.e. each item loads on a single factor and all the non-salient loadings are equal to zero. The second is that all salient loadings (λ) are equal. Therefore, if the p_1 first items load only onto the first factor F_1, the p_2 following items load onto F_2, …, the p_m following onto F_m, …, and the p_m last items onto F_M, ($\sum_{m=1}^{M} p_m = p$), then all the answers to a p item scale can be modelled as:

$$\begin{cases} \forall j \in [1, p_1], \quad y_j = \lambda' F_1 + \varepsilon_j \\ \forall j \in [(p_1 + 1), p_2], \quad y_j = \lambda' F_2 + \varepsilon_j \\ \vdots \\ \forall j \in [(p_{(m-1)} + 1), p_m], \quad y_j = \lambda' F_m + \varepsilon_j \\ \vdots \\ \forall j \in [(p_{(M-1)} + 1), p_M], \quad y_j = \lambda' F_M + \varepsilon_j \end{cases}$$

where $\forall j \in [1, p]$, $\varepsilon_j \sim \mathcal{N}(0,1)$ and $\varepsilon_j \perp \varepsilon_{(j'\neq j)}$

and $\forall m \in [1, M]$, $F_m \sim \mathcal{N}(0,1)$ and $F_m \perp \varepsilon_j$

In this model, the coefficient λ' is not directly equal to the salient loadings. Indeed, in order to preserve the variances of the y_j equal to unity, standardization is required using the

factor $\frac{1}{\sqrt{1+\lambda'^2}}$. Individual data can therefore be simulated in a matrix where each row represents the answers of one individual to all p items in the scale and each column represents the answers of the N individuals to one item. If i represents subjects (i = 1 to N), the answer of the subject i to the item j is:

$$\forall i \in [1, N],\ \forall m \in [1, M],\ \forall j \in [(p_{(m-1)} + 1), p_m],\ y_{ij} = \frac{\lambda' F_{mi} + \varepsilon_{ij}}{\sqrt{1+\lambda'^2}}$$

To introduce correlations between factors in this simulation model, each factor is modelled using a term specific to each factor ($f_m \sim \mathcal{N}(0,1)$) and a term common to all factors ($C \sim \mathcal{N}(0,1)$):

$$F_m = a_m f_m + b_m C$$

Thus, the proportions of each of these terms, a_m and b_m, make it possible to control for the factor inter-correlation levels with solely the constraint that $a_m^2 + b_m^2 = 1$ to preserve the variances of factors equal to unity. A last stage is necessary to obtain a non-symmetrical distribution of categorical data, as for data encountered in a real internal validity study on a psychiatric scale, for example, answers to a four-point Likert scale. The conversion of the y_{ij} into integral numbers from one to four is performed using three breakpoints in their distribution $\mathcal{N}(0,1)$. For each item j, these three breakpoints are ($-1+\delta_j$), ($0+\delta_j$), and ($1+\delta_j$) where δ_j is drawn from a uniform distribution between [-0.5, 0.5] to introduce asymmetry and thus simulate floor and ceiling effects. The data simulation was performed using R software 2.6.2.; vectors ε_j, f_m and C were generated using the function `rnorm` and δ_j using `runif`.

References

* References marked with an asterisk were included in the literature review

*Adachi N., Onuma T., Nishiwaki S., Murauchi S., Akanuma N., Ishida S., Takei N. (2000). Inter-ictal and post-ictal psychoses in frontal lobe epilepsy: a retrospective comparison with psychoses in temporal lobe epilepsy. *Seizure*; **9**, 328-35, DOI: 10.1053/seiz.2000.0413S1059-1311(00)90413-8 [pii]

Aleamoni L. M. (1973). Effects of size of sample on eigenvalues, observed communalities, and factor loadings. *J Appl Psychol*; **58**, 266-9, DOI: 10.1037/h0035429

*Basker M., Moses P. D., Russell S., Russell P. S. (2007). The psychometric properties of Beck Depression Inventory for adolescent depression in a primary-care paediatric setting in India. *Child Adolesc Psychiatry Ment Health*; **1**, 8, DOI: 1753-2000-1-8 [pii]10.1186/1753-2000-1-8

*Beck A. T. (1991). Relationship between the Beck Anxiety Inventory and the Hamilton Anxiety Rating Scale with anxious outpatients. *J Anxiety Disord*; **5**, 213-23, DOI: 10.1016/0887-6185(91)90002-B

Beck A. T., Ward C. H., Mendelson M., Mock J., Erbaugh J. (1961). An inventory for measuring depression. *Arch Gen Psychiatry*; **4**, 561-71.

*Bell M. D., Lysaker P. H., Beam-Goulet J. L., Milstein R. M., Lindenmayer J. P. (1994). Five-component model of schizophrenia: assessing the factorial invariance of the positive and negative syndrome scale. *Psychiatry Res*; **52**, 295-303, DOI: 0165-1781(94)90075-2 [pii]

*Bonicatto S., Dew A. M., Soria J. J. (1998). Analysis of the psychometric properties of the Spanish version of the Beck Depression Inventory in Argentina. *Psychiatry Res*; **79**, 277-85, DOI: S0165-1781(98)00047-X [pii]

*Bonilla J., Bernal G., Santos A., Santos D. (2004). A revised Spanish version of the Beck Depression Inventory: psychometric properties with a Puerto Rican sample of college students. *J Clin Psychol*; **60**, 119-30, DOI:10.1002/jclp.10195

Brown T. A. (2006). Confirmatory factor analysis for applied research. New York: The Guilford Press.

*Castro-Costa E., Uchoa E., Firmo J. O., Lima-Costa M. F., Prince M. (2008). Association of cognitive impairment, activity limitation with latent traits in the GHQ-12 in the older elderly. The Bambui Health and Aging Study (BHAS). *Aging Clin Exp Res*; **20**, 562-8, DOI: 5323 [pii]

Cattell R. B. (1978). The scientific use of factor analysis in behavioral and life sciences. New York Plenum press.

*Chapman L. K., Williams S. R., Mast B. T., Woodruff-Borden J. (2009). A confirmatory factor analysis of the Beck Anxiety Inventory in African American and European American young adults. *J Anxiety Disord*; **23**, 387-92, DOI: S0887-6185(08)00218-1 [pii]10.1016/j.janxdis.2008.12.003

Comrey A. L. (1978). Common Methodological Problems in Factor Analytic Studies. *J Consult Clin Psych*; **46**, 648-59, DOI: 10.1037/0022-006X.46.4.648

Comrey A. L., Lee H. B. (1992). A first course in factor analysis. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Costello A. B., Osborn J. W. (2005). Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis. *Practical Assessment, Research & Evaluation*; **10**.

*Dagnan D., Jahoda A., McDowell K., Masson J., Banks P., Hare D. (2008). The psychometric properties of the Hospital Anxiety and Depressions Scale adapted for use with people with intellectual disabilities. *J Intell Disabil Res*; **52**, 942-9, DOI: JIR1053 [pii]10.1111/j.1365-2788.2008.01053.x

*Dawkins N., Cloherty M. E., Gracey F., Evans J. J. (2006). The factor structure of the Hospital Anxiety and Depression Scale in acquired brain injury. *Brain Injury*; **20**, 1235-9, DOI: V041611313646066 [pii]10.1080/02699050601076414

Everitt B. S. (1975). Multivariate analysis: the need for data, and other problems. *Brit J Psychiat*; **126**, 237-40, DOI: 10.1192/bjp.126.3.237

Fabrigar L. R., Wegener D. T., MacCallum R. C., Strahan E. J. (1999). Evaluating the Use of Exploratory Factor Analysis in Psychological Research. *Psychol Methods*; **4**, 272-99, DOI: 10.1037/1082-989X.4.3.272

Fan X., Thompson B. (2001). Confidence Intervals about Score Reliability Coefficients, Please: An EPM Guideline Editorial. *Educ Psychol Meas*; **61**, 517-31, DOI: 10.1177/0013164401614001

*Farrell G. A. (1998). The mental health of hospital nurses in Tasmania as measured by the 12-item General Health Questionnaire. *J Adv Nurs*; **28**, 707-12, DOI: 10.1046/j.1365-2648.1998.00735

Fedorowicz V. J., Falissard B., Foulon C., Dardennes R., Divac S. M., Guelfi J. D., Rouillon F. (2007). Factors associated with suicidal behaviors in a large French sample of inpatients with eating disorders. *Int J Eat Disorder*; **40**, 589-95, DOI: 10.1002/eat.20415

Feldt L. S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika*; **30**, 357-70, DOI: 10.1007/BF02289499

Floyd F. J., Widaman K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychol Assessment*; **7**, 286-99, DOI: 10.1037/1040-3590.7.3.286

Ford J. K., MacCallum R. C., Tait M. (1986). The application of exploratory factor analysis in applied psychology: a critical review and analysis. *Personnel Psychology*; **39**, 291-314, DOI: 10.1111/j.1744-6570.1986.tb00583.x

*Fresan A., De la Fuente-Sandoval C., Loyzaga C., Garcia-Anaya M., Meyenberg N., Nicolini H., Apiquian R. (2005). A forced five-dimensional factor analysis and concurrent validity of the Positive and Negative Syndrome Scale in Mexican schizophrenic patients. *Schizophr Res*; **72**, 123-9, DOI:10.1016/j.schres.2004.03.021

*Friedman S., Samuelian J. C., Lancrenon S., Even C., Chiarelli P. (2001). Three-dimensional structure of the Hospital Anxiety and Depression Scale in a large French primary care population suffering from major depression. *Psychiatry Res*; **104**, 247-57, DOI: S0165-1781(01)00309-2 [pii]

*Gabryelewicz T., Styczynska M., Pfeffer A., Wasiak B., Barczak A., Luczywek E., Androsiuk W., Barcikowska M. (2004). Prevalence of major and minor depression in elderly persons with mild cognitive impairment--MADRS factor analysis. *Int J Geriatr Psych*; **19**, 1168-72, DOI: 10.1002/gps.1235

*Galinowski A., Lehert P. (1995). Structural validity of MADRS during antidepressant treatment. *Int Clin Psychopharm*; **10**, 157-61, DOI: 10.1097/00004850-199510030-00004

*Gorenstein C., Andrade L., Vieira Filho A. H., Tung T. C., Artes R. (1999). Psychometric properties of the Portuguese version of the Beck Depression Inventory on Brazilian college students. *J Clin Psychol*; **55**, 553-62, DOI: 10.1002/(SICI)1097-4679(199905)55:5<553::AID-JCLP3>3.0.CO;2-D [pii]

Gorsuch R. L. (1983). Factor Analysis London: Lawrence Erlbaum Associates.

Green S. A., Yang Y. (2009). Reliability of summed item scores using structural equation modeling: an alternative to coefficient alpha. *Psychometrika*; **74**, 155-67, DOI: 10.1007/S11336-008-9099-3

*Grunebaum M. F., Keilp J., Li S., Ellis S. P., Burke A. K., Oquendo M. A., Mann J. J. (2005). Symptom components of standard depression scales and past suicidal behavior. *J Affect Disorders*; **87**, 73-82, DOI: S0165-0327(05)00073-X [pii] 10.1016/j.jad.2005.03.002

*Guadagnoli E., Velicer W. F. (1988). Relation of sample size to the stability of component patterns. *Psychol Bull*; **103**, 265-75, DOI: 10.1037/0033-2909.103.2.265

Hamilton M. (1960). A rating scale for depression. *J Neurol Neurosurg Psychiatry*; **23**, 56-62,

*Hankins M. (2008). The factor structure of the twelve item General Health Questionnaire (GHQ-12): the result of negative phrasing? *Clinical Practice and Epidemiology in Mental Health* **4**, 10, DOI: 1745-0179-4-10 [pii]10.1186/1745-0179-4-10

*Harvey P. D., Davidson M., White L., Keefe R. S., Hirschowitz J., Mohs R. C., Davis K. L. (1996). Empirical evaluation of the factorial structure of clinical symptoms in schizophrenia: effects of typical neuroleptics on the brief psychiatric rating scale. *Biol Psychiatry*; **40**, 755-60, DOI: 0006-3223(95)00486-6 [pii]10.1016/0006-3223(95)00486-6

*Helm H. W., Jr., Boward M. D. (2003). Factor structure of the Beck Depression Inventory in a university sample. *Psychol Rep*; **92**, 53-61.

Hogarty K. Y., Hines C. V., Kromrey J. D., Ferron J. M., Mumford K. R. (2005). The Quality of Factor Solutions in Exploratory Factor Analysis: The Influence of Sample Size, Communality, and Overdetermination *Educ Psychol Meas*; **65**, 202-6, DOI: 10.1177/0013164404267287

*Honey G. D., Sharma T., Suckling J., Giampietro V., Soni W., Williams S. C., Bullmore E. T. (2003). The functional neuroanatomy of schizophrenic subsyndromes. *Psychol Med*; **33**, 1007-18, DOI:10.1017/S0033291703007864

*Hu Y., Stewart-Brown S., Twigg L., Weich S. (2007). Can the 12-item General Health Questionnaire be used to measure positive mental health? *Psychol Med*; **37**, 1005-13, DOI:10.1017/S0033291707009993

*Iwata N., Mishima N., Okabe K., Kobayashi N., Hashiguchi E., Egashira K. (2000). Psychometric properties of the State-Trait Anxiety Inventory among Japanese clinical outpatients. *J Clin Psychol*; **56**, 793-806, DOI: 10.1002/(SICI)1097-4679(200006)56:6<793::AID-JCLP8>3.0.CO;2-4 [pii]

*Iwata N., Mishima N., Shimizu T., Mizoue T., Fukuhara M., Hidano T., Spielberger C. D. (1998). Positive and negative affect in the factor structure of the State-Trait Anxiety Inventory for Japanese workers. *Psychol Rep*; **82**, 651-6, DOI: 10.2466/PR0.82.2.651-656

Jackson J. E. (1991). A user's guide to principal components. New York: John Wiley & sons.

*Jo S. A., Park M. H., Jo I., Ryu S. H., Han C. (2007). Usefulness of Beck Depression Inventory (BDI) in the Korean elderly population. *Int J Geriatr Psych*; **22**, 218-23, DOI: 10.1002/gps.1664

*Kabacoff R. I., Segal D. L., Hersen M., Van Hasselt V. B. (1997). Psychometric properties and diagnostic utility of the Beck Anxiety Inventory and the State-Trait Anxiety Inventory with older adult psychiatric outpatients. *J Anxiety Disord*; **11**, 33-47, DOI: S0887618596000333 [pii]

*Kay S. R., Sevy S. (1990). Pyramidical model of schizophrenia. *Schizophr Bull*; **16**, 537-45, DOI:10.1016/0165-1781(94)90075-2

*Kilic C., Rezaki M., Rezaki B., Kaplan I., Ozgen G., Sagduyu A., Ozturk M. O. (1997). General Health Questionnaire (GHQ12 &GHQ28) : psychometric properties and factor structure of the scales in a Turkish primary care sample. *Soc Psychiatry Psychiatr Epidemiol*; **32**, 327-31, DOI: 10.1007/BF00805437

*Killgore W. D. (1999). Empirically derived factor indices for the Beck Depression Inventory. *Psychol Rep*; **84**, 1005-13.

*Lachar D., Bailley S. E., Rhoades H. M., Espadas A., Aponte M., Cowan K. A., Gummattira P., Kopecky C. R., Wassef A. (2001). New subscales for an anchored version of the Brief Psychiatric Rating Scale: construction, reliability, and validity in acute psychiatric admissions. *Psychol Assessment*; **13**, 384-95, DOI: 10.1037/1040-3590.13.3.384

*Lancon C., Reine G., Llorca P. M., Auquier P. (1999). Validity and reliability of the French-language version of the Positive and Negative Syndrome Scale (PANSS). *Acta Psychiatr Scand*; **100**, 237-43, DOI: 10.1111/j.1600-0447.1999.tb10851.x

*Lee K. H., Harris A. W., Loughland C. M., Williams L. M. (2003). The five symptom dimensions and depression in schizophrenia. *Psychopathology*; **36**, 226-33, DOI: 10.1159/000073447 PSP2003036005226 [pii]

*Lindenmayer J. P., Czobor P., Volavka J., Lieberman J. A., Citrome L., Sheitman B., McEvoy J. P., Cooper T. B., Chakos M. (2004). Effects of atypical antipsychotics on the syndromal profile in treatment-resistant schizophrenia. *J Clin Psychiat*; **65**, 551-6.

Loo R. (1983). Caveat on Sample Sizes in Factor Analysis. *Percept Mot Skills*; **56**, 371-4.

*Lopez-Castedo A., Fernandez L. (2005). Psychometric properties of the Spanish version of the 12-item General Health Questionnaire in adolescents. *Percept Mot Skills* **100**, 676-80.

*Loza B., Kucharska-Pietura K., Kopacz G., Debowska G. (2003). Factor structure of paranoid schizophrenia: a prospective study. *Psychopathology*; **36**, 132-41, DOI: 10.1159/000071258 PSP2003036003132 [pii]

*Lykouras L., Oulis P., Psarros K., Daskalopoulou E., Botsis A., Christodoulou G. N., Stefanis C. (2000). Five-factor model of schizophrenic psychopathology: how valid is it? *Eur Arch Psy Clin N*; **250**, 93-100, DOI: 10.1007/s004060070041

MacCallum R. C., Widaman K. F., Zhang S., Hong S. (1999). Sample Size in Factor Analysis. *Psychol Methods*; **4**, 84-99, DOI: 10.1037/1082-989X.4.1.84

Mundfrom D. J., Shaw D. G., Ke T. L. (2005). Minimum Sample Size Recommendations for Conducting Factor Analyses *International Journal of Testing*; **5**, 159-68, DOI: 10.1207/s15327574ijt0502_4

*Munoz D. J., Chen E., Fischer S., Roehrig M., Sanchez-Johnson L., Alverdy J., Dymek-Valentine M., le Grange D. (2007). Considerations for the use of the Beck Depression

Inventory in the assessment of weight-loss surgery seeking patients. *Obes Surg*; **17**, 1097-101, DOI: 10.1007/s11695-007-9185-0

Nunnaly J. C. (1978). Psychometric theory. New York McGraw-Hill.

*Olden M., Rosenfeld B., Pessin H., Breitbart W. (2009). Measuring depression at the end of life: is the Hamilton Depression Rating Scale a valid instrument? *Assessment*; **16**, 43-54, DOI: 1073191108320415 [pii] 10.1177/1073191108320415

*Pallant J. F., Bailey C. M. (2005). Assessment of the structure of the Hospital Anxiety and Depression Scale in musculoskeletal patients. *Health Qual Life Out*; **3**, 82, DOI: 1477-7525-3-82 [pii] 10.1186/1477-7525-3-82

*Parker R. D., Flint E. P., Bosworth H. B., Pieper C. F., Steffens D. C. (2003). A three-factor analytic model of the MADRS in geriatric depression. *Int J Geriatr Psych*; **18**, 73-7, DOI: 10.1002/gps.776

Peterson R. A. (1994). A Meta-Analysis of Cronbach's Coefficient Alpha. *J Consum Res*; **21**, 381-91, DOI: 10.1086/209405

*Powell R. (2003). Psychometric properties of the Beck Depression Inventory and the Zung Self Rating Depression Scale in adults with mental retardation. *Ment Retard*; **41**, 88-95, DOI: 10.1352/0047-6765(2003)041<0088:PPOTBD>2.0.CO;2

R Development Core Team (2008). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for statistical Computing.

Revelle W. (2008). R Documentation: Procedures for Personality, Psychometric, and Psychological Research. Help pages for package 'psych' version 1.0-58: Principal Axis Factor Analysis: CRAN.

*Salamero M., Marcos T., Gutierrez F., Rebull E. (1994). Factorial study of the BDI in pregnant women. *Psychol Med*; **24**, 1031-5.

*Salokangas R. K., Honkonen T., Stengard E., Koivisto A. M. (2002). Symptom dimensions and their association with outcome and treatment setting in long-term schizophrenia. Results of the DSP project. *Nord J Psychiat*; **56**, 319-27, DOI: 10.1080/080394802760322079

*Sanchez-Lopez Mdel P., Dresch V. (2008). The 12-Item General Health Questionnaire (GHQ-12): reliability, external validity and factor structure in the Spanish population. *Psicothema*; **20**, 839-43.

*Serretti A., Jori M. C., Casadei G., Ravizza L., Smeraldi E., Akiskal H. (1999). Delineating psychopathologic clusters within dysthymia: a study of 512 out-patients without major depression. *J Affect Disorders*; **56**, 17-25, DOI: S0165-0327(99)00056-7 [pii]

*Shek D. T. (1990). Reliability and factorial structure of the Chinese version of the Beck Depression Inventory. *J Clin Psychol*; **46**, 35-43, DOI: 10.1002/1097-4679(199001)46:1<35::AID-JCLP2270460106>3.0.CO;2-W

Sijtsma K. (2009a). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*; **74**, 107-20, DOI: 10.1007/S11336-008-9101-0

Sijtsma K. (2009b). Reliability beyond theory and into practice. *Psychometrika*; **74**, 169-73, DOI: 10.1007/S11336-008-9103-Y

*Smith A. B., Selby P. J., Velikova G., Stark D., Wright E. P., Gould A., Cull A. (2002). Factor analysis of the Hospital Anxiety and Depression Scale from a large cancer population. *Psychol Psychother*; **75**, 165-76.

*Steer R. A., Beck A. T., Brown G. (1989). Sex differences on the revised Beck Depression Inventory for outpatients with affective disorders. *J Pers Assess*; **53**, 693-702, DOI: 10.1207/s15327752jpa5304_6

*Steer R. A., Kumar G., Ranieri W. F., Beck A. T. (1995). Use of the Beck Anxiety Inventory with adolescent psychiatric outpatients. *Psychol Rep*; **76**, 459-65.

*Steer R. A., Rissmiller D. J., Ranieri W. F., Beck A. T. (1993). Structure of the computer-assisted Beck Anxiety Inventory with psychiatric inpatients. *J Pers Assess*; **60**, 532-42, DOI: 10.1207/s15327752jpa6003_10

*Uslu R. I., Kapci E. G., Oncu B., Ugurlu M., Turkcapar H. (2008). Psychometric properties and cut-off scores of the Beck Depression Inventory-II in Turkish adolescents. *J Clin Psychol Med S* **15**, 225-33, DOI:10.1007/s10880-008-9122-y

Velicer W. F., Fava J. L. (1998). Effects of variable and subject sampling on factor pattern recovery *Psychol Methods*; **3**, 231-51, doi: 10.1037/1082-989X.3.2.231

Velicer W. F., Peacock A. C., Jackson D. N. (1982). A Comparison of Component and Factor Patterns: A Monte Carlo Approach. *Multivar Behav Res*; **17**, 371-88, DOI: 10.1207/s15327906mbr1703_5

*Ventura J., Nuechterlein K. H., Subotnik K. L., Gutkind D., Gilbert E. A. (2000). Symptom dimensions in recent-onset schizophrenia and mania: a principal components analysis of the 24-item Brief Psychiatric Rating Scale. *Psychiatry Res*; **97**, 129-35, DOI: S0165178100002286 [pii]

*Villalta-Gil V., Vilaplana M., Ochoa S., Dolz M., Usall J., Haro J. M., Almenara J., Gonzalez J. L., Lagares C. (2006). Four symptom dimensions in outpatients with schizophrenia. *Compr Psychiatry*; **47**, 384-8, DOI: S0010-440X(06)00019-8 [pii] 10.1016/j.comppsych.2006.01.005

*Wang Y. P., Andrade L. H., Gorenstein C. (2005). Validation of the Beck Depression Inventory for a Portuguese-speaking Chinese community in Brazil. *Braz J Med Biol Res*; **38**, 399-408, DOI: S0100-879X2005000300011 [pii] /S0100-879X2005000300011

*Werneke U., Goldberg D. P., Yalcin I., Ustun B. T. (2000). The stability of the factor structure of the General Health Questionnaire. *Psychol Med*; **30**, 823-9, DOI: 10.1017/S0033291799002287

Widaman K. F. (1993). Common Factor Analysis Versus Principal Component Analysis: Differential Bias in Representing Model Parameters? *Multivar Behav Res*; **28**, 263-311, DOI: 10.1207/s15327906mbr2803_1

*Woolrich R. A., Kennedy P., Tasiemski T. (2006). A preliminary psychometric evaluation of the Hospital Anxiety and Depression Scale (HADS) in 963 people living with a spinal cord injury. *Psychol Health Med*; **11**, 80-90, DOI: J07M627G23552610 [pii] 10.1080/13548500500294211

Tables

Table 1: References included and numbers of references extracting the same number of factors for each scale

Scale	References	Total	Number of factors					
			2	**3**	**4**	**5**	**6**	**7**
PANSS	(Bell *et al.*, 1994; Fresan *et al.*, 2005; Honey *et al.*, 2003; Kay *et al.*, 1990; Lancon *et al.*, 1999; Lee *et al.*, 2003; Lindenmayer *et al.*, 2004; Loza *et al.*, 2003; Lykouras *et al.*, 2000; Salokangas *et al.*, 2002; Villalta-Gil *et al.*, 2006)	11	-	-	1	8	-	2
BPRS	(Adachi *et al.*, 2000; Harvey *et al.*, 1996; Lachar *et al.*, 2001; Ventura *et al.*, 2000)	4	-	-	2	1	1	-
BAI	(Beck, 1991; Chapman *et al.*, 2009; Kabacoff *et al.*, 1997; Steer *et al.*, 1995; Steer *et al.*, 1993)	5	4	-	1	-	-	-
STAI	(Iwata *et al.*, 2000; Iwata *et al.*, 1998; Kabacoff *et al.*, 1997)	3	2	1	-	-	-	-
HAMA	(Beck, 1991; Serretti *et al.*, 1999)	2	2	-	-	-	-	-
HAMD	(Grunebaum *et al.*, 2005; Olden *et al.*, 2009)	2	-	-	1	1	-	-
MADRS	(Gabryelewicz *et al.*, 2004; Galinowski *et al.*, 1995; Lee *et al.*, 2003; Parker *et al.*, 2003; Serretti *et al.*, 1999)	5	3	2	-	-	-	-
BDI	(Basker *et al.*, 2007; Bonicatto *et al.*, 1998; Bonilla *et al.*, 2004; Gorenstein *et al.*, 1999; Grunebaum *et al.*, 2005; Helm *et al.*, 2003; Jo *et al.*, 2007; Killgore, 1999; Munoz *et al.*, 2007; Powell, 2003; Salamero *et al.*, 1994; Shek, 1990; Steer *et al.*, 1989; Uslu *et al.*, 2008; Wang *et al.*, 2005)	15	9	2	3	-	-	1
HADS	(Dagnan *et al.*, 2008; Dawkins *et al.*, 2006; Friedman *et al.*, 2001; Pallant *et al.*, 2005; Smith *et al.*, 2002; Woolrich *et al.*, 2006)	6	4	2	-	-	-	-
GHQ	(Castro-Costa *et al.*, 2008; Farrell, 1998; Hankins, 2008; Hu *et al.*, 2007; Kilic *et al.*, 1997; Lopez-Castedo *et al.*, 2005; Sanchez-Lopez Mdel and Dresch, 2008; Werneke *et al.*, 2000)	8	5	3	-	-	-	-
Total		61	29	10	8	10	1	3

Table 2: Percentage of variance accounted for by each factor and numbers of references used to estimate the means for each scale

Scale		Factors						
		F_1	F_2	F_3	F_4	F_5	F_6	F_7
PANSS	Mean	25.8	12.8	8.8	6.8	5.8	3.6	3.6
	(Minimum – Maximum)	(14.5 – 41.2)	(8.7 – 18.6)	(6.1 – 13.4)	(3.9 – 11.1)	(3.6 – 9.3)	(3.6 – 3.7)	(3.6 – 3.7)
	Number of references	11	11	11	11	10	2	2
BPRS	Mean	18.9	14.0	10.4	8.3	6.9	6.7	
	(Minimum – Maximum)	(12.8 – 23.3)	(9.3 – 17.2)	(8.7 – 11.7)	(6.7 – 10.0)	(6.1 – 7.8)	(. - .)	
	Number of references	4	4	4	4	2	1	
BAI	Mean	37.6	7.2	6.2	5.2			
	(Minimum – Maximum)	(36.3 – 39.5)	(4.4 – 7.7)	(. - .)	(. - .)			
	Number of references	5	5	1	1			
STAI	Mean	32.2	9.6	6.0				
	(Minimum – Maximum)	(29.8 – 34.3)	(7.4 – 11)	(. - .)				
	Number of references	3	3	1				
HAMA	Mean	26.9	8.2					
	(Minimum – Maximum)	(20.4 – 33.5)	(6.4 – 10)					
	Number of references	2	2					
HAMD	Mean	12.8	11.3	10.7	8.5	9.4		
	(Minimum – Maximum)	(12.6 – 13.0)	(11.2 - 11.4)	(10.4 – 11.0)	(7.3 – 9.8)	(. - .)		
	Number of references	2	2	2	2	1		
MADRS	Mean	33.7	15.6	10.6				
	(Minimum – Maximum)	(25.1 – 41.1)	(10.4 – 26.9)	(10.2 – 1.0)				
	Number of references	5	5	2				
BDI	Mean	29.4	8.8	6.8	5.4	6.0	5.5	4.9
	(Minimum – Maximum)	(22.9 – 34.5)	(5.9 – 25.1)	(5.0 – 6.1)	(4.9 – 6.1)	(. - .)	(. - .)	(. - .)
	Number of references	15	15	6	4	1	1	1
HADS	Mean	34.3	13.1	8.4				
	(Minimum – Maximum)	(23.6 – 41.4)	(11.4 – 16.4)	(8.1 – 8.6)				
	Number of references	6	6	2				
GHQ	Mean	39.6	13.0	9.2				
	(Minimum – Maximum)	(30.3 – 50.9)	(8.5 – 25.9)	(8.6 – 9.8)				

	Number of references	8	8	3				
Total	Mean	30.4	11.3	8.7	6.9	6.3	4.9	4.0
	(Minimum – Maximum)	(12.6 - 50.8)	(4.4 – 26.9)	(5.0 – 13.4)	(3.9 – 11.1)	(3.6 – 9.4)	(3.6 – 6.7)	(3.6 – 4.9)
	Number of references	61	61	32	22	14	4	3

Table 3: Mean of the percentages of items per factor (IQR: Interquartile Range)

Number of factors in the scale		Factors						
		F_1	F_2	F_3	F_4	F_5	F_6	F_7
2	Mean	55.7	39.0					
	IQR	[50.0 - 59.2]	[33.3 - 42.9]					
3	Mean	43.2	34.8	20.2				
	IQR	[40.4 - 49.4]	[27.8 - 41.3]	[16.7 - 24.1]				
4	Mean	29.1	26.6	19.3	20.3			
	IQR	[27.0 - 33.0]	[22.9 - 32.2]	[16.3 - 20.6]	[14.8 - 23.7]			
5	Mean	22.6	20.1	15.7	16.7	15.1		
	IQR	[20.0 - 25.8]	[16.7 - 22.5]	[13.3 - 19.2]	[16.7 - 19.7]	[12.7 - 16.7]		
6	Mean	16.7	22.2	16.7	16.7	11.1	16.7	
	IQR	[. - .]	[. - .]	[. - .]	[. - .]	[. - .]	[. - .]	
7	Mean	24.1	22.4	12.5	15.2	12.1	6.5	7.1
	IQR	[19.5 - 26.7]	[20.2 - 25.2]	[10.5 - 15.5]	[13.3 - 16.9]	[9.8 - 13.3]	[5.0 - 8.1]	[4.0 - 9.0]

Table 4: Values of the three criteria after PCA in the case of a three-factor scale (σ_λ: standard deviation of the salient loadings obtained after rotation over the 10 000 simulations, $R_{\%}$: percentage of simulations in which all the items in the scale load on the right factor, $W_{\%}$: mean of percentages of items loading on the wrong factor in the scale after rotation over the 10 000 simulations, **-** : < 5.10^{-2})

Sample size		Number of items							
		10	**15**	**20**	**25**	**30**	**35**	**40**	**45**
50	σ_λ	0.182	0.161	0.144	0.136	0.130	0.127	0.124	0.123
	$R_{\%}$	48.4	48.5	51.1	50.7	51.3	51.5	50.6	49.5
	$W_{\%}$	9.3	6.5	4.4	3.4	2.7	2.3	2.1	1.9
100	σ_λ	0.111	0.097	0.092	0.088	0.087	0.086	0.084	0.083
	$R_{\%}$	88.8	92.6	94.6	95.9	96.4	97.1	96.9	97.1
	$W_{\%}$	1.5	0.6	0.3	0.2	0.1	0.1	0.1	0.1
150	σ_λ	0.081	0.075	0.072	0.071	0.069	0.069	0.068	0.068
	$R_{\%}$	97.8	99.4	99.5	99.7	99.8	99.7	99.8	99.8
	$W_{\%}$	0.3	0.1	-	-	-	-	-	-
200	σ_λ	0.067	0.063	0.062	0.061	0.060	0.059	0.059	0.059
	$R_{\%}$	99.5	99.9	99.8	99.8	99.9	99.9	99.9	99.8
	$W_{\%}$	0.1	-	-	-	-	-	-	0.1
300	σ_λ	0.052	0.050	0.051	0.049	0.049	0.049	0.048	0.048
	$R_{\%}$	99.9	100.0	99.8	99.9	99.9	99.8	100.0	99.8
	$W_{\%}$	0.1	-	0.1	-	-	0.1	-	-
500	σ_λ	0.039	0.039	0.039	0.039	0.038	0.037	0.037	0.038
	$R_{\%}$	99.9	100.0	99.9	99.9	99.9	99.9	99.9	99.9
	$W_{\%}$	-	-	-	-	-	-	-	-
1000	σ_λ	0.029	0.027	0.027	0.027	0.027	0.027	0.026	0.026
	$R_{\%}$	99.9	100.0	100.0	99.9	100.0	99.9	100.0	100.0
	$W_{\%}$	0.1	-	-	-	-	-	-	-

Table 5: Values of the three criteria after EFA in the case of a three-factor scale (σ_λ: standard deviation of the salient loadings obtained after rotation over the 10 000 simulations, $R_{\%}$: percentage of simulations in which all the items of the scale load on the right factor, $W_{\%}$: mean of percentages of items loading on the wrong factor in the scale after rotation over the 10 000 simulations, **-** : < 5.10^{-2})

Sample size		Number of items							
		10	**15**	**20**	**25**	**30**	**35**	**40**	**45**
50	σ_λ	0.226	0.187	0.164	0.153	0.144	0.138	0.134	0.131
	$R_{\%}$	31.1	34.9	40.9	43.7	45.3	47.3	47.1	46.7
	$W_{\%}$	14.9	10.5	6.6	4.8	3.5	2.8	2.4	2.1
100	σ_λ	0.159	0.125	0.109	0.101	0.096	0.093	0.091	0.089
	$R_{\%}$	70.7	86.3	92.7	95.0	95.8	96.5	96.6	96.9
	$W_{\%}$	4.4	1.3	0.4	0.2	0.1	0.1	0.1	0.1
150	σ_λ	0.128	0.098	0.086	0.080	0.077	0.075	0.073	0.072
	$R_{\%}$	89.8	98.7	99.4	99.7	99.6	99.8	99.7	99.8
	$W_{\%}$	1.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
200	σ_λ	0.109	0.082	0.073	0.069	0.066	0.064	0.063	0.062
	$R_{\%}$	96.4	99.8	99.8	99.9	99.9	99.9	99.9	99.9
	$W_{\%}$	0.4	-	-	-	-	-	-	-
300	σ_λ	0.086	0.065	0.059	0.056	0.054	0.052	0.051	0.051
	$R_{\%}$	99.6	99.9	99.9	99.9	99.9	99.9	100.0	99.9
	$W_{\%}$	-	-	-	-	-	-	-	-
500	σ_λ	0.063	0.050	0.045	0.043	0.041	0.040	0.039	0.040
	$R_{\%}$	99.9	99.9	100.0	99.9	100.0	99.9	100.0	99.8
	$W_{\%}$	-	-	-	-	-	-	-	0.1
1000	σ_λ	0.043	0.034	0.032	0.030	0.029	0.028	0.028	0.027
	$R_{\%}$	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0
	$W_{\%}$	-	-	-	-	-	-	-	-

Table 6: Sample size required to meet the three criteria thresholds for quality of factor solutions (- : >1000)

Method of factor extraction	Number of factors	Number of items							
		10	**15**	**20**	**25**	**30**	**35**	**40**	**45**
PCA	**2**	300	300	300	300	300	300	250	250
	3	350	350	350	300	300	300	300	300
	4	400	400	350	350	350	350	350	350
EFA	**2**	500	400	350	300	300	300	300	300
	3	800	500	450	400	350	350	350	350
	4	-	-	600	500	450	400	400	400

Figures

Figure 1: Box-plot of the percentage of variance accounted for by each factor, according the factor rank in the scale, in all the references

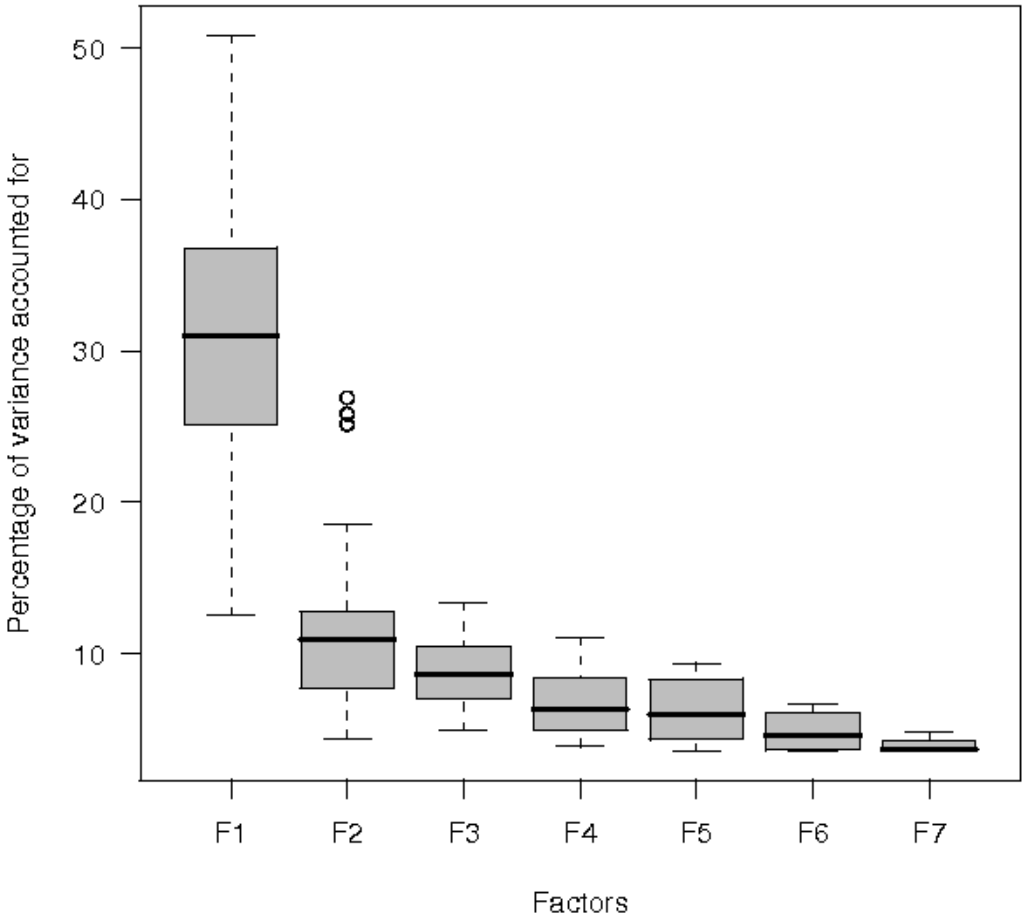


Figure 2: Path diagram for the three-factor simulation model with 10 items

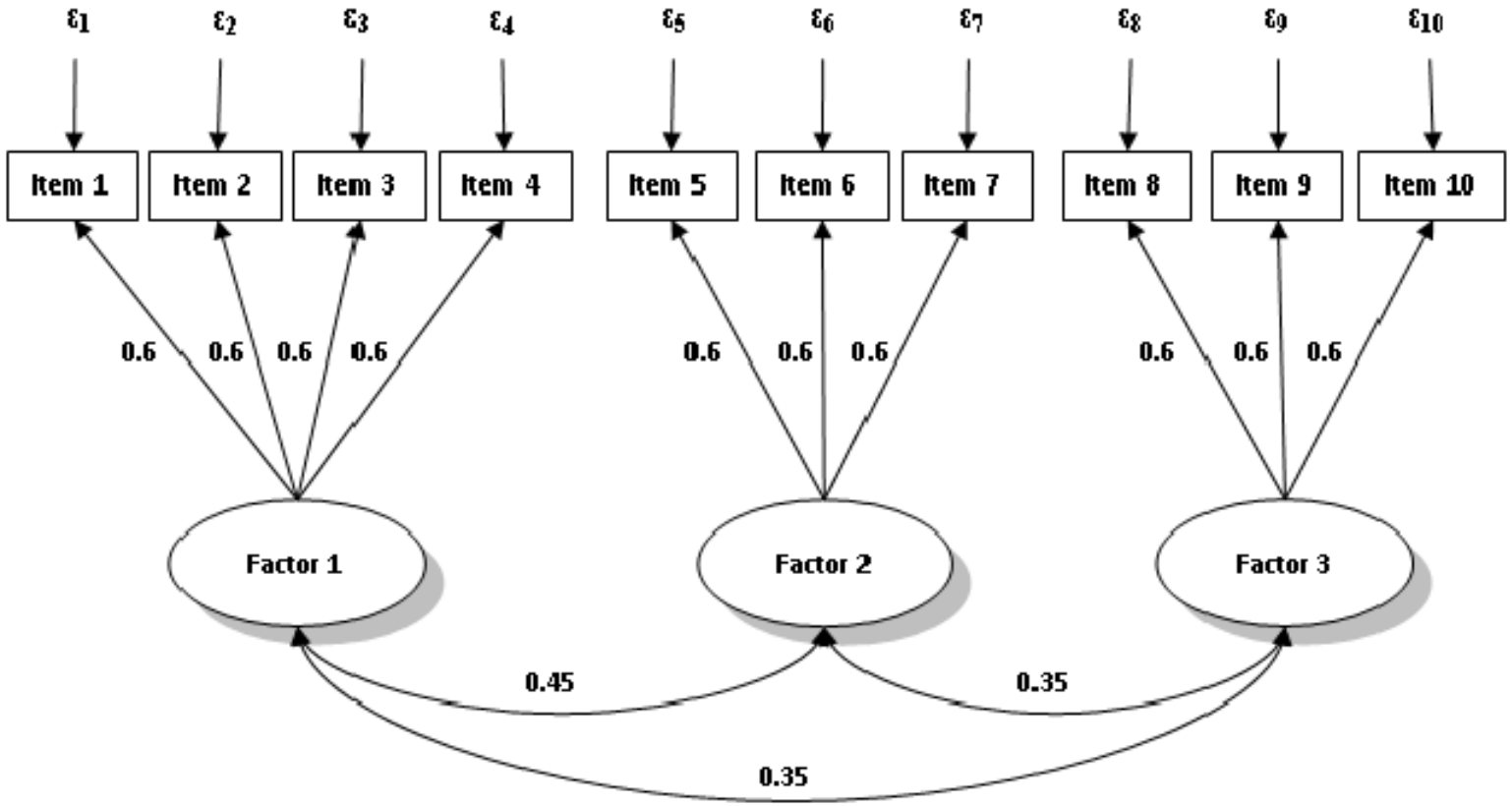


Figure 3: Mean of the values of the salient loadings after rotation on the 10000 simulations in relation to sample size. Example of a three-factor scale

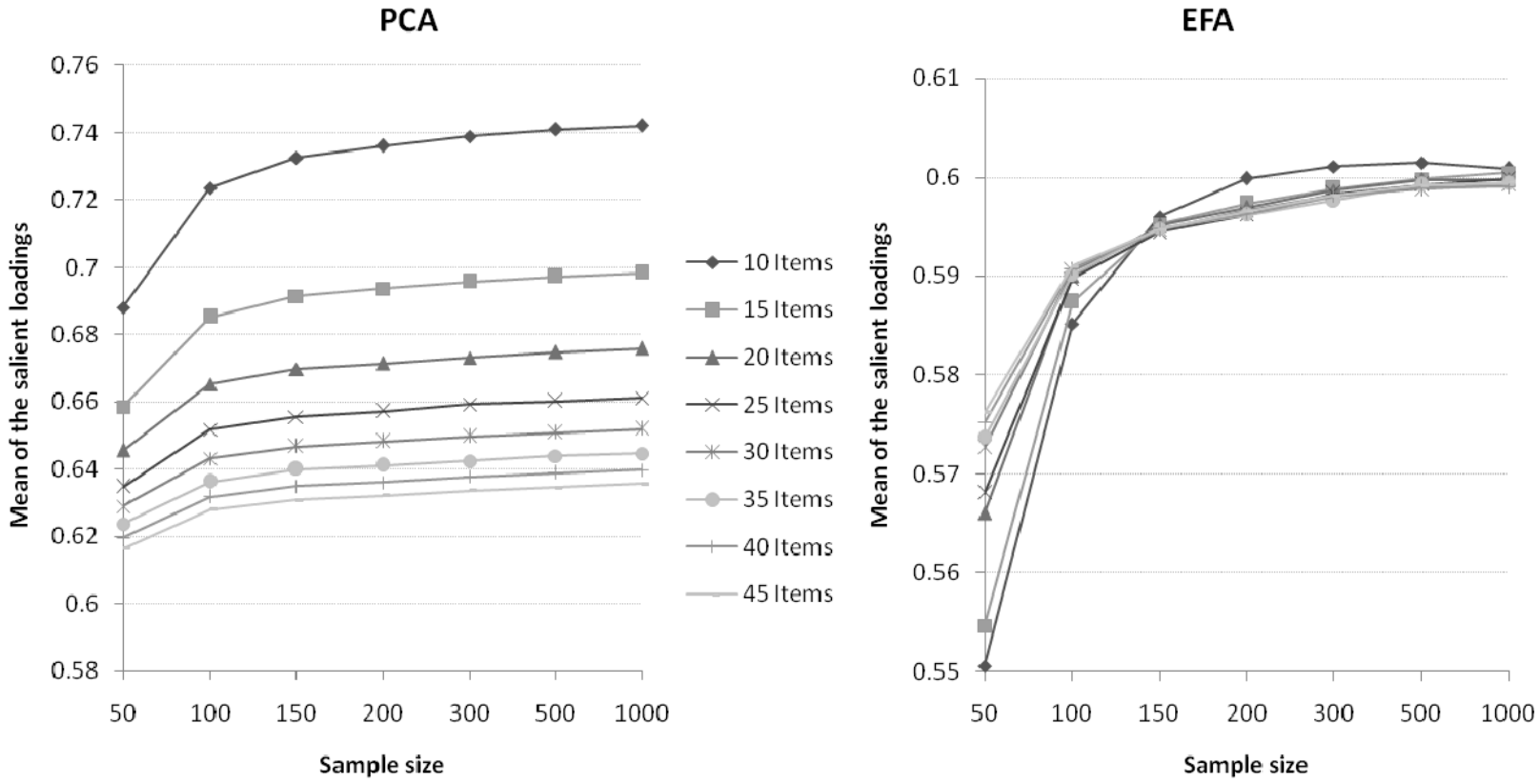


Figure 4: Mean of the standard deviations of the loadings over the 10 000 samples in relation to sample size in the case of Principal Component Analysis (PCA) or Exploratory Factor Analysis (EFA) followed by a promax rotation for the Beck Depression Inventory (BDI) and the Hamilton Depression Rating Scale (HAMD)

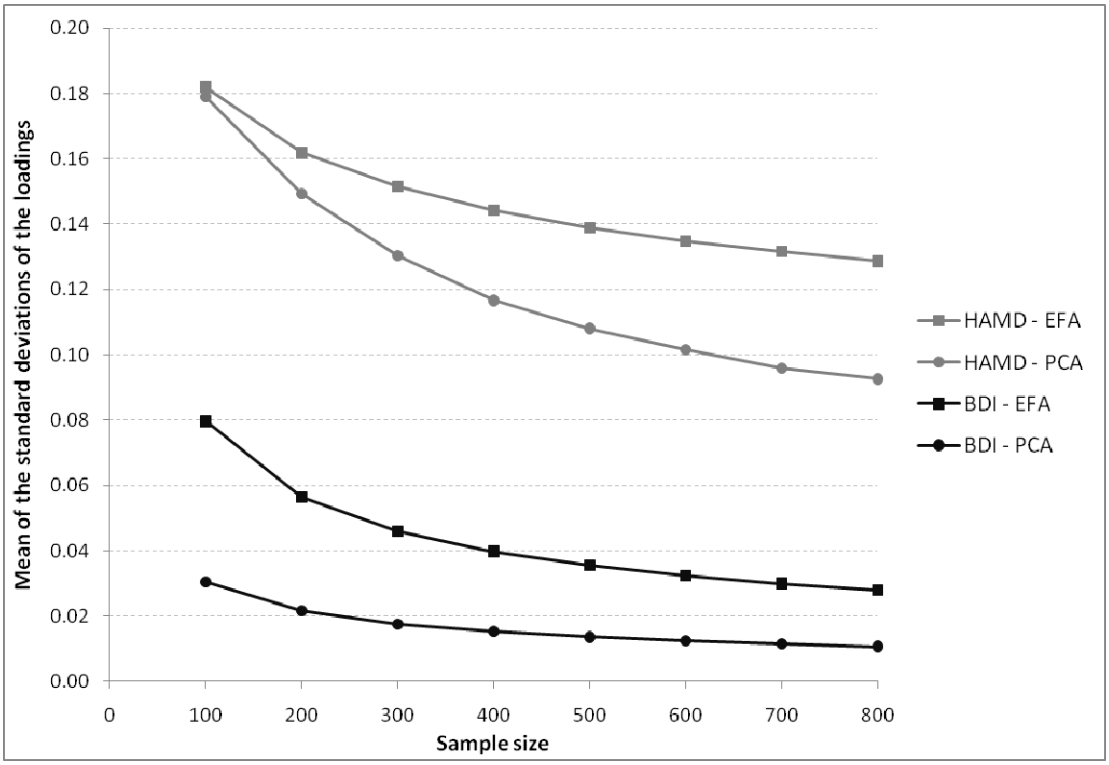
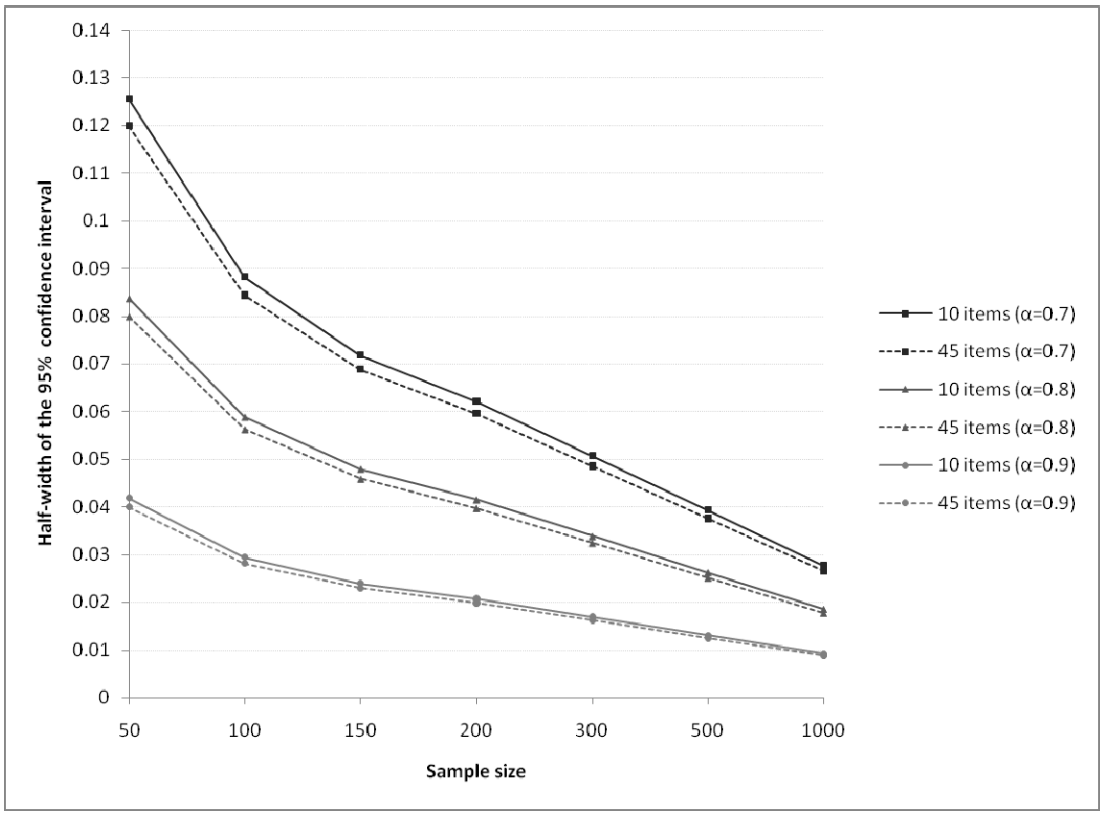


Figure 5: Half-width of the 95% confidence interval of Cronbach's alpha coefficient for three expected values (α) in relation to the sample size and the number of items



ANNEXE 2: THE MINIMAL CLINICALLY IMPORTANT DIFFERENCE DETERMINED USING ITEM RESPONSE THEORY MODELS: AN ATTEMPT TO SOLVE THE ISSUE OF THE ASSOCIATION WITH BASELINE SCORE (ARTICLE 2)

Alexandra Rouquette, Myriam Blanchin, Véronique Sébille, Francis Guillemin, Sylvana M Côté, Bruno Falissard and Jean-Benoit Hardouin.

ABSTRACT

Objective: Determining the Minimal Clinically Important Difference (MCID) of questionnaires on an interval scale, the Trait Level (TL) scale, using Item Response Theory (IRT) models could overcome its association with baseline severity. The aim of this study was to compare the Sensitivity (Se), Specificity (Sp) and Predictive Values (PV) of the MCID determined on the Score scale (MCID-Sc) or on the TL-scale (MCID-TL).

Study design and setting: The MCID-Sc and MCID-TL of the MOS-SF36 general health subscale were determined for deterioration and for improvement on a cohort of 1170 patients using an anchor-based method and a partial credit model. The Se, Sp and PV were calculated using the global rating of change (the anchor) as the gold standard test.

Results: The MCID-Sc magnitude was smaller for improvement (1.58 points) than for deterioration (-7.91 points). The Se, Sp and PV were similar for MCID-Sc and MCID-TL in both cases. However, if the MCID was defined on the score scale as a function of a range of baseline scores, its Se, Sp and PV were consistently higher.

Conclusion: This study reinforces the recommendations concerning the use of a MCID-Sc defined as a function of a range of baseline scores.

KEYWORDS: Minimal clinically important difference; Questionnaires; Sensitivity and specificity; Item response theory; Rasch models; Patient-reported outcomes

TEXT BOX

What is new?

- The Minimal Clinically Important Difference (MCID) defined as a function of a range of baseline scores leads to a better classification of individuals having experienced “at least a minimally important change” versus “no change" over time than the MCID defined without considering the baseline severity
- Determining the MCID using Item Response Theory (IRT) models does not greatly enhance its Sensitivity (Se), Specificity (Sp) and Predictive Values (PV) compared with its determination on the score scale
- The lack of interval scale properties of the score is not fully responsible for the MCID dependence on baseline severity

INTRODUCTION

Multi-item questionnaires are increasingly used in longitudinal studies to measure perceived health status and to assess its changes over time. Indeed, clinicians and policy makers are more and more interested in integrating patient's perspective and experience of disease and illness in the evaluation of treatments, interventions or public health policies [1-5]. However, a major limitation to the use of these measurement instruments in clinical research or epidemiological studies is their interpretability [6-17]. For instance, what is the meaning of a two-point reduction over a six month period when anxiety is assessed with a 20 points scale? Is it a trivial or a meaningful difference? The Minimal Clinically Important Difference (MCID) is a concept defined to help with the interpretation of observed differences obtained in longitudinal studies using questionnaires-[18].

The best method for determining the MCID of a questionnaire is still under debate, however, anchor-based methods are recommended by numerous authors as they compare observed score differences to external criteria that have clinical relevance [7, 11-12, 14, 19]. These criteria can be indicators of clinical response or of illness evolution but the most used are patient-based Global Ratings of Change (GRC) since they provide a simple measure of the significance of change from the individual perspective [3, 13-14, 19-20]. In practice, multiple anchors are more and more often used in the same study [21-23].

Several issues are, nevertheless, still complicating the MCID determination, especially its variations among populations, estimations approaches, etc., and raise questions about the existence of a unique questionnaire-specific MCID [11, 24-27]. One of these issues concerns the influence of the subjects' baseline score (BS) on the MCID value calculated using the anchor-based methods, as it has been shown in various studies [9, 27-36]. For that matter, various authors have recommended to define the MCID as a function of a range of BS rather than one single MCID [12, 14, 19, 30, 35]. Thus, to be able to conclude on the meaningfulness of someone's change, different MCID values should be considered depending on the subject's BS.

Several origins to this phenomenon have been mentioned [12]. The first one can be explained in psychophysical terms and is in relation with the subjective feature of the MCID concept: the subject's perception of a clinically meaningful change can be different depending on his/her baseline severity [37]. The second one is related to the statistical nature of the MCID concept and is called regression to the mean that describes the statistical tendency of extreme scores to become less extreme at follow-up [19, 30]. At last, two other potential origins of the MCID association with baseline severity concern the score itself (possibly weighted sum of the items responses) used as a measure of the construct (i.e. pain, anxiety, etc.) evaluated by the questionnaire. One of these origins is due to the upper and lower bounds of this score which are responsible for the floor/ceiling effects: patients whose BS is close to the ends of the scale are not able to register a large change because such a change would exceed the span of the scale [12, 25, 36-37]. The other one concerns the scale level of the score which has not necessarily the interval scale properties. With an interval scale, units along the scale are equal to one another [4, 36, 38-39]. The present study focuses on the potential lack of interval scale properties of the score and its role in the MCID dependence to BS phenomenon. Indeed, if the score scale is not an interval scale, interpretation of score differences can vary depending on the different portions of the scale.

Models from the Item Response Theory (IRT) are convenient tools to analyze questionnaire data and express the results on an interval scale. In this theory, the construct measured by the questionnaire, called latent trait, is assessed by a quantitative variable with interval scale properties, the Trait Level (TL) [40]. Thus, if the questionnaire measures anxiety (the latent trait) for example, an x-unit difference represents the same quantity whatever its location on the TL-scale (low, medium or high level of anxiety). If our hypothesis concerning the role of the interval scale properties in the MCID association with baseline severity is true, the MCID determination on the TL-scale using an IRT model, could therefore avoid this phenomenon. We could, thus, expect fewer misclassifications of individuals having experienced "at least a minimally important change" versus "no change" over time than with the MCID determined on the score scale.

The aim of our study is therefore to compare the Sensitivity (Se), Specificity (Sp), Positive and Negative Predictive Values (PPV and NPV) of the MCID determined on the Score scale (MCID-Sc) and the MCID determined on the TL-scale (MCID-TL) using an IRT model and an anchor-based method in which the external criteria is considered as the gold standard test.

METHODS

Data source

Data came from a French multicenter longitudinal prospective SATISQOL study composed of 1709 hospitalised patients, enrolled between October 2008 and September 2010, younger than 75 years-old and attending surgery or medical intervention for a chronic illness of one of the following systems: cardiovascular, musculoskeletal, nephrology, urology, digestive, pulmonary or endocrine. To be included, patients needed to speak French, have sufficient cognitive function to complete a self-administered questionnaire and exhibit symptoms of their chronic illness for, at least, six months. They were excluded if they did not have a therapeutic intervention during their hospitalisation.

Demographic information (age, sex, diagnosis, etc.), self-reported satisfaction with care (French version of the Patient Judgements of Hospital Quality questionnaire [41-42]) and quality of life (French version of the Medical Outcomes Study Short Form-36 questionnaire - MOS-SF36 [43-44]) were obtained during hospitalisation. Six months later, patients were asked to fill in the MOS-SF36 questionnaire again during a scheduled medical consultation. The study was approved by the ethic committee of Lorraine, France, and all the patients gave their informed consent to participate.

Questionnaire

The MOS-SF36 is a generic 36-items questionnaire divided into eight subscales addressing physical, mental and social health, and one item assessing health transition. To ensure the construct's unidimensionality required by the IRT model used in this study, analyses were performed on the five items of the General Health (GH) subscale. Each of these items was rated on an ordinal scale with five categories. The score, ranging from zero (worst

perceived general health) to 100, was computed as recommended by the MOS-SF36 user's guide [43]. Likewise, an individual mean imputation was performed if there were less than three missing responses in the GH-subscale as advocated.

The item assessing health transition at the six-month follow-up was chosen to be used as the GRC : "Compared to six months ago, how would you rate your health in general now?" Patients could choose between five responses: "Much better", "Somewhat better", "About the same", "Somewhat worse" and "Much worse".

Patients with three or more missing responses in the GH-subscale or who did not answer to the GRC at the six-month follow-up were excluded from the sample used for the analyses.

Analyses

Since it is well known that the amount and quality of change is likely to be different for improvement as compared to deterioration, the following analyses were performed in both circumstances [14-15, 19, 25, 28].

MCID-Sc determination

Changes in general health over the six-month interval were computed as the difference between baseline (T1) and six-month (T2) GH-subscale score. The MCID-Sc was computed as the mean score change from T1 to T2 in the subgroup of patients who answered "Somewhat better" to the GRC (SB group) for improvement and in the subgroup of patients who answered "Somewhat worse" (SW group) for deterioration. The dependence of score change to the BS was evaluated using Pearson's correlation coefficients. Polychoric correlation coefficients were used to assess association between score change and responses to the GRC.

Since it is recommended by various authors, a MCID-Sc composed of several values according to a range of Baseline Scores was determined: the MCID-Sc_{BS} [12, 14, 19, 30, 35]. Concretely, the MCID-Sc_{BS} was defined as the three means of score change from T1 to T2 for patients having a BS in the first third ([0 – 33]), the second third (]33 – 67[) or the higher third of the scale ([67 – 100]), in the SB group for improvement and in the SW group for deterioration.

MCID-TL determination

Assumptions of IRT

IRT models rely on three fundamental assumptions: unidimensionality, local independence and monotonicity. The unidimensionality of the GH subscale was checked, at each assessment time, using an eigenvalues analysis and the fit examination of a Confirmatory Factor Analysis (CFA) model with one factor. The Root Mean Square Error Approximation (RMSEA, acceptable fit if <0.06), the Comparative Fit Index (acceptable fit if >0.95), the Tucker Lewis Index (acceptable fit if >0.95) and the Standardized Root Mean Square Residual (acceptable fit if <0.08) were examined to evaluate the fit of the CFA model [45]. A non parametric IRT analysis was also performed by fitting a Monotonely Homogeneous Model of Mokken to our data. A good fit, evaluated by the Loevinger's H coefficients, indicates that the three IRT fundamental assumptions are [46]. Finally, the internal consistency of the GH subscale was checked by the computation of the Cronbach's alpha coefficient which was considered as acceptable if it was higher than 0.7 [47].

Fit of the Partial Credit Model (PCM) and item parameters estimation

A PCM, an IRT model for polytomous data (cf. supplementary material), was fitted on the data at T1 and T2 separately. A PCM was chosen because it is a model of the Rasch family which is very commonly used in the field of health related questionnaires [48]. This model defines MxJ item parameters with M the number of the response categories of the J items of the scale. In this model, the concept measured by the scale is represented by a random variable following a normal distribution. Fit tests, based on a chi-squared comparison, are known to be highly susceptible to large sample sizes. The PCM fit was, thus, adjusted for an expected sample of 400 individuals at both assessment times, which is a large enough sample to estimate the parameters of a PCM [49].

Measurement invariance of the GH subscale was checked using comparisons of the item parameters confidence intervals at both assessment times. As recommended if measurement invariance is met, averaged item parameters from across the two assessment

times were obtained by fitting a PCM on a data set made up of the T1 and T2 data sets [50-51].

MCID-TL determination

A latent regression IRT model was used to assess the TL mean variation over time within the SB/SW groups (cf. supplementary material). The MCID-TL was thus defined as the time effect on the TL-scale (TL mean change from T1 to T2) in the SB group for improvement and in the SW group for deterioration, respectively. To classify patients as having experienced “at least a minimally important change” or “no change", these MCID-TL had to be translated onto the score scale. A PCM was thus used to provide the relationship between the TL and the expected score at the GH subscale (cf. supplementary material). Using this translation tool, the score difference equivalent to the MCID-TL was determined for each BS varying from 0.5 to 99.5 by an increment of 0.5. Thus, knowing each patient’s BS, it was possible to determine if his/her score change over the six-month interval was larger than the MCID-TL or not. Due to the logistic form of the PCM, it was not possible to translate the MCID-TL for extreme BS (0 or 100); therefore it was approximated to the value obtained for the nearest BS (0.5 or 99.5 respectively).

Se, Sp, PPV and NPV computation

Each patient of the whole sample was classified as having experienced “at least a minimally important change” or “no change" over the six-month interval using the MCID-Sc, the MCID-Sc_{BS} and the MCID-TL classifications. Se, Sp, PPV and NPV were thus computed using the patient’s response at the GRC as the gold standard classification.

Software

Descriptive analysis, graphs, factor analysis and non parametric IRT analysis were performed using Stata®/MP 12.1 and the Microsoft® Office Excel 2007 spreadsheet program [52-53]. The items parameters and the PCM fit were estimated using RUMM® 2030 [54]. Finally, the SAS software 9.3 was used to estimate the MCID-TL values using the longitudinal form of the PCM with mixed effects [55].

RESULTS

At baseline, 1709 patients (877 men – 56.1%, 686 women – 43.9%, missing information for 146 patients) were entered. The average age of the participants was 55.7 (Standard Deviation – SD=14.0) years with a range of 18 to 80 years. At six-month follow-up, the response rate was 89.4%, i.e. 1528 patients. Amongst them, 58 did not answer to the GRC at T2 and 300 had more than two missing responses to the GH-subscale at T1 or T2, leaving 1170 patients for the analysis. The average GH-subscale score was 52.1 (SD=22.4) at T1 and 51.7 (SD=23.3) at T2. In the **figure 1** is depicted a histogram of the BS which was lower than or equal to 33 for 269 (23.0%) patients and higher than or equal to 67 for 372 (31.8%) patients.

[Figure 1 near here]

MCID-Sc determination

The response to the GRC was "Much better" for 266 (22.7%) patients, "Somewhat better" for 360 (30.8%), "About the same" for 401 (34.3%), "Somewhat worse" for 112 (9.6%) and "Much worse" for 31 (2.6%). The MCID-Sc of the GH subscale was equal to 1.58 (Standard Error - SE=0.76) points for improvement and to -7.91 (SE=1.26) points for deterioration. To notice, the mean score change in the group of patients considered as stable (who rated their health as "About the same" compared with six months ago) was -3.16 (SE=0.68). Polychoric correlation between score change and responses to the GRC was equal to -0.29.

Pearson's correlation between the score change and the BS was equal to -0.35 in the SB group and to -0.62 in the SW group. Box plots in **figure 2** show the variation of the score change over the six-month interval depending on the BS in the SB and SW groups. Globally, for improvement, the higher the BS, the smaller the score change. Conversely, for deterioration, the higher the BS, the larger the score change.

Means of score change specified in **figure 2** for each subgroup of patients defined by their BS were used to determine the MCID-Sc_{BS}. For instance, the MCID-Sc_{BS} for improvement was equal to 8.4 (SE=1.4) if the BS was included in [0 – 33] and to 2.5 (SE=1.0)

if it was included in]33 – 67[in the SB group. If the BS was included in [67 – 100] in the SB group, the MCID-Sc_{BS} for improvement was set to zero since the mean score change was negative in this subgroup.

[Figure 2 near here]

MCID-TL determination

At both times, only one eigenvalue was higher than one and the ratio of the first to the second eigenvalue was higher than four. All the criteria indicated an acceptable fit for the one factor CFA model, except the RMSEA which was equal to 0.088 at T1 and to 0.102 at T2. However, all the Loevinger's H coefficients did not detect any violation of the fundamental IRT assumptions. Finally, a good internal consistency was found at both assessment times with a Cronbach's alpha coefficient equal to 0.81 at T1 and to 0.84 at T2.

The assumptions of a good PCM fit to the data were not rejected at 5% (p=0.19 at T1 and p=0.32 at T2). The measurement invariance of the GH-subscale was assumed since the confidence interval of the 20 item parameters estimated at T1 overlapped with their confidence interval estimated at T2. The MCID-TL for improvement was estimated at 0.0839 (SE=0.0443) and at -0.4806 (SE=0.0833) for deterioration. It can be noted that the mean TL change in the group of patients considered as stable was equal to -0.1919 (SE=0.0426).

In the **figure 3** is depicted the relationship between the expected GH subscale score and the TL whose logistic shape is typical of the Rasch family models. Using this translation tool, it was possible to translate the MCID-TL on the score for each BS and represent it, as in the **figure 4** on the X-axis with the BS on the Y-axis. For example, a patient with a score of 20 on the GH subscale at baseline should have undergone a 1.5 points increase on his/her score at T2 to be classified as having experienced a minimal clinically important improvement using the MCID-TL whereas a patient with a BS equal to 80 should have undergone a 0.5 points increase.

[Figure 3 and figure 4 near here]

Se, Sp, PPV and NPV calculation

The Se, Sp and predictive values for the MCID-Sc, MCID-Sc_{BS} and the MCID-TL are shown for improvement and deterioration in **table 1**. All these values but one are lower than 80%.

[Table 1 near here]

DISCUSSION

Our study was designed to evaluate the advantages of IRT models for the determination of the MCID of the MOS-SF36 questionnaire GH subscale in a sample of hospitalized patients suffering from a chronic disease and undergoing a therapeutic intervention. In our study, the use of IRT models does not improve the Se, Sp and predictive values of the MCID-TL compared to the MCID-Sc, except for deterioration where its Se and predictive values seem slightly increased. For the MCID-Sc_{BS}, observed Se, Sp, and predictive values are consistently higher than for MCID-Sc or MCID-TL.

The overall lack of superiority of the MCID-TL compared to the MCID-Sc can be explained in considering **figures 1 and 3**. Indeed, in the **figure 3**, it can be seen that the relationship between the GH subscale score and the TL is quasi linear for a score ranged from 20 to 80 approximately. It means that, in this score range, the scale level of the GH subscale score nearly reaches the interval scale level. Moreover, in the study sample, 965 (82.5%) patients had a BS ranged in]20 – 80], as it can be seen in **figure 1**. It follows that few misclassifications of individuals having experienced “at least a minimally important change” versus “no change" over time, using the MCID-Sc, can be explained by the lack of interval scale properties of the score scale in our study. However, the magnitude of the MCID is another important factor to consider. Indeed, this magnitude is approximately five times larger for deterioration than for improvement. The slightly better MCID-TL’s performances in the case of deterioration suggested in our study could result from its magnitude since the lack of interval scale properties of the score could lead to more distortions in a large difference than in a small difference, i.e. the larger the quantity measured, the larger the discrepancy observed between its measures on the score scale or on the TL-scale.

The other important result of our study concerns the better results obtained with the MCID-Sc_{BS}. Further research should be done to disentangle the origins of this phenomenon and to determine if it could be explained by a different perception of change depending on the baseline severity. Indeed, for example, the MCID decrease with the increasing BS observed in the case of improvement could result from the ceiling effect as well as from the Regression To the Mean (RTM) phenomenon. In concrete terms, the ceiling effect is due to a lack of items able to measure a minimal clinically significant improvement for patients with an already high score at baseline. The score change observed for these patients is, therefore, lower than the change which would have been observed if there had been no ceiling effect. Although this effect is smaller than on the score scale, the use of the LT is also subject to floor and ceiling effects and it might be another reason for the lack of superiority of the MCID-TL compared to the MCID-Sc [56]. The RTM phenomenon is responsible for a higher probability of negative change score for patients in the upper part of the BS distribution (statistical tendency of extreme scores to become less extreme at follow-up). In our study, the RTM could explain the negative mean change score (-4.8) observed in the subgroup of patients with a BS comprised in [67 – 100] in the SB group (i.e. a decreasing mean score on the GH subscale from T1 to T2 whereas patients rated their health in general on the GRC at T2 as better than at T1).

One of the most cited limits of the anchor-based method concerns the validity of the anchor [19, 27, 29, 57]. In our study, the weak values of the Se, Sp, ~~and~~ predictive values and correlations observed between score change from T1 to T2 and the GRC raise questions about the validity of the MOS-SF36 health transition item used as an anchor. In the MOS-SF36 questionnaire, the response to this item is not used to compute the score of the other eight dimensions assessed and, consequently, of the GH subscale. This item's face validity is obviously good to assess change on the construct supposed to be measured by the GH-subscale. However, the mean change in the subgroup of patients considered as stable (health in general rated as "About the same" compared with six months ago) was negative on the score scale (-3.16) as well as on the TL-scale (-0.19). These results raise different questions [27]: is the construct measured by the GH-subscale the same as the "health in general" referred to in the GRC? Has this GRC still the same meaning for the patients when assessing their health six

month ahead (recall bias)? Finally, does response shift in one or several items of the GH-subscale occur from T1 to T2? Further analysis should be done to clarify these issues. Another limit should be discussed concerning the heterogeneity of diseases in the cohort used in this study. The use of a more valid anchor and/or a more homogeneous clinically-defined cohort may have improved the values of the Se, Sp, PPV and NPV for each of the MCID values calculated with the three different methods but would unlikely have favoured one method over another.

To our knowledge, this work is the first one which uses IRT models to determine the MCID on the TL. These models are powerful tools that make the measurement of subjective phenomenon on an interval level scale possible. However, our study shows that, for the GH-subscale of the MOS-SF36 questionnaire, the ability of a single MCID value to classify individuals as having experienced “at least a minimally important change” versus “no change" over time, is not enhanced if the MCID is determined on the TL-scale compared with the MCID-Sc. Furthermore, the recommendations done by various authors concerning the use of several MCID values according to the baseline severity (MCID-Sc_{BS}) values are reinforced by our results [13, 15, 22, 30, 35]. Methods to determine the number of values for the MCID-Sc_{BS} which leads to the highest Se and Sp for a scale should be developed. The choice of this number should obviously be balanced with the logistical challenge of a large number of values in practice, especially with separate MCID values for improvement and deterioration.

Fundings

The French National Research Agency, under reference N-2010-PRSP-008-01, supported this study. The SATISQOL cohort project was supported by an IRESP (Institut de recherche en santé publique) grant from Inserm, and a PHRC (Programme Hospitalier de Recherche Clinique) national grant from French Ministry of Health, France.

Conflict of interest: none declared

SUPPLEMENTARY MATERIAL

The Partial Credit Model

The PCM is an IRT model for polytomous data belonging to the Rasch models, in which the probability of a response y_l to an item j ($j = 1, \dots, J$) with l categories ($l = 1, \dots, m_j$) for the subject i ($i = 1, \dots N$) is a function of the subject's TL (denoted θ_i). It can be written:

$$P(Y_{ij} = y_l/\theta_i, \delta_{jl}) = \frac{exp\left(y_l\theta_i - \sum_{l=1}^{y}\delta_{jl}\right)}{\sum_{c=0}^{m_j} exp\left(c\theta_i - \sum_{l=1}^{c}\delta_{jl}\right)}$$

where δ_{jl} is the item parameter associated to the response category l of the item j [61]. The relationship between the TL and the expected score on the scale, $E(S)$, can be calculated using the following equation [41]:

$$E(S) = \sum_{j=1}^{J}\sum_{l=1}^{m_j} y_l \times P(Y_j = y_l/\theta, \delta_{jl})$$

The Longitudinal Mixed Partial Credit Model

If θ is considered as a random variable having, for example, a normal distribution $N(\mu, \sigma^2)$, the PCM is a mixed-effects logistic model in which the parameters to be estimated are μ (the TL mean in the sample), σ (the TL standard error in the sample) and δ_{jl} (the item parameters)[62]. For repeated measurements, a longitudinal form of the mixed-effects PCM has been developed as it has yet been done for the Rasch model by Blanchin et al [53, 63]. The probability of a response in the category y of an item j at time t ($t = 1, \dots T$) can be written as:

$$P(Y_{ij}^{(t)} = y^{(t)}/\theta_i^{(t)}, \delta_{jl}) = \frac{exp\left(y^{(t)}\theta_i^{(t)} - \sum_{l=1}^{y}\delta_{jl}\right)}{\sum_{c=0}^{m_j} exp\left(c\theta_i^{(t)} - \sum_{l=1}^{c}\delta_{jl}\right)}$$

The item parameters δ_{jl} are assumed to be constant within time assessments (i.e. measurement invariance is assumed) and can be estimated or are considered as known and fixed in the model. The other parameters to be estimated are the $\mu^{(t)}$ parameters (the TL mean in the

sample at time t), $\sigma^{(t)}$ (the TL standard error in the sample at time t) and $\sigma^{(tt\prime)}$ (the covariance between $\theta^{(t)}$ and $\theta^{(t\prime)}$) with $t \neq t'$. The distribution of the TL is assumed to be a multinormal distribution of dimension T. The mean change over time of the construct measured by the questionnaire in the entire sample can be evaluated by the evolution of $\mu^{(t)}$ across each time t.

Dichotomous group variables (G with realisation g_i for the subject i) can be introduced in this model to be able to assess this time effect in various groups in the sample:

$$P(Y_{ij}^{(t)} = y^{(t)} | \theta_i^{(t)}, \delta_{jl}, \alpha, \beta^{(t)}) = \frac{exp\left(y^{(t)}\theta_i^{(t)} + \boldsymbol{\alpha g_i} + \boldsymbol{\beta^{(t)} g_i} - \sum_{l=1}^{y} \delta_{jl}\right)}{\sum_{c=0}^{m_j} exp\left(c\theta_i^{(t)} + \boldsymbol{\alpha g_i} + \boldsymbol{\beta^{(t)} g_i} - \sum_{l=1}^{c} \delta_{jl}\right)}$$

With the identifiability constraint $\beta^{(1)} = 0$, in this model, α represents the mean difference between the two groups at first time ($t = 1$) and $\alpha + \beta^{(t)}$ represents the mean difference between the two groups at time t. Means of the TL at each time t and t', and in each group are indicated in **table 2**. If the variable G indicates the response to the GRC (0 indicating "About the same" and 1 indicating "Somewhat better"), the MCID-TL for improvement is equal to the mean change of the TL from time t to time t' in the group $G = 1$, that is $(\mu^{(t')} - \mu^{(t)}) + (\beta^{(t')} - \beta^{(t)})$.

[Table 2 near here]

REFERENCES

1. Clancy CM and Eisenberg JM. (1998) Outcomes research: measuring the end results of health care. Science. 282:245-6.

2. Roger VL. (2011) Outcomes research and epidemiology: the synergy between public health and clinical practice. Circ Cardiovasc Qual Outcomes. 4:257-9.

3. US Department of Health and Human Services (USDHHS). Guidance for industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. (2009); Available from: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatory/Information/Guidances/UCM193282.pdf.

4. McHorney CA. (1997) Generic health measurement: past accomplishments and a measurement paradigm for the 21st century. Ann Intern Med. 127:743-50.

5. Ellwood PM. (1988) Shattuck lecture--outcomes management. A technology of patient experience. N Engl J Med. 318:1549-56.

6. Guyatt GH and Cook DJ. (1994) Health status, quality of life, and the individual. JAMA. 272:630-1.

7. Liang MH. (2000) Longitudinal construct validity: establishment of clinical meaning in patient evaluative instruments. Med Care. 38:II84-90.

8. Norman GR, Stratford P and Regehr G. (1997) Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol. 50:869-79.

9. Stucki G, Daltroy L, Katz JN, Johannesson M and Liang MH. (1996) Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. J Clin Epidemiol. 49:711-7.

10. Beaton DE, Bombardier C, Katz JN, Wright JG, Wells G, Boers M, *et al.* (2001) Looking for important change/differences in studies of responsiveness. OMERACT MCID

Working Group. Outcome Measures in Rheumatology. Minimal Clinically Important Difference. J Rheumatol. 28:400-5.

11. Beaton DE, Boers M and Wells GA. (2002) Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. Curr Opin Rheumatol. 14:109-14.

12. Copay AG, Subach BR, Glassman SD, Polly DW, Jr. and Schuler TC. (2007) Understanding the minimum clinically important difference: a review of concepts and methods. Spine J. 7:541-6.

13. Cook CE. (2008) Clinimetrics Corner: The Minimal Clinically Important Change Score (MCID): A Necessary Pretense. J Man Manip Ther. 16:E82-3.

14. Revicki D, Hays RD, Cella D and Sloan J. (2008) Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol. 61:102-9.

15. Beaton DE, van Eerd D, Smith P, van der Velde G, Cullen K, Kennedy CA, *et al.* (2011) Minimal change is sensitive, less specific to recovery: a diagnostic testing approach to interpretability. J Clin Epidemiol. 64:487-96.

16. de Vet HC, Terluin B, Knol DL, Roorda LD, Mokkink LB, Ostelo RW, *et al.* (2010) Three ways to quantify uncertainty in individually applied "minimally important change" values. J Clin Epidemiol. 63:37-45.

17. Ferreira ML, Herbert RD, Ferreira PH, Latimer J, Ostelo RW, Nascimento DP, *et al.* (2011) A critical review of methods used to determine the smallest worthwhile effect of interventions for low back pain. J Clin Epidemiol. 65:253-61.

18. Jaeschke R, Singer J and Guyatt GH. (1989) Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials. 10:407-15.

19. Crosby RD, Kolotkin RL and Williams GR. (2003) Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol. 56:395-407.

20. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, *et al.* (2007) Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 60:34-42.

21. Yost KJ, Sorensen MV, Hahn EA, Glendenning GA, Gnanasakthy A and Cella D. (2005) Using multiple anchor- and distribution-based estimates to evaluate clinically meaningful change on the Functional Assessment of Cancer Therapy-Biologic Response Modifiers (FACT-BRM) instrument. Value Health. 8:117-27.

22. Purcell A, Fleming J, Bennett S, Burmeister B and Haines T. (2010) Determining the minimal clinically important difference criteria for the Multidimensional Fatigue Inventory in a radiotherapy population. Support Care Cancer. 18:307-15.

23. Sloan JA. (2005) Assessing the minimally clinically significant difference: scientific considerations, challenges and solutions. COPD. 2:57-62.

24. Revicki DA, Cella D, Hays RD, Sloan JA, Lenderking WR and Aaronson NK. (2006) Responsiveness and minimal important differences for patient reported outcomes. Health Qual Life Outcomes. 4:70.

25. Hays RD and Woolley JM. (2000) The concept of clinically meaningful difference in health-related quality-of-life research. How meaningful is it? Pharmacoeconomics. 18:419-23.

26. Beaton DE. (2003) Simple as possible? Or too simple? Possible limits to the universality of the one half standard deviation. Med Care. 41:593-6.

27. Terwee CB, Roorda LD, Dekker J, Bierma-Zeinstra SM, Peat G, Jordan KP, *et al.* (2010) Mind the MIC: large variation among populations and methods. J Clin Epidemiol. 63:524-34.

28. Stratford PW, Binkley JM, Riddle DL and Guyatt GH. (1998) Sensitivity to change of the Roland-Morris Back Pain Questionnaire: part 1. Phys Ther. 78:1186-96.

29. de Vet HC, Ostelo RW, Terwee CB, van der Roer N, Knol DL, Beckerman H, *et al.* (2007) Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. Qual Life Res. 16:131-42.

30. Crosby RD, Kolotkin RL and Williams GR. (2004) An integrated method to determine meaningful changes in health-related quality of life. J Clin Epidemiol. 57:1153-60.

31. Stratford PW, Binkley J, Solomon P, Finch E, Gill C and Moreland J. (1996) Defining the minimum level of detectable change for the Roland-Morris questionnaire. Phys Ther. 76:359-65; discussion 66-8.

32. Lauridsen HH, Hartvigsen J, Manniche C, Korsholm L and Grunnet-Nilsson N. (2006) Responsiveness and minimal clinically important difference for pain and disability instruments in low back pain patients. BMC Musculoskelet Disord. 7:82.

33. Jensen MP, Chen C and Brugger AM. (2003) Interpretation of visual analog scale ratings and change scores: a reanalysis of two clinical trials of postoperative pain. J Pain. 4:407-14.

34. ten Klooster PM, Drossaers-Bakker KW, Taal E and van de Laar MA. (2006) Patient-perceived satisfactory improvement (PPSI): interpreting meaningful change in pain from the patient's perspective. Pain. 121:151-7.

35. Tubach F, Ravaud P, Baron G, Falissard B, Logeart I, Bellamy N, *et al.* (2005) Evaluation of clinically relevant changes in patient reported outcomes in knee and hip osteoarthritis: the minimal clinically important improvement. Ann Rheum Dis. 64:29-33.

36. Bird SB and Dickson EW. (2001) Clinically significant changes in pain along the visual analog scale. Ann Emerg Med. 38:639-43.

37. Baker DW, Hays RD and Brook RH. (1997) Understanding changes in health status. Is the floor phenomenon merely the last step of the staircase? Med Care. 35:1-15.

38. Stevens SS. (1946) On the Theory of Scales of Measurement. Science. 103:677-80.

39. Samsa G, Edelman D, Rothman ML, Williams GR, Lipscomb J and Matchar D. (1999) Determining clinically important differences in health status measures: a general approach with illustration to the Health Utilities Index Mark II. Pharmacoeconomics. 15:141-55.

40. Embretson SE and Reise SP. Item Response Theory for Psychologists: L. Erlbaum Associates; 2000.

41. Nguyen Thi PL, Briançon S, Empereur F and Guillemin F. (2002) Factors determining inpatient satisfaction with care. Soc Sci Med. 54:493-504.

42. Rubin HR, Ware JE, Jr., Nelson EC and Meterko M. (1990) The Patient Judgments of Hospital Quality (PJHQ) Questionnaire. Med Care. 28:S17-8.

43. Leplège A, Ecosse E, Coste J, Pouchot J and Perneger T. Le questionnaire MOS SF-36: Manuel de l'utilisateur et guide d'interprétation des scores: Estem; 2001.

44. Ware JE, Jr. and Sherbourne CD. (1992) The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care. 30:473-83.

45. Hu L and Bentler PM. (1999) Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal. 6:1-55.

46. Sijtsma K and Molenaar IW. Introduction to Nonparametric Item Response Theory: SAGE Publications; 2002.

47. Cronbach LJ. (1951) Coefficient alpha and the internal structure of a test. Psychometrika 16:297–334.

48. Anthoine E, Moret L, Regnault A, Sébille V and Hardouin J-B. (Submitted) How PRO measures are psychometrically validated? A review of publications on primary validation.

49. Smith AB, Rush R, Fallowfield LJ, Velikova G and Sharpe M. (2008) Rasch fit statistics and sample size considerations for polytomous data. BMC Med Res Methodol. 8:33.

50. Wright BD. (1996) Comparison requires stability. Rash Measurement Trans. 10:506.

51. Norquist JM, Fitzpatrick R, Dawson J and Jenkinson C. (2004) Comparing alternative Rasch-based methods vs raw scores in measuring change in health. Med Care. 42:I25-36.

52. Hardouin J-B, Bonnaud-Antignac A and Sébille V. (2011) Nonparametric item response theory using Stata. Stata Journal. 11:30-51.

53. StataCorp LP. Stata Statistical Software: Release 12.1. College Station, TX2012.

54. Andrich D, Sheridan BS and Luo G. Rumm2030: Rasch Unidimensional Measurement Models [computer software]. Perth, Western Australia: RUMM Laboratory; 2010.

55. SAS Institute Inc. Procedures Guides. Cary, NC: SAS Institute Inc; 2010.

56. Revicki DA and Cella DF. (1997) Health status assessment for the twenty-first century: item response theory, item banking and computer adaptive testing. Qual Life Res. 6:595-600.

57. Kemmler G, Giesinger J and Holzner B. (2011) Clinically relevant, statistically significant, or both? Minimal important change in the individual subject revisited. J Clin Epidemiol. 64:1467-8.

FIGURES AND TABLES

Figure 1: Histogram of the general health (GH) subscale score at baseline

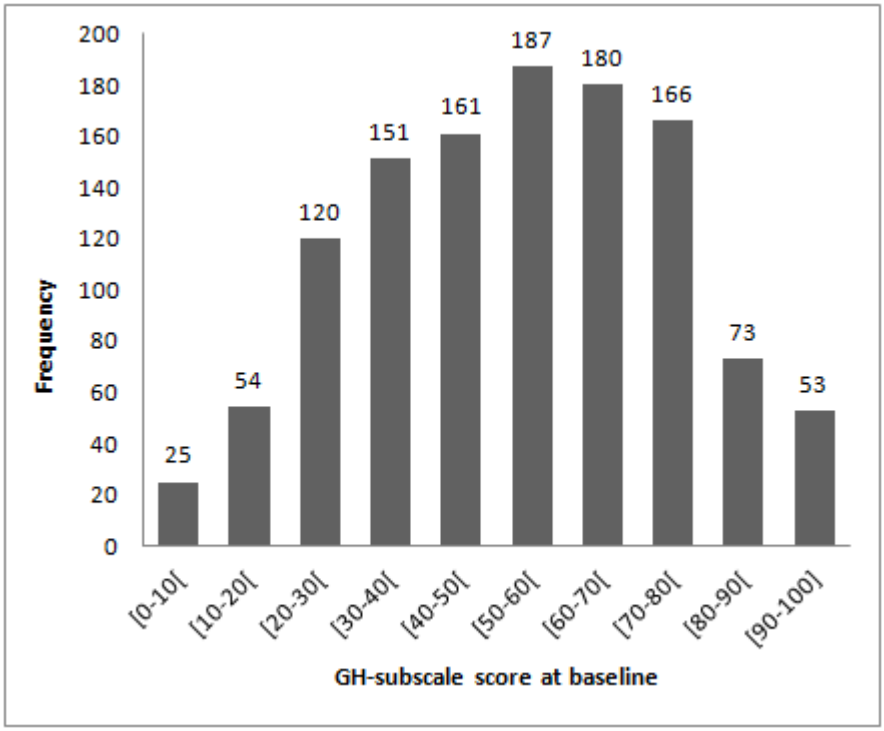


Figure 2: Box plots of the general health subscale score change from time 1 to time 2, depending on the score at time 1, in the subgroups of patients who answered "Somewhat better" (Improvement) or "Somewhat worse" (Deterioration) to the global rating of change (μ: Mean, SE: Standard Error, N: Number of patients)

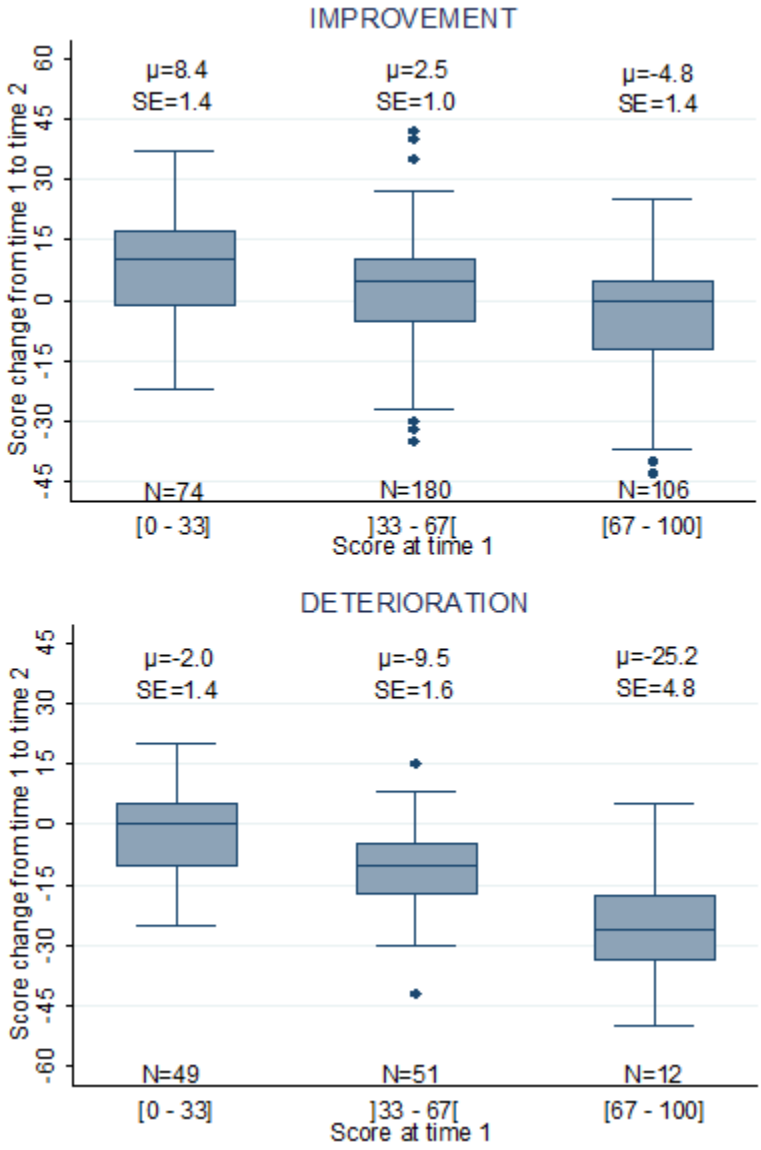


Figure 3: Expected general health subscale score depending on the trait level

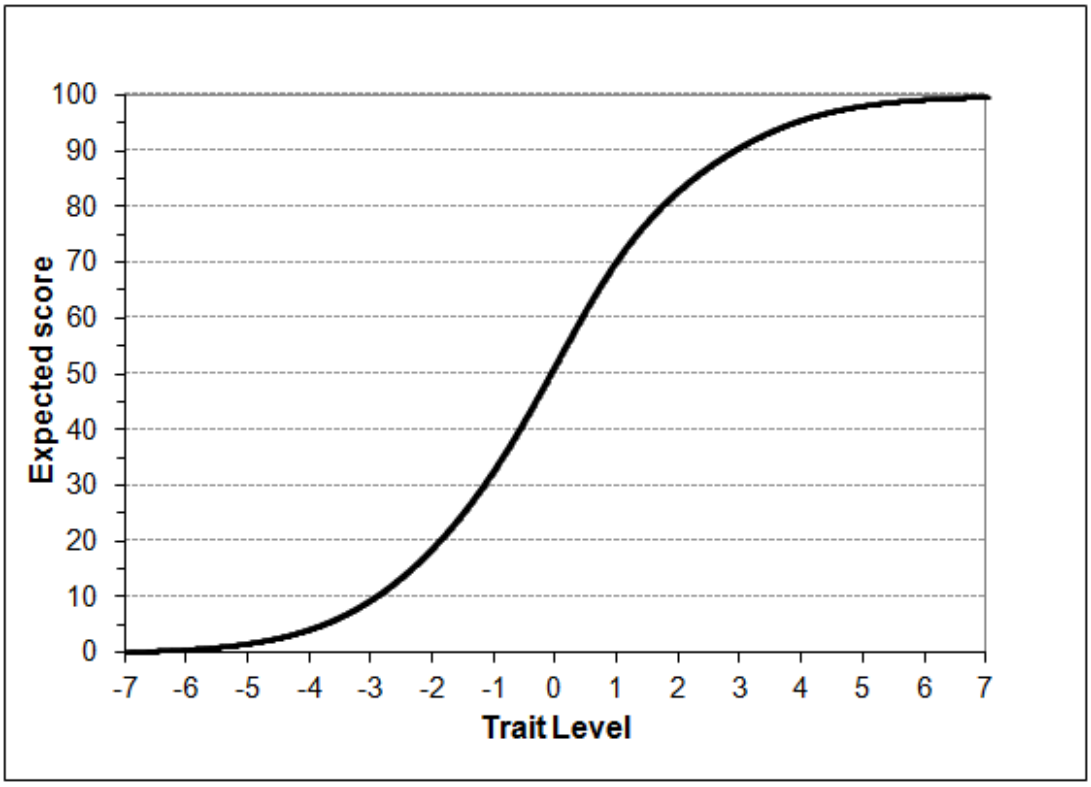


Figure 4: Minimal Clinically Important Difference determined on the Trait Level for improvement (MCID-TL= 0.0839) and deterioration (MCID-TL= -0.4806), translated on the score scale, depending on the baseline score

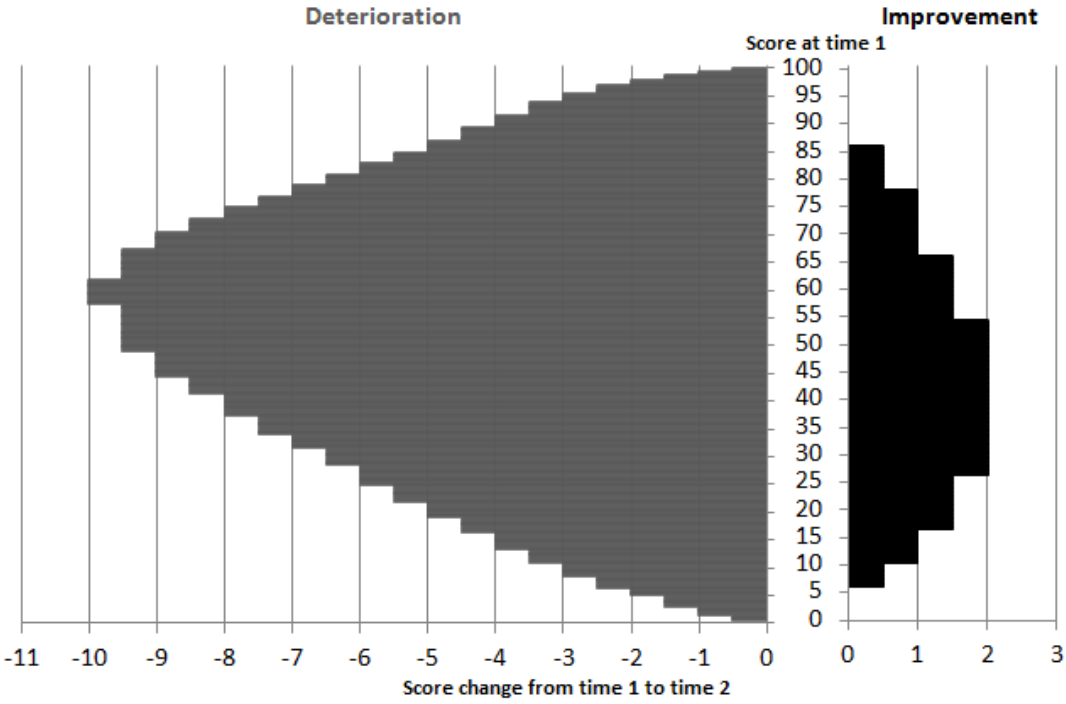


Table 1: Sensitivity (Se), Specificity (Sp), Positive Predictive Value (PPV) and Negative Predictive Value (NPV) for the Minimal Clinically Important Difference determined on the score scale (MCID-Sc), on the trait level (MCID-TL) or defined as a range of values on the score scale according to the Baseline Score (MCID-Sc_{BS}) of the general health subscale for people who rated their health as better (Improvement) or worse (Deterioration) compared to six month ago.

		Se [$CI_{95\%}$]	Sp [$CI_{95\%}$]	PPV [$CI_{95\%}$]	NPV [$CI_{95\%}$]
Improvement	MCID-Sc	54.6% [50.7 – 58.5]	65.6% [60.9 – 70.2]	71.3% [67.2 – 75.3]	48.1% [43.9 – 52.3]
	MCID-Sc_{BS}	56.6% [52.7 – 60.4]	68.6% [64.0 – 73.1]	73.8% [69.8 – 77.7]	50.3% [46.1 – 54.5]
	MCID-TL	54.6% [50.7 – 58.5]	65.8% [61.2 – 70.5]	71.4% [67.4 – 75.5]	48.2% [44.0 – 52.4]
Deterioration	MCID-Sc	44.1% [35.9 – 52.2]	65.8% [61.2 – 70.5]	31.5% [25.1 – 37.9]	76.7% [72.3 – 81.2]
	MCID-Sc_{BS}	53.2% [45.0 – 61.3]	75.8% [71.6 – 80.0]	43.9% [36.5 – 51.3]	81.9% [78.0 – 85.9]
	MCID-TL	51.1% [42.9 – 59.2]	63.8% [59.1 – 68.5]	33.5% [27.2 – 39.8]	78.5% [74.1 – 83.0]

$CI_{95\%}$: Confidence Interval 95%

Table 2: Means of the Trait Level (TL) at each time t and t' and in each group G when a dichotomous group variable is introduced in the longitudinal mixed Partial Credit Model (for example, G indicates the response to the Global Rating of Change: 0 indicating "About the same" and 1 indicating "Somewhat better")

	Group	
	$G = 0$	$G = 1$
Time t	$\mu^{(t)}$	$\mu^{(t)} + \alpha + \beta^{(t)}$
Time t'	$\mu^{(t\prime)}$	$\mu^{(t\prime)} + \alpha + \beta^{(t\prime)}$
Mean change of the TL (Time effect)	$\mu^{(t\prime)} - \mu^{(t)}$	$\left(\mu^{(t')} - \mu^{(t)}\right) + (\beta^{(t')} - \beta^{(t)})$