UNIVERSITE BRETAGNE LOIRE

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Antoine VANIER

Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université de Nantes
sous le sceau de l'Université Bretagne Loire

École doctorale : ED 502 « Biologie, Santé »

Discipline : Recherche clinique, Innovation technologique, Santé publique
Spécialité : Biostatistique
Unité de recherche : EA 4275 SPHERE – "methodS in Patient-centered outcomes and HEalth REsearch"

Soutenue le 18 Octobre 2016
Thèse N° :

The concept, measurement, and integration of response shift phenomenon in Patient-Reported Outcomes data analyses

On certain methodological and statistical considerations

JURY

Rapporteurs :	**Frans OORT**, Full Professor, University of Amsterdam
	Mirjam SPRANGERS, Full Professor, University of Amsterdam
Examinateurs :	**Nancy MAYO,** Full Professor, McGill University Montreal
	Carolyn SCHWARTZ, Adjunct Research Professor, Tufts University Boston
Directeur de Thèse :	**Jean-Benoit HARDOUIN**, Associate Professor-ScD, University of Nantes
Co-directeur de Thèse :	**Véronique SÉBILLE**, Full Professor, University of Nantes

Research team where the thesis was conducted

EA 4275 SPHERE – “methodS in Patient-centered outcomes and HEalth REsearch”

Directrice : Pr Véronique Sébille

Université de Nantes – Institut de Recherche en Santé IRS2

22, Boulevard Bénoni-Goullin

44200 Nantes

Scientific output

- **Original articles published in international peer-reviewed journals**

Vanier A, Sébille V, Blanchin M, Guilleux A, Hardouin JB. Overall performance of Oort's procedure for response shift detection: a pilot simulation study. *Quality of Life Research* 2015 24(8):1799-1807

Vanier A, Leplège A, Hardouin J.B., Sébille V, Falissard B. Semantic primes theory may be helpful in designing questionnaires such as to prevent response shift. *Journal of Clinical Epidemiology* 2015;68(6):646-54

- **Book Chapter**

Vanier A, Falissard B, Sébille V, Hardouin JB**.** The complexity of interpreting changes observed over time in Health-Related Quality of Life: a short overview of 15 years of research on response shift theory. Accepted. To be published in Guillemin F, Leplège A, Briançon S, Spitz E, Coste J. (2016). *Perceived health and adaptation in chronic disease. Stakes and future challenge*. New-York-NY. Taylor and Francis

- **Other article in international peer-reviewed journal with other work on response shift**

Guilleux A, Blanchin M, Vanier A, Guillemin F, Falissard B, Schwartz CE, Hardouin JB, Sébille V. RespOnse Shift ALgorithm in Item response theory (ROSALI) for response shift detection with missing data in longitudinal patient-reported outcomes studies. *Quality of Life Research* 2015;24(3):553-64

- **Communications**

(Oral) Vanier A, Leplège A, Hardouin J.B, Sébille V, Falissard B (2014). Semantic primes, une possibilité pour éviter le Response Shift ? Nancy, 3-6 Juin, *Colloque APEMAC, Journée des jeunes chercheurs, Maladie chroniques, adaptation et santé perçue : enjeux et prospective*

(Oral) Vanier A, Leplège A, Hardouin J.B., Sébille V, Falissard B (2013). Some conceptual issues concerning Response Shift in HRQL measurement, Nantes, 7 Juin, *International Workshop "Response Shift and subjective measures in health sciences". Scientific meetings of the university of Nantes*

(Poster) Vanier A, Sébille V, Blanchin M, Guilleux A, Hardouin J.B (2014). Power and Type-I-Error of global assessment of response shift occurrence using Likelihood-Ratio Test or information criteria in Oort's procedure: a simulation study. Berlin, 15-18 Octobre, *21th Annual Conference of International Society of Quality Of Life*

(Poster) Vanier A, Leplège A, Hardouin J.B, Sébille V, Falissard B (2014). Semantic primes: a possible way to avoid Response Shift? Berlin, 15-18 Octobre, *21th annual conference of International SOciety of Quality Of Life*

- **Award**

Prize of the best oral communication. Nancy, 3-6 Juin 2014, *Colloque APEMAC, Journée des jeunes chercheurs, Maladie chroniques, adaptation et santé perçue : enjeux et prospective*

- **Overall quantitative output as a MD specialized in public health (biostatistics/epidemiology) / researcher**

Up to this date, 11 original articles published in peer-reviewed journals, 1 book chapter, 36 communications (oral or poster), 6 original articles in review process

Remerciements (Aknowledgments)

Note to English-speakers: As this section is not a part of the core of the manuscript and is dedicated to personal thanks, it will be written in my native tongue (French).

Maintenant que la rédaction du manuscrit est achevée, il s'agit de m'atteler à l'exercice le plus ardu du doctorant (il requiert en effet un niveau de diplomatie florentine qui ne peut probablement pas être appréhendé par une théorie psychométrique existante, si moderne soit-elle) : les remerciements. L'importance de cette tâche est d'ailleurs reflétée par la position cruciale de cette section dans le document (la suite est de moindre importance…). J'espère sortir victorieux de cette épreuve, en effet, mon niveau de fonctionnement social mesuré par la SF-36 pourrait considérablement chuter à la suite d'un si facile, mais ô combien crucial oubli.

Lorsque j'ai rédigé ma thèse d'exercice pour l'obtention du diplôme d'état de docteur en médecine il y a maintenant près de quatre ans, je m'étais attaché à suivre un ordre chronologique quant aux remerciements des différentes personnes qui avaient influées sur mon parcours universitaire du début jusqu'à ce moment. Bien évidemment, quatre ans plus tard, les lois fondamentales de la causalité n'ayant pas changé, cet ensemble de personnes est toujours à remercier. Toutefois, ayant maintenant quatre ans de plus, et étant donc de fait forcément plus cynique, aigri et débonnaire (c'est bien l'évolution naturelle de toute personne selon l'âge non ?), je ne réitèrerai pas au complet cet exercice et me concentrerai cette fois-ci sur l'ensemble des personnes ayant gravité dans l'univers plus local de la rédaction de ce présent manuscrit. Par ailleurs, pour des raisons que les circonvolutions de mon cortex préfrontal ignorent (un déni d'oisiveté en vérité), la présente méthodologie pour choisir l'ordre des remerciements sera un tirage au sort non aléatoire sans remise avec une probabilité de succès attribuée à chacun des individus selon une distribution que je qualifierai de chaotique.

Les deux grands gagnants ex-aequo de ce tirage au sort sont donc Jean-Benoit Hardouin et Véronique Sébille, soit mes directeurs de thèse. Je vous remercie infiniment car vous faîtes partie probablement des rares personnes capables de maîtriser une délicate équation qui est de me laisser explorer mes lubies du moment (j'avoue ne pas avoir spécialement prévu au début que j'utiliserai le terme « molécule sémantique » dans une thèse de biostatistique…) tout en ne transigeant pas quant à la rigueur scientifique et à l'honnêteté intellectuelle (et à la bonne humeur aussi). En ce sens, vous êtes des mentors, au sens noble du terme. Toutefois, je confesse pour cette fois-ci avoir peut-être un peu abusé, et je fais la promesse solennelle (les écrits

restent) à Jean-Benoit que je serai plus focalisé quant à mon futur – j'espère – excitant post-doctorat.

Je remercie ensuite bien évidemment Mirjam Sprangers, Frans Oort, Nancy Mayo et Carolyn Schwartz, les rapporteurs et examinateurs de mon jury de thèse. C'est bien simple, sans votre existence le sujet principal de ma thèse n'existerait tout simplement pas, aussi, j'avoue être particulièrement honoré que mon modeste travail soit jugé par des acteurs aussi majeurs de la recherche sur le response shift.

Tout proche de ces personnalités de choc, je remercie Alice Guilleux et Myriam Blanchin, respectivement doctorante et ingénieure de recherche dans l'équipe, qui avec les personnes suscitées de l'unité et moi-même avons formé le petit groupe initial de gens qui ont décidé de s'attaquer à ce drôle de truc qui s'appelle le response shift. Merci pour toute l'aide apportée, pour les discussions enrichissantes, et aussi pour avoir le bon goût de faire semblant de supporter mes blagues de post-adolescent mal fini pendant nos réunions de travail.

Je remercie ensuite Bruno Falissard, qui doit assumer sa part de responsabilité quant à l'orientation exotique, mais rafraichissante, qu'a prise une partie de ces travaux originaux de thèse. Je remercie de même Alain Leplège, qui a su me guider quand mon esprit s'égarait dans les limbes de la sémantique et de « l'épistémologie » (probablement un peu fort pour qualifier le travail, disons plutôt un travail méthodologique avec un soupçon d'épistémologie).

Je remercie ensuite Francis Guillemin, qui est visiblement suffisamment sûr de lui pour avoir accepté la rédaction d'un chapitre d'un ouvrage sous sa direction par quelqu'un d'aussi chérubin que moi, j'espère avoir été à la hauteur.

Même si pour des raisons géographiques j'ai depuis trois ans moins de contacts avec le reste de l'équipe (mais préparez-vous, je reviens bientôt…), je remercie bien sur tous les membres de l'EA 4275 pour tous les agréables moments partagés, notamment les journées scientifiques toujours sémillantes. J'espère qu'avec mon retour à plein temps dans les locaux Nantais, de futures collaborations s'envisageront.

Je remercie ensuite Sophie Tézenas du Montcel, Jean-Luc Kop et Jacques Pouchot qui ont acceptés de faire partie de mon comité de thèse annuel sans aucune autre contrepartie que de l'altruisme scientifique. Je n'ai malheureusement pas eu le temps de développer certaines des idées suggérées pendant ces réunions, mais elles ne sont pas pour autant tombées dans l'oreille d'un sourd.

Je remercie Pierre Rufat, ainsi que de nouveau Sophie Tézenas du Montcel, pour m'avoir permis de finaliser sereinement ce manuscrit tout en bénéficiant de mon cadre de travail habituel au BIOSPIM.

Je remercie Alain Mallet (et Sophie de nouveau.), avec qui j'ai majoritairement participé à l'enseignement de biostatistique de l'Université Pierre et Marie Curie Paris 6. Avoir bénéficié de leur tutorat m'a sans aucun doute permis de faire le point sur les fondamentaux d'une discipline dont je tente peu à peu de prétendre que je fais partie des gens qui sont censés avoir un certain niveau d'expertise.

Je remercie tous mes confrères du BIOSPIM qui m'ont permis de me ragaillardir pendant nos pauses déjeuner alors que la rédaction de ce manuscrit s'apparentait parfois à la réalisation d'un travail herculéen.

Je remercie ma famille, qui quatre ans après ma première thèse a toujours l'obligation de me supporter, et qui a d'ailleurs remplie cette fonction avec un brio certain durant une année 2016 qui fut un peu particulière.

Je remercie tous mes amis, qui visiblement, choisissent toujours de me supporter pour une raison cosmique quelconque. Je remercie particulièrement Flora, qui m'a donné un cadre sécurisant à un moment critique, me permettant de finaliser mes travaux sereinement.

Enfin, comme toujours, je remercie Sam, peu importe les changements d'étiquettes passé, présent, futur (espérons-le) et au-delà, peu importe le plan de réalité, nous veillerons toujours l'un sur l'autre, comme une paire de quarks perdue dans un univers quantique stochastique.

Préface (Preface)

Note to English-speakers: As this section is not a part of the core of the manuscript and is dedicated to personal thoughts, it will be written in my native tongue (French).

Après avoir fini la rédaction de ce manuscrit, j'ai ressenti le besoin de lui adjoindre quelques réflexions plus personnelles, éventuellement vaguement existentielles, quant à ma relation avec le monde de la recherche biomédicale. Aussi, une préface me sembla de bon aloi. En tout premier lieu, j'indique donc explicitement qu'il s'agit de remarques personnelles, en marge du corpus scientifique de ce manuscrit. C'est aussi en ce sens que je la rédige à la première personne du singulier, au lieu de l'utilisation plus traditionnelle du « nous » impersonnel universitaire.

Ces quelques remarques personnelles sont particulièrement motivées d'une part par mes interactions avec mes confrères cliniciens, d'autre part par de plus ou moins plaisantes discussions avec des individus d'extractions variées (oui comme c'est une préface j'ai décidé d'être snob). Il se trouve qu'étant, de par ma fonction d'Assistant Hospitalo-Universitaire en biostatistique, partie prenante d'une pétillante activité d'expertise méthodologique et de réalisation d'analyses de données au service de mes confrères d'un groupe hospitalier jouissant d'une certaine reconnaissance, je suis parfois interrogé, avec plus ou moins de bienveillance, quant à mon activité de recherche. Il en va de même lorsque je rencontre un nouvel acolyte de l'espèce humaine dans un cadre social quelconque. Étant un individu peu craintif (une bonne pâte quoi…), je suis toujours enclin à tenter d'expliquer en quoi ce que je fais me semble justifié.

Quel que soit le jugement porté par chacun de ces individus sur l'intérêt de ce que je fais, une remarque semble tout de même relativement constante. Il apparaît pour beaucoup de personnes que ce que je fais est catégorisé comme plutôt « fondamental ». Aussi, nombreuses ont été les personnes qui sont perplexes car elles semblent ressentir un décalage entre mes travaux de recherche et le fait que ma formation première est d'être médecin. Une première remarque s'impose : « fondamental » par rapport à quoi ? En effet, le laboratoire de manière générale ne se présente pas comme faisant des travaux fondamentaux, et ça n'est pas forcément la définition que j'aurai des miens. Tout numéricien qui a passé sa carrière à optimiser l'algorithme de Newton-Raphson afin qu'il converge en moins d'itérations, ou, plus abstrait, ceux qui sont réellement passionnés par dédier leur vie à toute théorie algébrique obscure se

gausseraient (ha ha, blague de statisticien…) de me voir prétendre que mes travaux puissent être qualifiés de « fondamentaux ». Néanmoins, admettons la prémisse que relativement à mes confrères cliniciens, mes travaux de recherche leur semblent se situer plus en amont que les leurs, et que pour tout un chacun j'ai vaguement l'air de « faire des maths ».

En général, cette constatation semble souvent être fortement corrélée avec deux interrogations plus ou moins manifestes ou latentes selon les individus : premièrement à quoi concrètement servent mes travaux immédiatement pour la « vraie » société, deuxièmement, est-il réellement nécessaire qu'une collectivité publique quelconque me verse sans sourciller un salaire (je précise que vu le contexte économique contraint actuel, je comprends que la question soit légitime) ? Ce qui amène à poser la question de l'utilité de mes travaux, ainsi que la question de mes réelles motivations à les conduire.

En ce qui concerne l'utilité, encore une fois ne considérant pas personnellement mener des travaux « fondamentaux » et travaillant dans le domaine de la santé, je pourrai me contenter d'apporter une réponse enfonçant des portes ouvertes : mes travaux, ceux du laboratoire, et l'ensemble des travaux menés sur la planète dans le domaine biomédical tentent de contribuer lentement, très lentement, mais surement à faire en sorte que l'on soigne mieux les gens, et que la population humaine soit globalement en meilleure santé. Maintenant, j'ai bien conscience que des travaux méthodologiques ne soient pas forcément ceux qui soient le plus facilement identifiables comme menant directement à une amélioration de la santé. Aussi, je suppose qu'il me serait demandé d'apporter des précisions. Ce que j'ai fait, avec plus ou moins de succès, à toute personne qui me l'a demandée. Ce que j'ai par ailleurs, avec un succès qui sera déterminé par le jury, tenter de faire dans les parties introductives de ce manuscrit de thèse, aussi je ne reviendrai pas sur cette question (en résumé, voir partie 1 et partie 2 de ce manuscrit).

Toutefois, continuons d'assumer la prémisse que mes travaux puissent être classés relativement à l'échelle de valeurs de certaines personnes comme « fondamentaux ». Quel est donc leur utilité contrairement à l'appliqué ou au translationnel ? J'avoue que si vraiment je considérais mes travaux comme tels, alors je dirai que leur unique utilité serait de proposer des idées, dans le sens de modèles ou de cadres théoriques, afin de tenter de résoudre des incohérences ou des manques d'une théorie préexistante (ce que peut être un micro chouïa, et de façon extrêmement modeste, le travail sur semantic primes et response shift). Et qu'à ce moment-là, en réalité, je me moquerais bien de déterminer leur utilité finale, au sens de contribution à la société (même si je pourrais en avoir une idée). Certes, je peux envisager que cette remarque interpelle (voir même qu'elle provoque un hérissement capillaire des personnes

ayant la responsabilité des agences de moyens qui sont enclines à ce que les dossiers de demande de financement contiennent absolument un encart « bénéfices attendus »), mais je considère que si on adhère un tant soit peu à l'épistémologie Kuhnienne, ça va tout de suite beaucoup mieux (oui je suis snob, mais aussi lâche, aussi je n'ai aucun scrupule à invoquer en joker ce cher Thomas Kuhn qui est fréquemment considéré comme un des plus grands épistémologues du vingtième siècle, démontrant ainsi mon inculture crasse en matière d'épistémologie fine).

J'invoquerais d'ailleurs plus spécifiquement le volet social de l'épistémologie de ce cher bon vieux Thomas. Si on considère que la vérité scientifique à un instant *t* ne peut être déterminée comme un absolu, mais comme un travail de consensus social au sein de la communauté scientifique, alors on se détend du slip quant à la nécessite de définir immédiatement l'utilité des travaux fondamentaux. Dans ce cadre, il me semble que l'on puisse assez sereinement envisager qu'un fondamentaliste (oui je surfe sur le « zeitgeist » de mon époque pour faire des blagues de mauvais goût) doit uniquement se préoccuper de proposer des idées abstraites originales, cohérentes, et respectant les critères d'un discours scientifique, dans le but de résoudre des incohérences théoriques ou expérimentales. Ce sera ensuite, lentement, décennies après décennies à l'ensemble de la société de déterminer l'éventuelle utilité ou non (c'est plus souvent non que oui pour un chercheur lambda comme moi d'ailleurs soyons honnêtes) de ces idées. Il me semble d'ailleurs qu'en étudiant avec une très grossière loupe l'histoire des sciences que des exemples récents puissent illustrer ces propos. Je ne crois pas qu'il fut demandé à l'ensemble des gens qui ont participé au début du siècle dernier à mettre au point la physique quantique de mette dans l'encart « bénéfices attendus » de leur hypothétique demande de financement : « permettra de faire de l'IRM » ou « sera responsable d'un tiers de l'économie mondiale du 21^{è} siècle ». Il me semble que ces personnes ont simplement voulu en premier lieux résoudre des incohérences entre observations et théorie.

Aussi, puisque je peux me permettre (je laisse à chacun la possibilité de juger de la puissance de cette bien maigre argumentation) de considérer qu'il n'est pas forcément nécessaire de faire de la recherche satisfaisant pourtant aux critères de fondement d'un discours scientifique en se posant la question de son utilité immédiate, quid donc des motivations d'un chercheur ?

En général, sur ce terrain, lorsque je tente malgré tout du mieux que je peux de justifier de la place de mes travaux dans le champ de la recherche en santé, et que la sauce prend, alors

la motivation semble en découler de fait : j'ai une envie irrépressible de contribuer au bonheur des gens sur fond de coucher de soleil en Technicolor.

C'est flatteur, ça semble altruiste, mais j'avoue qu'au fond de moi, je sens bien que mes motivations sont bien plus égoïstes. Aussi je profite de cet espace pour faire mon coming-out. Certes, je n'exclus pas que de multiples motivations non linéaires complexes m'animent, cependant j'en identifie une en particulier qui me tient à cœur : ça m'amuse. Pour des raisons qui tiennent bien évidemment de toutes les raisons possibles que l'on peut invoquer lorsque l'on utilise le modèle biopsychosocial de la santé, je suis un individu qui prend un plaisir certain à se poser des questions et à tenter de les résoudre aux moyens d'une certaine expertise scientifique. A vrai dire, j'invoquerais presque le déterminisme, voir la fatalité : je suis là où j'en suis car il me semble que c'est la seule chose que je sais faire vaguement (ce sera aux membres de mon jury de déterminer si je suis apte). Et au passage, je crois que beaucoup des personnes qui naviguent avec moi dans la même galère sont animées par cette motivation, même si aujourd'hui cela peut sembler presque indécent de le dire. D'ailleurs, je remercie la plupart des gens du laboratoire, notamment Véronique et Jean-Benoit, qui sont des gens pour qui cette motivation compte. Enfin, je ferai la remarque que si on tente de connecter cette motivation à mes élucubrations premières sur l'utilité dans le cadre de l'épistémologie Kuhnienne, alors on pourrait se rendre compte qu'il n'y a pas incompatibilité entre les deux, peut-être même envisager que c'est une assez bonne motivation à faire de la recherche, du moment que rigueur et honnêteté intellectuelle sont au rendez-vous.

Aussi, au final, je le dis haut et fort : oui ça n'est pas déplaisant d'imaginer que peut être je vais un peu contribuer à la recherche en santé, mais bon avant tout, ce manuscrit est la résultante de mon égoïste amusement. Et je compte bien continuer.

Table of contents

Table of tables

Table of figures

List of abbreviations

1PLM	1 Parameter Logistic Model
2PLM	2 Parameter Logistic Model
3PLM	3 Parameter Logistic Model
AIC	Akaike's Information Criterion
BIC	Bayesian Information Criterion
BP	Bodily Pain
CART	Classification and Regression Trees
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CTT	Classical Test Theory
df	Degrees of Freedom
DWLS	Diagonally Weighted Least Squares
EBM	Evidence-Based Medicine
EFA	Exploratory Factor Analysis
EQ-5D-3L	EuroQol-5D-3L
FDA	Food and Drug Administration
GH	General Health
GRE	Graduate Record Examination
HADS	Hospital Anxiety and Depression Scale
HEI-Q	Health Education Impact Questionnaire
HRQL	Health-Related Quality of Life
IC	Information Criterion
ICD	International Classification of Diseases
IQ	Intellectual Quotient
IRT	Item Response Theory
LLRA	Linear Logistic model with Relaxed Assumptions
LRT	Likelihood-Ratio Test
MCS	Mental Component Score
MH	Mental Health
ML	Maximum Likelihood
MLM	Robust Maximum Likelihood with Satorra-Bentler Correction
NHP	Nottingham Health Profile
OBI1	Overall Behavior Indicator 1
OBI2	Overall Behavior Indicator 2
OP	Oort's Procedure
PCOR	Patient-Centered Outcomes Research
PCS	Physical Component Score
PF	Physical Functioning
PF	Principal Factor
PGI	Patient-Generated Index
PRO	Patient-Reported Outcomes
QLDS	Quality of Life Depression Scale
QLQ-C30	European Organization for Research and Treatment of Cancer QLQ-C30
QoL	Quality of Life

QOLAP	Quality of Life Appraisal Profile
RCT	Randomized Control Trial
RE	Role Emotional
RMSEA	Root Mean Square Error of Approximation
ROSALI	RespOnse Shift detection Algorithm using Item response theory
RP	Role Physical
RS	Response Shift
SABIC	Sample size Adjusted Bayesian Information Criterion
SEIQoL	Schedule for the Evaluation of Individual Quality of Life
SEM	Structural Equation Modeling
SF	Social Functioning
SF-36	Medical Outcomes Study 36-Items Short Form
SI	international System of Units
SIP	Sickness Impact Profile
VAS	Visual Analogous Scale
VT	Vitality
WHO	World Health Organization

Part 1: Introduction

“After all, if you have made all these efforts to come back, it must be because you really love tests” - GlaDos, a fictional hysterical, yet sometimes whimsical, artificial intelligence

1. Scope

Currently, it is taken for granted the practice of medicine has to be based on coherent scientific theoretical knowledge leading to falsifiable and verifiable empirical proofs (i.e. *the hypothetical-deductive model* [1,2]). This necessity has been particularly emphasized since the middle of the 19th century. One pioneer of this approach is the physiologist Claude BERNARD who insisted on the need to prove medical knowledge with observable facts (which was the foundation of what was called *experimental medicine*) and was one of the first to propose blind experiments [3]. This need to support medical practice by empirical facts has been even more justified since the occurrence of the concept of *Evidence-Based Medicine* (EBM) which emerged in the 1960s and was formalized at the beginning of the 1990s [4,5].

EBM is an approach to medical practice intended to optimize decision-making by emphasizing the use of evidences from well designed and conducted research. In the EBM framework, evidences are classified by their epistemological strengths. In this framework, *meta-analyses* and well conducted *Randomized Controlled Trials* (RCTs) are considered as the strongest types of evidences to prove the value of a new therapeutic strategy [6].

In 1972, Archie COCHRANE published *Effectiveness and Efficiency: random reflections on health services* [7]. In this book, he described the lack of use of RCTs, which has led to the support of many medical practices only assumed to be effective. Use of RCTs has exponentially increased since they are considered as the methodological gold standard to prove the efficacy of new therapeutic strategies [8]. In parallel, the development of *epidemiology* has also emphasized the need to measure population's health state or to assess the effectiveness of public health programs (i.e. programs aiming healthy people and developed as primary or secondary level of prevention; e.g. systematic screening for breast cancer) [9]. Regarding sick individuals, on the opposite side, there has been recently a heavy focus on the idea to provide *"the right treatment, for the right person, at the right time"* which has led to the concept of *personalized* (or *stratified* or *precision*) *medicine* [10].

Thus, measuring *outcomes* in health-related research is of paramount importance as it is clearly established the evaluation of *efficacy* and/or *effectiveness* and/or *efficiency* of a new medical intervention has to be adequately assessed [11]. Therefore, the process of defining, measuring and assessing outcomes is one of the key methodological characteristics of empirical health-related research. For example, in RCTs, the difference in evolution over time of a particular outcome between the groups compared is the criterion which defines efficacy (which

is usually called the *primary criterion*, and the corresponding outcome is usually called the *primary endpoint*) [11].

Outcomes used in health-related research are most of the time measured and analyzed using quantitative methods (i.e. biostatistics). Some examples of outcomes traditionally used in RCTs are rate of mortality, survival time (especially in oncology) or the value of a biological substance (e.g. blood sugar level, serum creatinine level…) [11]. In epidemiology, we can cite routine morbidity outcomes such as incidence rate or prevalence rate [9]. These types of outcomes can be categorized as *objective* as they measure a state, a propriety or a characteristic of the human body considered only as an object with a presence in the physical reality [12]. These types of outcomes are relative to a conceived object (i.e. the human body). Measuring these types of outcomes can be a relevant process. A decrease in the rate of mortality after treating an infectious disease with a new antibiotic can of course be considered as a relevant outcome. In addition, some medical conditions are defined by out-of-range value of a biological constant (e.g. high blood pressure) and therefore, the related biological constant is a straightforward measure of the efficacy of a therapy.

Nonetheless, objective outcomes cannot always be solely used as criteria to assess medical interventions. Indeed, health is currently defined as "*a state of complete physical, mental and social well-being*" [13]. This definition recognizes a patient as a thinking subject and emphasizes the need to incorporate the assessment of patients' preferences, feelings, perceptions and experiences in health related research. These individual perceptions are not always correlated with objective outcomes [14]. Therefore, there was a will since the middle of the 20th century to use *subjective* outcomes (i.e. subjective as related to a thinking subject [12]) when assessing various medical interventions. In addition, in some medical areas or specialties, patient's speech is the only way to access to symptoms and health state (e.g. most of psychiatric diseases cannot be explored or assessed with objective outcomes). Thus, in the last decades, there was an increasing development and use of instruments designed to measure *subjective concepts* (in the form of *self-administered questionnaires*, *semi-structured* or *structured interviews*) [15]. In particular, it has led to the development of *Patient-Reported Outcomes* (PRO) which are designed to measure concepts such as depression level, anxiety level, pain, or fatigue using the perspective of the patient only (i.e. without any intervention of the expertise of a health-care professional [16]). In the recent years, at least in the countries with a high standard of living, there was a dramatic increase in chronically-ill patients [17]. Moreover, in some medical areas such as oncology, objective outcomes are now less relevant (new therapies

do not necessarily result in significant improvements in survival but rather focus on decreasing adverse effects) [18]. Thus, there was since the beginning of the 1980s an increasing interest in measuring *Quality of Life* (QoL) [18].

Instruments of measure of objective or subjective concepts share some similarities. Indeed, a measurement device, irrespective of the outcome (objective or subjective), has to comply to certain properties to be considered having a sufficient level of quality [12]. When measuring the weight of a human body (an objective outcome) with a weighing scale, it is expected the scale will provide a *valid* measure of weight and therefore will not measure for example height. If conditions do not change between two measurements (e.g. if the same person measure his weight twice in a very short period of time) it is expected the scale will be *reliable* and therefore will provide the same value twice. Finally, if conditions change (e.g. if someone eats a huge quantity of food), it is expected the scale will be *sensitive to change* with a sufficient precision.

These aforementioned properties (validity, reliability and sensitivity to change) are also expected when designing and using a PRO instrument measuring a subjective outcome [12]. The design and use of PRO instruments with good measurement properties is the concern of a field called *psychometrics*, which started at the end of the 19th century when psychologists became interested in measuring certain psychological phenomena [19]. Today, many PRO instruments have achieved a sufficient level of these properties [18]. Certain instruments designed to assess depression (a subjective state of mind) can be considered as having a higher level of reliability or sensitivity to change than a manual sphygmomanometer designed to measure blood pressure (an objective phenomenon) [20]. Certain instruments designed to assess QoL have met these aforementioned properties [18].

However, certain key characteristics differentiate objective concepts from subjective ones. A first fundamental difference can be the strength of the *nomological theories* between these concepts [12]. A lot of objective concepts used in physics (e.g. length, temperature, speed, mass, spin…) are intertwined in very heavily stabilized and formalized theories (i.e. the standard model of physics and quantum field theory) that were refined throughout thousands of years of scientific development [21]. Some objective concepts in medicine have also achieved a high level of theoretical stability. The diagnosis of the death of a human body as a biological state is a concept with shared standard definitions and methods [22]. Myocardial infarction is a very well-defined medical event directly linked to the rise in the value of a biological marker (i.e. serum troponin level) [23]. In a broader way, the definition of diseases is the concern of

the International Classification of Diseases which is at its 10th revision [24]. But, this level of strength of the nomological theories regarding subjective concepts has been rarely or is not achieved [12]. Depression as a disease may be a heavily stabilized concept [25]. Contrariwise, certain authors still consider QoL as a loosely-defined concept [18,26]. In a broader way, modern cognitive sciences (i.e. psychology, neurosciences, linguistics, artificial intelligence and philosophy of the mind) which are the sciences which primary deal with subjective concepts have achieved significant progresses mostly since the 20th century only [27].

A second fundamental difference concerns the definition of the *unit of measurement*. The objective physical reality can be described using seven basics physical quantities (or dimensions, i.e. length, mass, time, electric current, temperature, luminous intensity, amount of substance), with a unit of measurement associated to each (i.e. the International System of Units (SI) [28]). Other physical quantities are an algebraic combination of some the seven aforementioned (e.g. speed is length per time). The unit of measurement of each of these seven quantities is either defined relatively to a *standard* (e.g. the kilogram is defined by the International Prototype of the Kilogram located at the International Bureau of Weights and Measures), or to a *very stable physical phenomenon* (e.g. the second is defined by an exact value of periods of a certain frequency of radiations from the caesium atom), or to a *constant of the laws of the universe* (e.g. the meter is defined relatively to the speed of light in vacuum (*c*) which is assumed to be a constant of the universe). Whatever these definitions, they share a similarity which is the fact they defined the unit of measurement in relationships with abstract objects or concepts of the physical reality, unrelated to the phenomenological experience of a thinking subject. In contrast, defining the metric of a *subjective measure*[1] is often of a different nature. When the concept is very well-defined and assessed via structured interview by a well-trained health-care professional, the unit of measurement may approach something defined relatively to a standard (e.g. the assessment of a depressive syndrome by a trained psychiatrist) [25]. However, especially when using PRO, the metric of the concept one wants to measure is relative to the *internal perspective* of the patient (e.g. it is the patient that has to define what means a level of 3 of fatigue on a *Visual Analogous Scale* (VAS) [29]).

These two aforementioned key differences between objective and subjective measures or outcomes lead to a major consequence that can be summarized shortly: the process of measuring a subjective concept is generally more *dynamic* by nature [26]. The meaning of

[1] This metonymy will replace the more adequate expression « measure of a subjective phenomenon » for the sake of simplicity throughout the rest of the manuscript

subjective concepts can vary between people and within people over time (they can be *reconceptualized*). For example, the meaning of what is QoL can be very different for the same person at 15 or 80 years old. Moreover, the metric of these measures is also susceptible to change between and within individuals over time. The scale of the measure can be *recalibrated* in individuals' mind. These changes in the meaning of a concept over time are probably particularly susceptible to occur after a salient medical event (e.g. after a chemotherapy, it can be expected for one particular individual that "being moderately exhausted" does not correspond to the same level of fatigue than before) [30].

Therefore, admitting the dynamic nature of measuring subjective concepts and especially the fact their meaning can change within people over time leads to a major issue when assessing a medical intervention using subjective quantitative outcomes: is the difference in the measured concept assessed over time actually a true reflection of a change in the concept being assessed, or to a change in the meaning of the concept (or to both)? For example, after a cancer treatment, is a person reporting a change from 60 to 70 actually have a better QoL in itself, or is it its internal meaning of what is QoL that have changed?

In the 1980s and 1990s, the results of certain empirical studies using PRO as outcomes illustrated this dynamic nature of measuring subjective concepts. Indeed, it was found what were initially called *"paradoxical and counter-intuitive findings"* like reporting of stable HRQL by patients with a life-threatening disease [31], or discrepancies between clinical measures of health and patients' own evaluations of their health [32].

Since, it became apparent this dynamic nature of measuring subjective concepts had to be integrated into the use and interpretation of PRO instruments. Thus, this possible change of a meaning of subjective concepts over time has led to the development of what is now known as *response shift theory*.

In health-related research, response shift has been defined in 1999 as "*a change in the meaning of one's self evaluation of a target construct over time*" [30]. Initially, the term response shift described the effect on the value of the measure of a subjective concept because of a change in the meaning of this concept over time in a person's mind after the occurrence of a salient medical event [30]. It has since progressively involved to what the sociologist Robert K. MERTON calls a "*middle range theory*" (i.e. a theory starting via the observation of empirical phenomena (as opposed to a broad abstract entity) resulting in the process of generating general statements that can be verified by data [33]) aiming at explaining how the *psychological*

adaptation to chronic conditions or salient medical events can lead to a change in the meaning of subjective concepts and therefore having an effect on scores measured [34]. Thus, theoretical and methodological developments regarding detection, interpretation and integration of response shift phenomenon have become a prime concern in the field of PRO measurement since the beginning of the 2000s [34].

2. Main objectives and contribution of this thesis work

The broad objective of this PhD thesis is to investigate certain methodological and statistical considerations regarding response shift concept, detection and integration into PRO data analyses.

If response shift phenomenon has been proposed in the field of psychology since the late 1970s [35], this concept has been translated into health-related research since the beginning of the 2000s only (the first theoretical model was published in 1999 [30]). However, the amount of work conducted on response shift is substantial both in quantity and variety. Researches about response shift in PRO measurement is still a young field, generating various debates [36]. As it is not a field with heavily stabilized certainties there are still debate and controversies on the meaning of response shift phenomenon within the academic community. In addition, as it is still a young concept, it is not necessarily an "easy to grasp" concept at a first level. Thus, *a first objective* of this thesis work was to propose a state of the art of works conducted on response shift theory at an international level, with the intent of covering a broad variety of the many aspects of research on response shift. Thus, this first work covers the historical, theoretical, dialectical, methodological and empirical perspectives of researches on response shift theory.

On a more statistical and methodological level, numerous methods have been developed to detect and/or measure and/or integrate response shift when analyzing PRO based data. Initially, most of these methods were *design-based methods* (i.e. these approaches are based on a specific experimental design or on the use of specific measurement tools), but there was since the middle of the 2000s a shift in favor of *statistical-based methods* (i.e. methods relying on statistical modeling to analyze response shift after having collected data) [34]. One of the most attractive methods to detect response shift is *Oort's procedure* (OP), published in 2005 [37]. OP is based on *Structural Equation Modeling* (SEM), a statistical modeling technique for testing and estimating different types of causal relations using a combination of quantitative

data (i.e. covariance ± mean structures) and qualitative hypotheses [38]. OP allows detection of response shift without the need of a specific design [37]. Nonetheless, as it relies on statistical modeling, it is by definition a probabilistic method. If OP has been used on several clinical datasets [39–50], little is known regarding its statistical performances (for example its capacity to detect actual response shift). Moreover, certain methodological choices of the OP can be questioned. Therefore, *a second objective* of the thesis work was to provide for the first time an assessment of the performances of the procedure via a pilot simulation study.

When the concept of response shift phenomenon has emerged in psychology and health-related research, it was primarily viewed as a *measurement bias* adding complexity in the assessment of a subjective concept [51]. It is now more viewed as a result of the process of *adaptation to illness* over time [34]. Nonetheless, there are still debates about the interpretation of the occurrence of response shift phenomenon, especially in regards to the type of empirical design (RCTs or epidemiological studies) [36]. In addition, some authors have pointed out response shift can be a process not only related to psychological phenomena in individuals' mind, but also linked to the characteristics of the concept being measured or to the PRO instrument designed to measure it [29,52,53]. In the particular context of RCTs, it can be hypothesized response shift phenomenon is a process adding complexity in the measure of the subjective experience about the tested treatment, if the efficacy of this treatment is supposed to be a direct effect. Thus, it leads to the question, if and when potentially useful, could the occurrence of response shift phenomenon be prevented by using an appropriate PRO instrument? Therefore, the *third and last* objective of this thesis work was to provide a hypothesis which assumes response shift phenomenon can be linked with the semantic complexity of the concepts being assessed and the items used to assess it and therefore be, in part, prevented by designing a PRO instrument with the least possible semantic complexity.

3. Layout of the thesis

After this introduction, a second part will expose the context and the theoretical prerequisites that are the basis of the measurement of subjective concepts. In short, the notion of health, Patient-Centered Outcomes Research, Patient-Reported Outcomes and Quality of Life will be defined. Then the basics in psychometrics, especially regarding the different measurement models, will be presented.

The third, fourth and fifth part of this manuscript will be each dedicated to one of the three main aforementioned objectives of this thesis work.

Lastly, a discussion will deepen certain limits and perspectives of the two original research papers associated to the thesis (the fourth and fifth part, as the third part is a state of the art redacted as a book chapter) before a conclusion.

Part 2: Context and theoretical prerequisites

"A ruler wears a crown while the rest of us wear hats, but which would you rather have when it's raining?" - Barrin, *a fictional wise master-wizard*

This second part will introduce the context and theoretical prerequisites related to this thesis work. First, we will introduce the context. In the broadest way, this is the assessment of outcomes in health-related research. Especially, we will see how the current definition of health implies to encompass the internal perspective of the patient and therefore subjective aspects. Then, we will see how this requirement has led to the development of the concept of Patient-Centered Outcomes Research. After, we will see how Patient-Reported Outcomes and Health-Related Quality of Life measures are tools that can help to achieve this requirement of assessing subjective aspects of health. Lastly, we will present the theoretical underpinnings of subjective measures, which are the concerns of a field called psychometrics.

1. Context

1.1 The notion of health

The current definition of *health* was proposed by the World Health Organization (WHO) in 1948. It was emphasized health "*is not merely the absence of disease*". Indeed, health was rather defined as "*a state of complete physical, mental and social well-being*" [13]. We can see this current definition of health is *multidomain*. These domains can relate to objective phenomena (i.e. *objective* as relative to a conceived object), or to subjective states of mind (i.e. *subjective* as relative to a thinking subject) [54]. Thus, to not be healthy can be linked to the occurrence of objective abnormalities (physical or biological), but also to the perception a person has of his/her own functioning or to a social malfunctioning (e.g. a demoralizing absence of social interactions).

This distinction between these objective and subjective aspects of health is paramount when assessing the efficacy of a therapeutic strategy or the effectiveness of health policies. It shows the aforementioned assessment cannot be comprehensive enough with objective outcomes only (e.g. mortality, survival, clinimetrics). It has also to encompass the internal perspective of the patient, i.e. the subjective aspects of health.

Indeed, the individual perception of health and disease is not always correlated with objective outcomes such as health state or the severity of a medical condition [14]. As an example, if it seems straightforward to assume chronic peripheral artery disease is objectively a more severe condition than a benign tendinitis, the perceived health associated with one or the other condition can vary greatly from one person to another. Someone suffering from a benign tendinitis who deeply values sport as a major component of his/her well-being can

perceive his/her health as much more deteriorated than a very sedentary person suffering from chronic peripheral artery disease [55]. This simple example emphasizes the need to incorporate patients' preferences, feelings, perceptions and experiences when assessing the impact of medical conditions on health or when assessing treatment options.

Therefore, this need to focus on the subjective aspects of health has recently led to the development of the concept of *Patient-Centered Outcomes Research (PCOR)*.

1.2. The concept of Patient-Centered Outcomes Research

The Patient-Centered Outcomes Research Institute (PCORI) was established by law in 2010 in the United States of America, as part of the Patient Protection and Affordable Care Act [56]. Its goals are funding patient-centered comparative clinical effectiveness research and extending the concept of patient-centeredness from health care delivery to health-care research. PCOR focuses on *"patient-centeredness, which is determined by the extent to which the decision-making needs, preferences, and characteristics of patients are addressed in diverse settings of health care"* [57]. Indeed, *"individuals are not only distinguished by their biological variability: they also differ greatly in terms of how the disease affects their life"*, which promotes a focus towards the *"personome"*, i.e. *"the influence of the unique circumstances of the person"* [58]. Although it is well recognized that psychological, cultural, behavioral, environmental, and economic factors might importantly influence human health and disease, the integration of these features from the patient perspective into health care research remains a challenge. Therefore, the evaluation of questions and outcomes meaningful and important to patients and caregivers is emphasized by PCOR.

The PCORI includes both a board of governors and a Methodology Committee. This Methodology Committee is charged with developing methodological standards for PCOR. In 2012, this Methodology Committee published a methodology report [57]. Four general areas were identified in order to enhance methods for PCOR:

1. *"prioritizing research questions,*
2. *using appropriate study designs and analyses,*
3. *fostering efficient dissemination and implementation of results,*
4. *incorporating patient perspectives throughout the research continuum".*

In order to achieve the latter aforementioned area, the Methodology Committee emphasized the need to refine guidelines for the development, validation, use and interpretation of *Patient-Reported Outcomes* (PRO) [57].

1.3. What are Patient-Reported Outcomes?

PRO are instruments used to elicit information and collecting that information into some form of structured data [59]. Most of the time, they provide a mean of quantifying qualitative information, but the qualitative information can also be analyzed through the use of qualitative methods. As subjects are not always ill and therefore not always "patients", it is also sometimes suggested PRO could mean *Person-Reported Outcomes* [18].

Whatever the meaning, the goal of PRO instruments is to gather information from people themselves. They collect information directly from the patient without interpretations by clinicians or others (even if sometimes a proxy person is asked to fill the PRO when the person of interest cannot be asked (e.g. in pediatrics or for patients in a coma)). Thus, PRO should not be confused with clinical rating scales (where the clinician knowledge is used to rate disease severity or treatment effects) [59]. Therefore, they have been defined by the Food and Drug Administration (FDA) as *"a measurement of any aspect of a patient's health status that comes directly from the patient (i.e. without the interpretation of the patient's responses by a physician or anyone else)"* [16]. As such, most PRO are *self-administered questionnaires* (although some of them are semi-structured or structured interviews). They usually take the form of single- and multi-item measurement scales. An *item* is a question asked to the person filling the PRO. The most used response formats are *Likert-scale* (Figure 1) or *Visual-Analogous Scale* (VAS) (Figure 2).

Figure 1. Some items of the Medical Outcomes Study 36-Items Short Form (SF-36), with Likert-scale response format *(Source: WARE and SHERBOURNE, Medical Care, 1992 [60])*

	Yes, Limited a Lot	Yes, Limited a Little	No, Not limited at All
3. **Vigorous activities**, such as running, lifting heavy objects, participating in strenuous sports	[1]	[2]	[3]
4. **Moderate activities**, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	[1]	[2]	[3]
5. Lifting or carrying groceries	[1]	[2]	[3]
6. Climbing **several** flights of stairs	[1]	[2]	[3]
7. Climbing **one** flight of stairs	[1]	[2]	[3]
8. Bending, kneeling, or stooping	[1]	[2]	[3]
9. Walking **more than a mile**	[1]	[2]	[3]
10. Walking **several blocks**	[1]	[2]	[3]
11. Walking **one block**	[1]	[2]	[3]
12. Bathing or dressing yourself	[1]	[2]	[3]

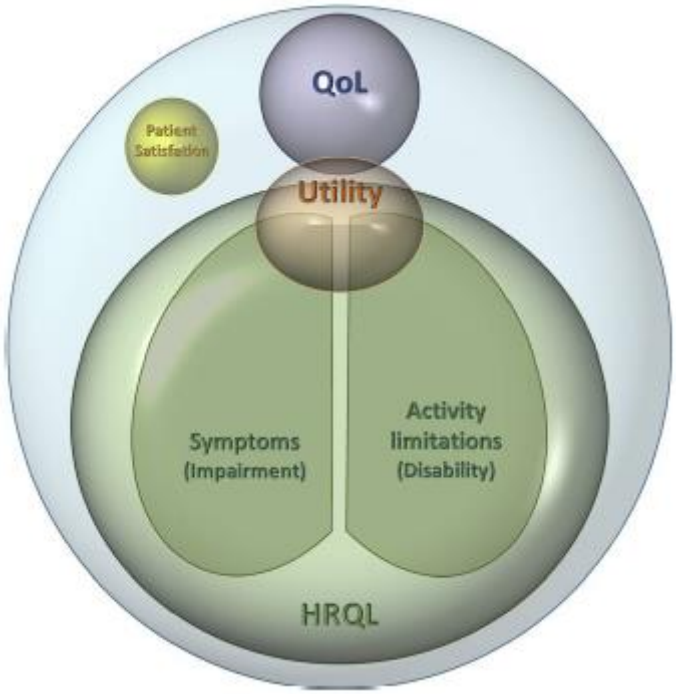
Figure 2. An item of the EuroQol (EQ)-5D-3L questionnaire, with VAS response format *(Source: RABIN and CHARRO, Annals of Medicine, 2001 [61])*

A PRO is not just an instrument used for gathering opinion. Rather, they are usually designed to measure a specific *concept.* The purpose of a PRO instrument is to map qualitative information onto a or several quantitative *scales* in order to measure the level of a concept [18,59]. Therefore, PRO are instruments that can help to measure subjective aspects of health. Thus, PRO are increasingly used in health-related research, whether as primary or secondary criteria in clinical trials, as indicators in epidemiological studies or for clinical practice [18]. Practitioners, researchers and policy makers are more and more interested by these *subjective measures* [18].

In health-related research, the range of concepts and outcomes that are covered by PRO is broad [59] (Figure 3). Some PRO are used to assess concepts related to impairment through the measurement of the level of symptoms. They cover domains like pain, fatigue, anxiety, depression, incontinence… An example of these instruments is the Hospital Anxiety and Depression Scale (HADS) which focuses on the measurement of the level of some anxiety and depressive symptoms [62]. Some are also used to assess disability or activity through the measurement of functioning. They cover domains like the activities of daily living. The Barthel Index of Disability is an example of these PRO [63].

Lastly, since the early 1980s, a lot of PRO are now designed to assess concepts that are named *Quality of Life* (QoL), or *Health-Related Quality of Life* (HRQL).

Figure 3. Examples of types of PRO currently used in health-related research, according to *MCKENNA, BMC Medicine, 2011 [59]*



1.4. The concept of Quality of Life or Health-Related Quality of Life

Currently, there isn't a universally accepted definition of what QoL is. Many definitions have been attempted, frequently emphasizing components of happiness and satisfaction with life [18]. However, some investigators argue that most people, at least in the Western world, are familiar with the expression "quality of life" and have an intuitive (although heterogeneous) understanding of what it comprises [18].

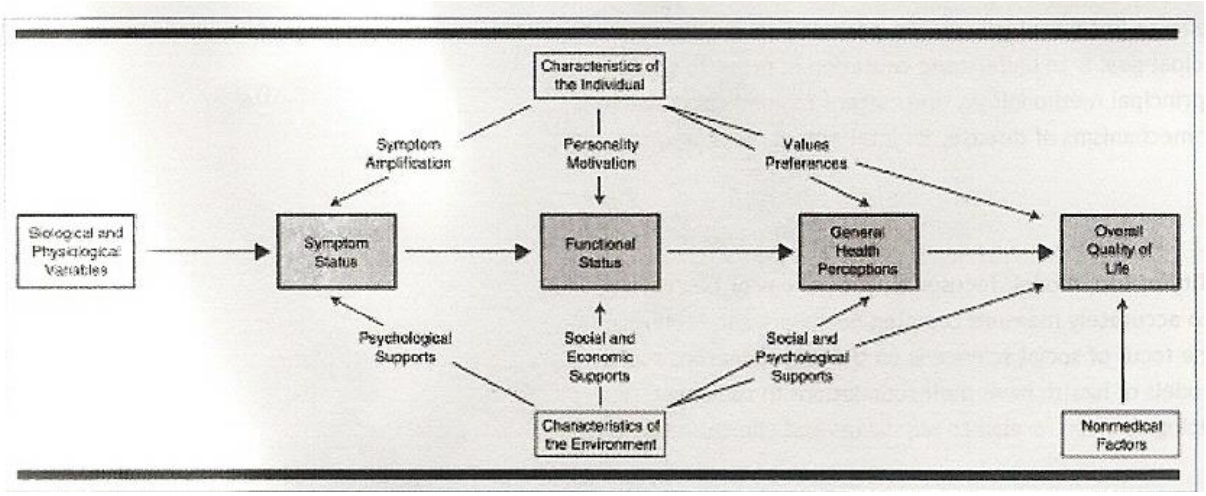
Some people equates QoL to personal *well-being*, but this view is not a consensus [64]. The concept of QoL also overlaps with the concept of PRO. Nonetheless, PRO is now frequently viewed as a broader concept: some PRO are designed to measure QoL, but some measure other concepts (e.g. personality, cognitive functioning…) [59]. QoL is also a term that can have different meaning depending on the area of application: it can relate to different characteristics from the view of an economist, a politician, a psychologist or a physician [18]. In the context of health-related research, people are rarely interested in assessing QoL in such a broad sense, and instead are focused only on the aspects of QoL that are affected by disease or treatment options. Thus, the term Health-Related Quality of Life (HRQL) is frequently used in order to remove ambiguity [18].

Although there is no current consensus, whether on the number of domains involved, or on the content of those domains [18,65], most researchers consider HRQL to be a *multidomain* concept: often including physical functioning, physical symptoms and toxicity, emotional functioning, cognitive functioning, role functioning, social well-being and functioning, sexual functioning and existential issues [18,65].

Several conceptual or theoretical models for HRQL have been proposed. None of them currently makes consensus [18]. Moreover, research about HRQL is still frequently performed without a clear reference to a definition of what the authors consider HRQL to be or to a conceptual model [65]. Indeed, in a systematic review performed in 2012, it was found on 148 papers about HRQL research the absence of any reference to a model in 48 studies, the use of a model specific only to the study conducted in 46 studies, and the use of a model referenced in at most four papers in 77 studies [65].

Nonetheless, some conceptual models for HRQL have some popularity. One of the most referenced is the *WILSON and CLEARY model*, published in 1995 [14] (Figure 4).

Figure 4. Conceptual model of relationships between patient outcomes, according to *WILSON and CLEARY, JAMA, 1995 [14]*



In this model, the outcomes that can be measured in health-related research are described as a continuum of increasing biological, social and psychological complexity. At one end of the continuum are biological and physiological variables such as serum creatinine level or hematocrit. The continuum progresses through more and more complex and integrated outcomes: from biological and physiological variables to symptoms status to functional status to general health perception to overall QoL [14]. General health perceptions (i.e. HRQL) represent an integration of all the previous health concepts and are a subjective rating [14]. Overall QoL integrates non-medical factors [14]. Characteristics of the individual (e.g. personality, values, preferences…) and characteristics of the environment (e.g. psychological, social and economic supports…) have an effect on the level of the different outcomes, thus leading to a subjective perceived general health perception and overall QoL [14]. This model was refined by FERRANS et al. in 2005 to better explicate the position of individual and environmental factors [66].

Other theoretical models for QoL have been proposed. In the *expectations model of CALMAN* [67], QoL is a measure of the difference between the hopes and expectations of a person and his/her present experience: i.e. a difference between perceived goals and actual goals. The Schedule for Evaluation of Individual Quality of Life (SEIQoL) [68] and the Patient Generated Index (PGI) [69] are two instruments using CALMAN's model as a conceptual basis. In the *HUNT and MCKENNA's needs model* [70], QoL is a measure of the ability and capacity of individuals to satisfy certain human needs (including such aspects as identity, status, self-esteem, affection, love, security, enjoyment, creativity, food, sleep, pain avoidance and activity). QoL is at its

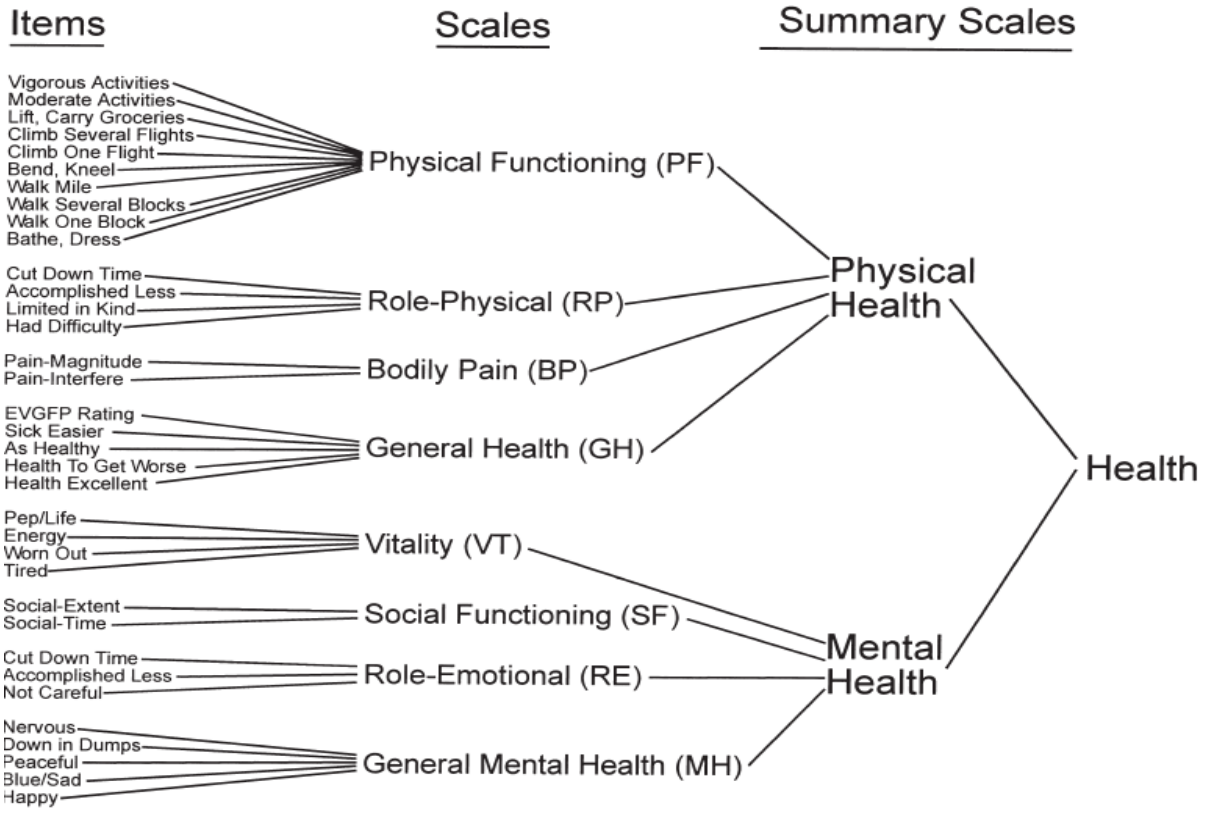
highest when a person fulfills all his/her needs and at its lowest when few needs are satisfied. The Quality of Life Depression Scale (QLDS), an instrument designed to measure QoL in the context of depression disorder is based on HUNT and MCKENNA's needs model [71].

Usually, a PRO measuring HRQL is composed of a combination of items dealing with aspects regarding impairment, disability and more subjective experiences about satisfaction with life, happiness or well-being [59]. Especially when assessing functioning, some items used in questionnaires for assessing QoL can be viewed as objective items (i.e. focusing on the human body as an object, e.g. a performance-based item like "what time do you need to walk up a flight of stairs?" or a perception-based item like "how often do you walk up stairs?") [29]. Other items, involving the process of rating an experience in comparison with an internal standard (i.e. evaluation-based item like "how difficult is it to walk up a flight of stairs?") are subjective ones [29].

One of the first questionnaire that is still sometimes described as a QoL instrument is the aforementioned Barthel Index of Disabillity [63]. Nonetheless, this instrument focuses on functional ability, physical functioning and activities of daily living. Therefore, it does not provide an adequate representation of patients' overall QoL. In the late 1970s and early 1980s, a second generation of questionnaires was developed. These instruments focused on physical functioning, physical and psychological symptoms, impact of illness, perceived distress and life satisfaction. Two examples of these instruments are the Sickness Impact Profile (SIP) [72] and the Nottingham Health Profile (NHP) [73]. Although these instruments are frequently described as QoL questionnaires, their authors did not claim them as such, as these instruments do not encompass some of the most subjective aspects of QoL (e.g. social well-being…). Therefore, they are viewed by some authors as general evaluation of health or health status [18].

One the most widely used HRQL instrument is the Medical Outcomes Study 36-Items Short Form (SF-36), developed in the early 1990s [60]. It is composed of 36 items. According to the SF-36, HRQL can be divided into two main subdomains at a first level [74]: a physical one (PCS: Physical Component Score) and a mental one (MCS: Mental Component Score). Each of these subdomain can be further divided into four subdomains at a second level. The PCS encompasses Physical Functioning (PF), Role Physical (RP), Bodily Pain (BP) and General Health (GH). The MCS encompasses Vitality (VT), Social Functioning (SF), Role Emotional (RE), and general Mental Health (MH) (Figure 5).

Figure 5. The structure of the SF-36 *(Source: KELLER et al., Journal of Clinical Epidemiology, 1998 [74])*



The aforementioned instruments are intended for general use, irrespective of the illness or condition of patients. They may also be used with healthy people. Some of them were developed to survey populations. They are *generic questionnaires* to measure HRQL [18]. As they intend to cover a wide range of conditions, they can be used to compare scores between medical conditions or against the general population. Nonetheless, they can fail to cover very specific aspects of the impact of a particular disease on one's HRQL [18]. Thus, a lot of *disease-specific* questionnaires have been developed. An example of one of the most used disease-specific questionnaire is the European Organization for Research and Treatment of Cancer QLQ-C30 (EORTC QLQ-C30) which focuses on cancer specific aspects.

HRQL measures are increasingly used in health-related research, whether as criteria of interest in comparative research (randomized control trials or others), as epidemiological indicators in population surveys, or for clinical practice [18]. Indeed, as aforementioned, these types of measures allow to capture subjective aspects about perceived health state level. A high level of HRQL has been identified as a goal for all people across all life stages by leading health organizations [75,76]. Thus, some researchers currently ague HRQL should be as often as possible one of the primary criteria of interest in health-care research [77].

Therefore, now that we have established the importance of the use of PRO and especially of the concept of HRQL in health-care research, we will further provide the necessary background to understand how it can be adequately measured. The scientific underpinnings of subjective measures are the concern of a field called *psychometrics*.

2. Theoretical prerequisites in psychometrics

2.1. Generalities

Psychometrics is the study of the measurement of psychological phenomena like abilities, skills, intellectual performances, personality or knowledge [19]. Its origin can be traced at the end of the 19th century, when psychologists became interested in quantifying psychological phenomena. One of the first concerns of the psychometric field was the search of a way to measure intelligence which led to the development of the Intelligence Quotient (IQ) [78]. Francis GALTON (1822-1911), James CATTEL (1860-1940), Charles SPEARMAN (1863-1975), Louis THURSTONE (1887-1955), Lee CRONBACH (1916-2001) are some of the pioneers of the discipline, which has strong ties with statistics and biostatistics. As researchers became increasingly interested in assessing the subjective aspects of health, the psychometric methods were progressively translated in the field of health-related research since the middle of the 20th century. This translation allowed the use and measurement of PRO.

In the psychometric field, a concept we want to measure is called a *construct* (e.g. anxiety, pain, fatigue, HRQL…) [79]. As for any measurement device, the metrological properties of a questionnaire have to be evaluated. There are three main properties that are needed to be achieved:

1. the measure has to be valid (*validity*: it has to measure what it is supposed to measure, e.g. if a questionnaire is designed to measure anxiety level, it has to not measure depression level);
2. the measure has to be reliable (*reliability*: a measure is said to have a high reliability if it produces similar results under consistent conditions. Measures that are reliable are accurate, reproducible, and consistent from one testing occasion to another);
3. the measure has to be *sensitive to change* (if conditions do change between two measurements, a measure is said to be sensitive to change if it captures the change with a sufficient level of precision) [12].

A PRO instrument has good psychometric properties only if these conditions are met. The assessment of the psychometric properties of a questionnaire is the concern of the validation phase of an instrument. It relies on statistical techniques. One key aspect of these statistical techniques is the choice of a *measurement model*. We will further develop some different measurement models in psychometrics.

2.2 The measurement models in psychometrics

The measurement model is the algebraic process used to transform the observed responses to all the items of a questionnaire into a unique numeric value: the measure of the construct [12]. Since the end of the 19th century, different theories have been developed, describing the relationships between the items and the measure of the construct. The first and most used measurement model is the *Classical Test Theory* (CTT).

2.2.a. The Classical Test Theory

Under this theory, the measure is a variable which is called the *score*. The score is taken to be an appropriate representation of the level of a construct we want to measure. The score is a combination of the responses to all items. Usually, this combination is simply the sum of the responses to the items (sometimes with different weights applied to each item).

This theory postulates the *observed score* (*S*) for an individual to each item is an assessment of the *true score* (*T*) of this individual. The observed score can be decomposed into two terms, the true score (*T*) and a *measurement error* (*E*) [80–82]:

$$S = T + E.$$

The expected value of the observed score is the true score:

$$E(S) = T.$$

Thus, the true score of an individual *i* is conceptualized as the mean of observed scores of a repeatedly administered test, with the mean of measurement errors on the *J* recurrences supposed to be equal to 0:

$$t_i = \frac{\sum_{j=1}^{J} s_{ij}}{J} \text{ with } \frac{\sum_{j=1}^{J} e_{ij}}{J} = 0 \text{ and } j = 1 \; to \; J.$$

Under CTT, it is assumed measurement errors have to be independent in all circumstances. Formally stated, it corresponds to three postulates:

1. the true score is not correlated to measurement error ($\rho_{T,E} = 0$),
2. measurement errors between two items of the same scale are uncorrelated ($\rho_{E_1,E_2} = 0$) (i.e. assumption of *local independence*),
3. measurement error between one item and true score of another item are uncorrelated ($\rho_{E_1,T_2} = 0$).

At an individual level, the precision of a score is the error variance of the recurrences [83].

At a population level, the true score is a random variable with a variance $\sigma^2(T)$. At this level, the precision of a test is defined by a *reliability index*, which is the variance of the true score divided by the variance of the observed score:

$$\rho_{s,t}^2 = \frac{\sigma^2(T)}{\sigma^2(S)}.$$

As measurement error is assumed to be independent of true score, the observed score variance can be written as:

$$\sigma^2(S) = \sigma^2(T) + \sigma^2(E).$$

Thus, the more the error variance decreases, the more the reliability index of the test increases [84]. As $\sigma^2(T)$ cannot be empirically known, the reliability index has to be estimated. The most used estimator is *Cronbach's alpha* [85].

If CTT is still the most used measurement model, it has been heavily criticized [86,87]. One of the criticisms concerns the postulates of independence about measurement error which are strict. In practice, whether they are met is rarely investigated. In addition, these postulates do not allow adequately taking into account certain systematic error phenomena, like learning when a questionnaire is repeatedly administered to the same person. The justification of the use of a questionnaire relies heavily on the value of its Cronbach's alpha. However, it is well-known the value of the Cronbach's alpha has a tendency to increase with the number of items in a questionnaire [12]. Therefore, under CTT, using a long questionnaire can be easily justified. Another criticism concerns the obtained score. Under CTT, the raw score can rarely achieve *interval scale property* (i.e. a unit difference characterizes the same amount of change in the concept measured whatever the initial position on the scale) [86]. A last criticism concerns the

definition of the true score under CTT. Indeed, in CTT the true score is defined as the mean of observed scores of a repeatedly administered test to an individual. Thus, the true score is defined by the items used to estimate observed score. The true score is synonymous of the operationalization of the measure. Therefore, different questionnaires designed to assess HRQL would indeed measure different HRQL concepts, each defined by the items used to measure it [88].

2.2.b. Models with latent variables

2.2.b.a What is a latent variable?

The broadest definition of what is a *latent variable* comes from the field of probability theory and statistics. In this field, a latent variable is defined as a variable "*for which there is no sample realization for at least some observations in a given sample*" [89]. A more restrictive and pragmatic definition of what is a latent variable can be found in social and behavioral sciences. Here, the notion of latent variable refers to a "theoretical" or "hypothetical" construct. Thus, latent variables are the abstract and unobservable concepts specified in theories [89].

Therefore, in psychometrics, a latent variable can be the construct someone wants to measure (e.g. anxiety, depression, HRQL…). This variable cannot be observed or measured directly, as opposed to *manifest variables* which are the items of a questionnaire. Since the beginning of the 20th century, several models incorporating the use of latent variables have been developed. One of these model is the "*common factor model*". We will now elaborate expansively on this model as it is the model used in the simulation study of this thesis work (Part 4).

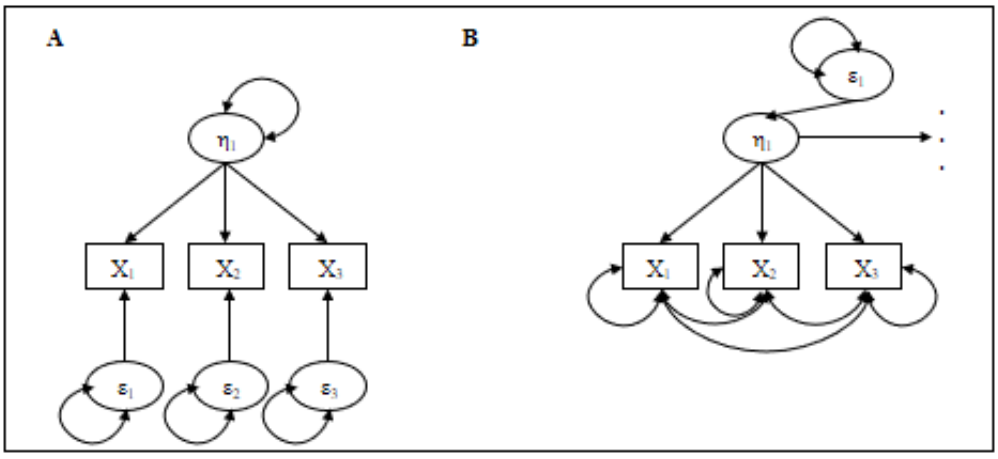
2.2.b.b. The common factor model

The work of Charles SPEARMAN on *factor analysis* at the very beginning of the 20th century (1904) was the starting point of the development of models with latent variables [90,91]. Later, Louis THURSTONE formalized the "common factor model" in 1947 [92]. In this model, a latent variable is called a *common factor* (e.g. HRQL). The common factor is an unobservable hypothetical construct that is assumed to explain the relationships between the manifest variables, which are called the *indicators* (i.e. the items). Thus, each subject is supposed to have a certain level of the common factor, and his/her responses to the items are supposed to be *indicative* of this level. The responses are supposed to *reflect* the level of the

common factor, as it is the common factor that is assumed to explain why a subject have responded in a particular way.

This aforementioned model, where the common factor is assumed to explain the relationships between the indicators, is called a *reflective model*. In this model, this is the common factor which is supposed to cause the responses to the indicators and therefore the indicators are taken to reflect the true value of the common factor. Nonetheless, another type of relationships between indicators and common factor can be used: a *formative model* [93]. In a formative model, these are the indicators that are assumed to cause the value of the common factor, i.e. they are supposed to *form* the common factor. An example of this type of model can be justified when the common factor is *an aggregate* of a collection of several indicators measuring something somewhat similar although their values are not assumed to be caused by a higher-order process (e.g. the measure of *a composite index* like the Human Development Index which is computed using indicators of life expectancy, level of education, income per capita etc. can be thought as a formative model). In psychometrics, models are usually reflective, as the common factor is generally assumed to cause the relationships between the indicators. Moreover, it can be argued the formative model can be used to describe relationships between manifest variables and common factor in specific context, but it is not a measurement model *per se*. Thus, from here to the end of this manuscript, *we will assume the common factor model as a reflective model*. Figure 6 displays path diagrams of a reflective and formative model.

Figure 6. A graphical representation of A. A reflective common factor model, B. A formative model



Note: This figure uses path diagram representation of common factor model derived from Structural Equation Modeling. I.f the reader is not familiar with these kinds of graphical representations, it will be introduced later in the same chapter of this manuscript (see Figure 7). **B:** The arrow from η_1 leading to three dots represents the need to have at least one indicator (direct or indirect) of $\eta 1$ for the model to be identified.

In the reflective model, each indicator is a linear function of one or several common factors and a *unique factor* (or residual factor). When there is only one common factor involved, the construct is said to be *unidimensional* (a unique latent variable is enough to explain all the correlations between the indicators, or, formally stated, if the latent variable is maintained at a fixed level, then all the indicators are independent [94]), otherwise the construct is said to be *multidimensional*. Common factors are continuous variables and explain the level of several indicators. The unique factor, also a continuous variable, is associated with one indicator only. It represents both the variability of the indicator that cannot be explained by common factors and measurement error.

The model can be written as:

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

The subscript j is for the indicators (i.e. the items, j = 1 to p) and the subscript m for common factors (m = 1 to M). y_j is the vector of responses of N subjects to the item j. Each common factor F_m is a vector of the factor level of N subjects. Each common factor loads (i.e. explains) on each indicator. Thus, the strength of the correlation between a common factor and an indicator is represented by a coefficient: the loading λ_{jm} which has an absolute value between zero and one. The unique factor to each item is a vector of length N (ε_j). A unique factor associated to an indicator is assumed to be independent of the unique factors associated to the other items (i.e. assumption of local independence) and independent of common factors. A unique factor is assumed to follow a Gaussian distribution with zero mean.

Initially, factor analysis was used in an exploratory manner (*EFA: Exploratory Factor Analysis*). The purpose of EFA is to search for the number and meaning of latent variables which can explain the variability and correlations of a set of manifest variables [90]. Here, the meaning of latent variables comes from data analyses [95]. The methods used to estimate the parameters of an EFA model on sample data are *Principal Factor* (PF) or *Maximum Likelihood* (ML). Because of the indeterminate nature of the common factor model (i.e. for any given multiple-factor model, there exist an infinite number of equally good-fitting solutions, each represented by a different factor loading matrix), once the appropriate number of factors has been determined, the extracted factors are generally rotated [90]. In psychometrics, the usual goal of the rotation is to propose the most readily interpretable solution in which each factor is defined by a subset of indicators that load highly on one factor, and each indicator has ideally

a high loading on one factor (the *primary loading*) only, and has close to zero loading on the remaining factors (no *cross-loadings*) [92]. Thus, the most commonly used type of rotation is called *varimax* rotation which maximizes the sum of the variances of the squared loadings (i.e. the proportion of variance in the indicator the factor solution explains) [90]. It is an orthogonal rotation: it implies the common factors are uncorrelated.

Later, *Confirmatory Factor Analysis* (CFA) has been developed and was formalized by Karl Gustav JÖRESKOG (1935-present) in 1969 [96]. In CFA, the number and meaning of the latent variables are defined prior to data analyses [95]. Here, the goal of the statistical analysis is to check if the model fits adequately the data. Thus, CFA became an important statistical tool in assessing the *structural validity* of a questionnaire and modeling responses to items [90].

There are several key differences between EFA and CFA [90]. First, in CFA, the model specifies which common factor(s) loads on which item(s) (whereas in EFA, each common factor loads on each item). Usually, an item is linked to one common factor only (there are no cross-loading). Then, if in EFA the unique factors are supposed to be independent to each other (before rotation, and after if an orthogonal rotation is used), it is possible in CFA to relax this assumption and to specify in a model some correlations between some unique factors (i.e. relaxing the assumption of local independence). Lastly, in CFA the common factors are usually assumed to be correlated to each other, while in EFA they are assumed to be uncorrelated when using an orthogonal rotation such as varimax.

A CFA model for the observed items of an arbitrary subject *i* can be written as:

$$y_i = \Lambda\eta_i + \varepsilon_i,$$

where y_i is a vector of observed items, η_i is a vector of unobserved common factors, ε_i is a vector of unique (or residual) factors and matrix Λ contains factor loadings [96]. Using sample data, the parameters of the model are usually estimated using ML (which assumes multivariate normality [97]). The goal of the fitting algorithm is to provide a set of parameter values which leads to the best possible prediction of the *observed variance-covariance matrix* between the items (i.e. the model predicts an *expected variance-covariance matrix* as close as possible to the observed one). If the model is just-identified or over-identified (i.e. the degrees of freedom (*df*) of the model are null or positive), then there is a *unique solution* of parameter values which minimizes this *fit function*:

$$F_{ML} = \ln|S| - \ln|\Sigma| + trace[(S)(\Sigma^{-1})] - p\,,$$

where $|S|$ is the determinant of the observed variance-covariance matrix, $|\Sigma|$ is the determinant of the expected variance-covariance matrix, *p* is the number of indicators [90]. The determinant is a scalar that reflects a generalized measure of the covariance between the entire set of variables contained in a matrix. The trace of a matrix is the sum of values on the diagonal (which is the sum of variances for a variance-covariance matrix). Thus, the goal of the fitting algorithm is to minimize the differences between these matrix summaries (i.e. the determinant and trace) for *S* and Σ. When there is perfect fit, the value of the fitting function is zero.

To achieve identification of the model, the metric of latent variable(s) must be scaled. It requires the imposition of what is called *identifiability constraints*. This can be done in two different ways: either a factor loading for each latent variable is constrained to be equal to one, or the variance of the latent variable is set to be equal to one [90].

The quality of the fit of an estimated CFA model on sample data can be assessed using fit indices [98]. The first developed fit index is the χ^2 which is calculated as:

$$\chi^2 = F_{ML}(N-1),$$

where *N* is the sample size. Under the null hypothesis $S = \Sigma$, this quantity is supposed to follow a χ^2 distribution with a known number of degrees of freedom. Thus, the fit of the model can be evaluated via statistical hypothesis testing. However, it is currently rarely used in applied research [90]. Indeed, as the *power* of the test increases with sample size, it leads to the rejection of adequate fit even when differences between *S* and Σ are negligible when sample size is large [90]. Thus, other fit indices have been developed. One popular fit index is the *Root Mean Square Error of Approximation* (RMSEA). The RMSEA is based on the idea that although $F_{ML}(N-1)$ asymptotically follows a central χ^2 distribution under the null hypothesis of perfect fit, it asymptotically follows a noncentral χ^2 distribution under the alternate hypothesis. Thus, the noncentrality parameter (*d*) of this distribution depends on how badly the model fits, and can be used to construct a fit index [98]. This *d* parameter can be estimated as:

$$d = max(\chi^2 - df, 0)/(N-1), \text{ and } RMSEA = \sqrt{d/df}.$$

Thus, the closer the RMSEA is to zero, the better the fit is. Used cut-off values can be: RMSEA < 0.05 is an excellent fit, RMSEA between 0.05 and 0.08 is an average fit and RMSEA > 0.08 is a mediocre fit [98]. Another popular fit index is the *Comparative Fit Index* (CFI) which is a measure of how the current model improves the fit relatively to a "null" model (a model assuming no covariance between the manifest variables) [98]. It is computed as:

$$CFI = \frac{\max(\chi_0^2 - df_0, 0) - \max(\chi_k^2 - df_k, 0)}{\max(\chi_0^2 - df_0, 0)},$$

with the subscript *0* for the "null" model and the subscript *k* for the tested model. The CFI can range from zero to one and the closer the value is to one, the better the fit is. Used cut-off value can be CFI > 0.95 as a criterion for good fit [98].

So, when assessing the structural validity of a questionnaire using the estimation of a CFA model on sample data, a questionnaire with adequate structural validity will be one with:

- sufficiently high factor loadings value for each item (e.g. factor loadings > 0.4),
- adequate fit according to fit indices [90].

A CFA model can be expanded to include more than the fit of the observed variance-covariance matrix between the manifest variables (i.e. the items). Indeed, it can also model the vector of means of the manifest variables (which is called adding a *mean structure* to the model) [90]. In this case, there is an additional set of parameters that are estimated which are called the *intercepts* and the corresponding CFA model for an arbitrary subject *i* can be written as:

$$y_i = \tau + \Lambda\eta_i + \varepsilon_i,$$

with τ a vector of intercepts (one for each indicator) [90]. Thus, an intercept corresponds to the value of an indicator when the common factor is equal to zero. As for the modeling of covariance structure, identifiability constraints are required to achieve identification. Here, either an intercept per common factor has to be constrained to be equal to zero, or the common factor(s) mean(s) has to be set to be equal to zero [90].

As ML assumes multivariate normality, it requires the indicators to be continuous variables, or, at least, categorical variables with a sufficient number of categories (at least seven) [99]. Nonetheless, several estimators have been developed to accommodate the use of categorical items. One effective and simple option is to use *robust Maximum-Likelihood* with *Satorra-Bentler correction* (MLM) [100]. Here, the χ^2 value of the model is corrected by a scaling correction factor (*c*) which is a general measure of the degree of multivariate kurtosis in the data. The standard errors are also corrected using a sandwich-type estimator. Thus, MLM produces adequate values of $\chi^2$and fit indices. It also adequately corrects the standard errors associated with the parameters.

Nonetheless, when there is a large departure from multivariate normality (e.g. when the number of categories is ≤ 4), it is possible for the parameter values to be biased [99]. Thus, other estimators are available. One popular option is called the "*three-stages approach*" [101].

Its fundamental assumption is the idea a categorical indicator is representative of a *latent continuous indicator (y*)* following a Gaussian distribution that has been truncated [99,102]. In a first step, some parameters called *thresholds (δ)* are estimated. These thresholds are the parameters defining how the categories of the manifest ordinal indicator and the latent continuous indicator are related. These thresholds are the values of the latent continuous indicator at which there is a switch from one category to another of the observed categorical indicator (thus, for an item with *k* categories, there are *k-1* thresholds to estimate). If *x* is the manifest indicator, in general, with *k* categories:

$$x = i \text{ if } \delta_{i-1} < y* < \delta_i, \text{where } \delta_0 \to -\infty, \text{and } \delta_k \to +\infty.$$

Within a CFA model, estimates of the thresholds can be obtained as:

$$\widehat{\delta_i} = \Phi^{-1}(p_1 + p_2 + \cdots + p_i), \qquad i = 1, \dots, k-1,$$

with $\Phi^{-1}$the inverse of the standard normal distribution function and $p_i$the estimated percentage of responses in category *i* [103].

A second step is the use of these thresholds for the estimation of *polychoric correlations* between the items (an estimation of the correlations between the items as if they were taken to be continuous). Let $\delta_1^{(1)}, \delta_2^{(1)}, \dots, \delta_{k-1}^{(1)}$ be the thresholds for variable y_1^* and let $\delta_1^{(2)}, \delta_2^{(2)}, \dots, \delta_{k-1}^{(2)}$ be the thresholds for variable y_2^*. The polychoric correlations can be estimated by maximizing the log-likelihood of the multinomial distribution [103]:

$$\ln(L) = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} n_{ij} \log(\pi_{ij}(\theta)),$$

where $\pi_{ij}(\theta) = \Pr(y_1^* = i, y_2^* = j) = \int_{\delta_{1-1}^{(1)}}^{\delta_1^{(1)}} \int_{\delta_{j-1}^{(1)}}^{\delta_j^{(2)}} \Phi_2(u,v) du dv,$

with $\Phi_2(u,v) = \frac{1}{2\pi\sqrt{(1-\rho^2)}} e^{-\frac{1}{2(1-\rho^2)}(u^2 - 2\rho uv + v^2)}$ the standard bivariate normal density with correlation ρ, θ a parameter vector $\theta = (\delta_1^{(1)}, \delta_2^{(1)}, \dots, \delta_{k-1}^{(1)}, \delta_1^{(2)}, \delta_2^{(2)}, \dots, \delta_{k-1}^{(2)}, \rho)$ and n_{ij} the sample size for categories *i* and *j* of the two variables.

Finally (third step), the estimated polychoric correlations and thresholds are used as data to estimate the model using a *Diagonally Weighted Least Squares* (DWLS) estimator [101]. Parameter values of the CFA model are found by minimizing this fit function [99]:

$$F_{DWLS} = (r - \hat{\rho})' W^{-1} (r - \hat{\rho}),$$

where r and $\hat{\rho}$ are vector containing sample-based and model-based relevant statistics (thresholds and correlations) and W is a weight matrix (a consistent estimator of the asymptotic variance-covariance matrix of the sample statistics; but when performing DWLS, only the diagonal of the matrix is used (variances) to decrease the probability of computational and convergence issues [99]).

This method imposes more identifiability constraints (the metric of the latent continuous indicators must be scaled), three parameterizations are usually used [102,103]:

1. the *delta (or marginal) parameterization* where the distribution of y^* is standardized,
2. the *theta (or conditional) parameterization* where the latent continuous indicator error variance is constrained to be equal to one,
3. the *alternative parameterization* (which requires an indicator with at least three categories) where two thresholds are constrained to be equal to zero and one.

This method can be computationally intensive on large models and sometimes leads to convergence issues [102].

CFA is currently thought as a special case of what is called *Structural Equation Modeling* (SEM) [104]. SEM is a very wide class of statistical modeling techniques for testing and estimating different types of causal relations using a combination of quantitative data (i.e. covariance ± mean structures) and qualitative hypotheses [38]. The early beginnings of SEM can be found in the work of Sewall WRIGHT (1889-1988) on *path analysis* in 1921 [105]. Path analysis is a method used to describe the directed dependencies among a set of manifest variables. It can be understood as a generalization of regression analysis. In regression analysis, a set of independent variables (which are in SEM language the *exogenous* variables) are used to explain the value of one dependent variable (which is in SEM language the *endogenous* variable). Path analysis allows modeling more complex directed relationships among a set of variables than regression analysis (e.g. a dependent variable can also explain the value of another variable, an independent variable can explain the value of more than one dependent variable…). Sewall WRIGHT formalized the rules of tracing directed relationships among a set of manifest variables, and a method to estimate the path coefficients (which are the coefficients estimating the strength of a relationship between two variables that are thought to be linked in the path model) [106]. The path coefficients are derived from the analysis of the observed variance-covariance matrix between the set of variables.

In the early 1970s, especially with the work of Karl Gustav JÖRESKOG, path analysis and the common factor model were unified under the SEM theory (this was called by JÖRESKOG as the LInear Structural RELationships (LISREL) model) [107]. Since, SEM can accommodate the use of latent variables.

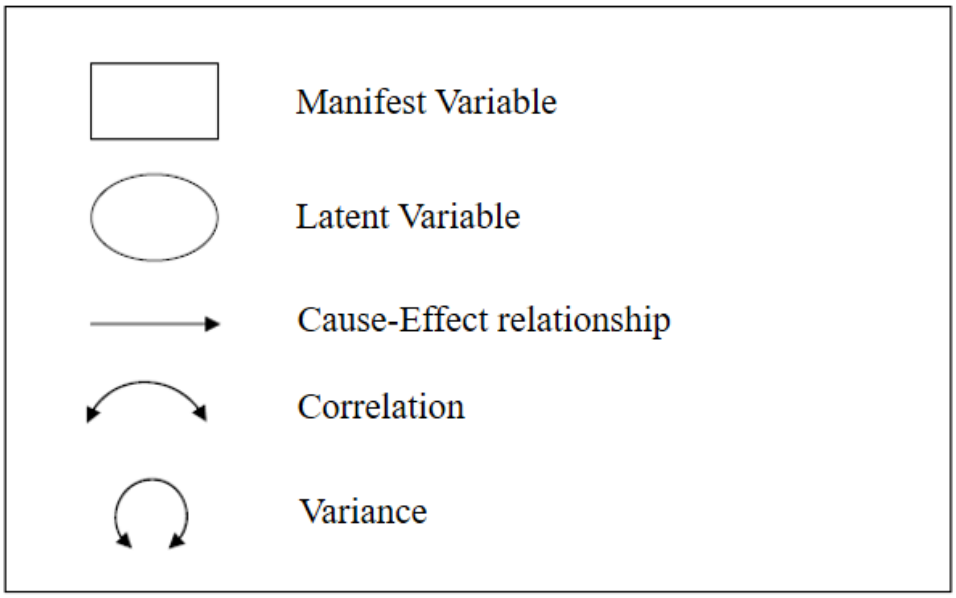
Currently, a comprehensive SEM is usually described as a compound of two main parts:

1. one or several *measurement model(s)*, which is/are the part of the SEM describing how latent variable(s) is/are measured by manifest indicators;
2. a *structural model*, which is the part of the SEM describing the structural relationships between latent and/or some manifest variables [38].

Therefore, a CFA model is a special case of SEM. It is a SEM restricted to a measurement model only [90].

SEM are usually displayed using graphical representations (which are called *path diagrams*) [108] (Figure 7).

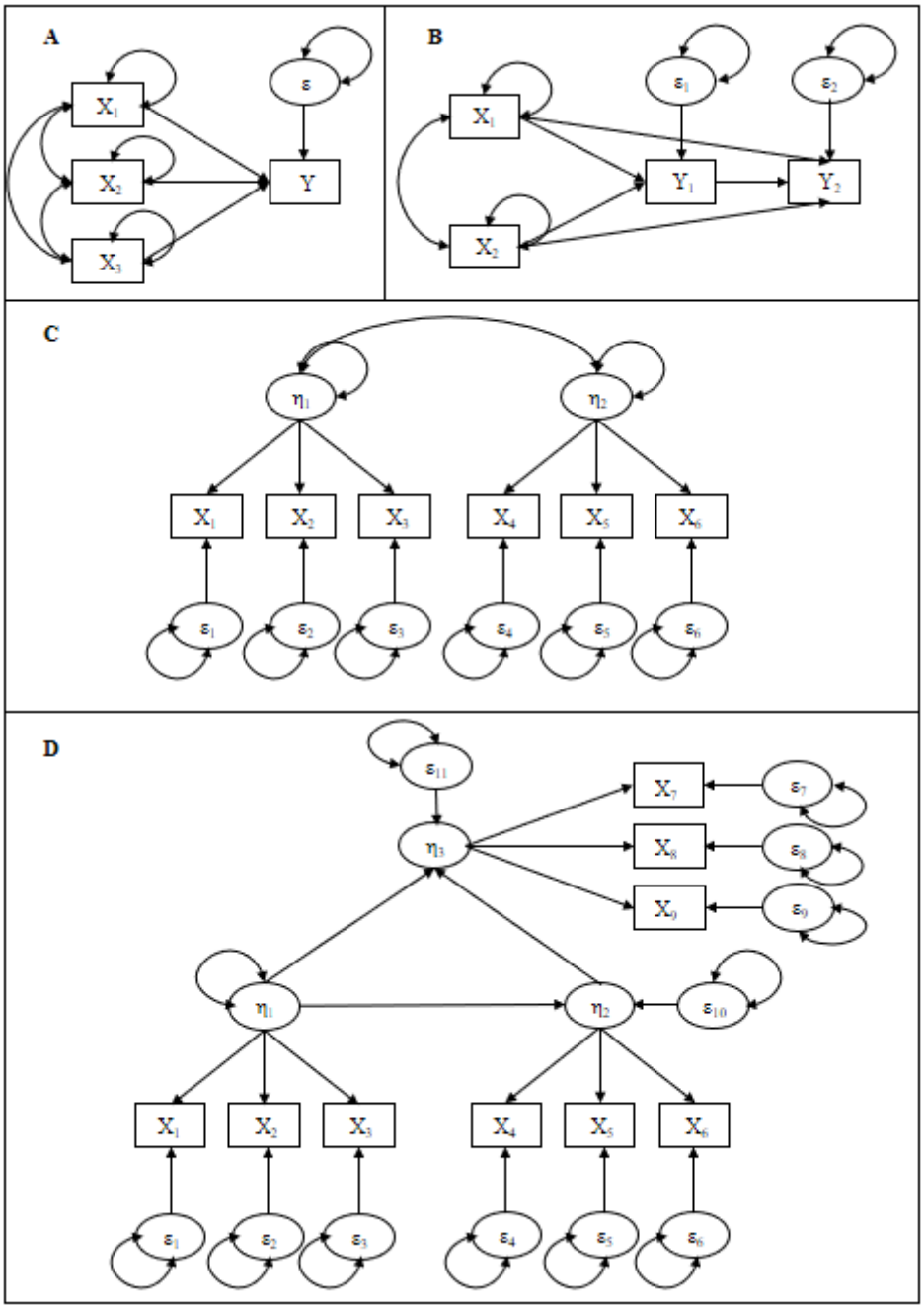
Figure 7. Symbols used in path diagrams



A path diagram contains enough information to translate it into the mathematical representation of the model. Manifest variables are represented by rectangles. Latent variables are represented by ellipses. A supposed causal relationship is represented by a straight single-headed arrow from the "cause" variable to the "effect" variable. Any variable receiving at least

one single-headed arrow is an endogenous variable. Otherwise, it is an exogenous variable. Two correlated variables (without any causal relationship specified) are linked with a curved double-headed arrow (only between exogenous variables). Residual factors are represented either using an ellipse and a straight single-headed arrow received by one endogenous variable, or by an arrow received by an endogenous variable without the depiction of the ellipse. The variance of an exogenous variable (including residual factors) can sometimes be symbolized by a curved double-headed arrow with both heads received by the same variable.

Figure 8. Theoretical examples of some Structural Equation Models. A: A regression model. B: A path analysis model. C: A Confirmatory Factor Analysis model (bidimensional structure with 3 items loading on each dimension). D: A comprehensive Structural Equation Model with 3 measurement models and a structural part



Now that we have described the theoretical underpinnings of the common factor model, we will further briefly introduce another class of latent variable measurement models which are *Item Response Theory models* (IRT). Indeed, although the common factor model is the measurement model under which the simulation study of this thesis work was performed (Part 4), IRT models nevertheless represent an important class of measurement models that can be

used in psychometrics and therefore cannot be omitted of these theoretical prerequisites. In addition, the data used in the simulation study of this thesis work were simulated using a *Rasch model*, which is a family of IRT models.

2.2.b.c. Item Response Theory models

Initially, work on IRT models began in the context of educational testing, and expanded at the beginning of the 1960s [109]. Thus, IRT models are the measurement models used for standardized tests of abilities in universities, like the Graduate Record Examnation (GRE) in the United States of America. Some of the pioneers of IRT models were the psychometrician Frederic LORD (1912-2000), and the mathematician Georg RASCH (1901-1980), who pursued parallel research independently. This measurement model is now increasingly used in health-related research.

IRT models have been first developed for binary items. They model the probability for an arbitrary subject *i* (*i* = 1 to *N*) to give a positive answer to item *j* (*j* = 1 to *p*) (which is usually coded as 1) [86]. This probability is supposed to be a function of two different types of parameters:

1. a *person parameter*, which is the level of the latent variable that is wanted to be measured (called "*latent trait*" in IRT),
2. some *item(s) parameter(s)*.

The logistic model with one item parameter (*1PLM*: one-parameter logistic model, which is mathematically equivalent to the *Rasch model*) can be written as:

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

where $\theta_i \sim N\left(0, \sigma_\theta^2\right)$ is the level of the subject i on the latent trait and δ_j is the *difficulty parameter* of the item *j* (which is the level of the latent trait at which a subject has a probability of 0.5 to give a positive answer to the item) [109].

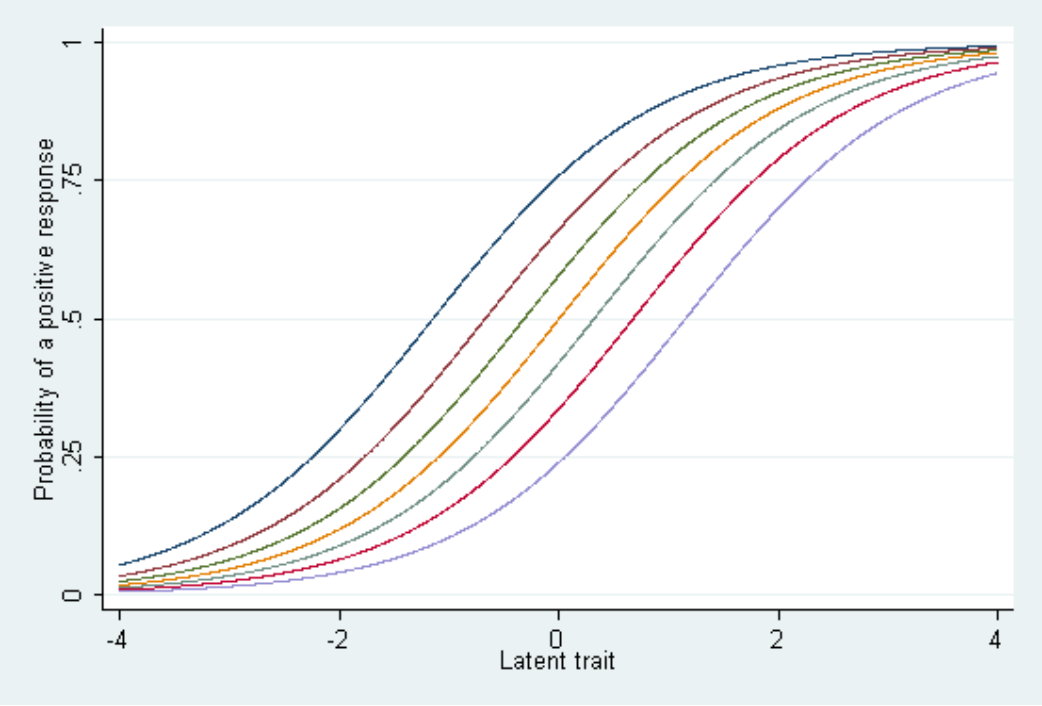
The aforementioned model can be extended with the adjunction of more item parameters like a *discrimination parameter* ν_j (2PLM: two-parameters logistic model, with ν_j representing the ability to discriminate two individuals with a latent trait level close to δ_j) or a *guessing parameter* c_j (3PLM: three-parameters logistic model, with c_j the value of the probability to give a positive answer when $\theta = -\infty$) [109].

The 1PLM and 2PLM can also be extended (Partial Credit Model, Generalized Partial Credit Model, Rating Scale Model, Graded Response Model…) to accommodate the use of nominal or ordinal items [109].

IRT models assume three fundamental hypotheses:

1. the assumption of local independence (the responses to the items are independent conditionally to the latent trait),
2. the assumption of unidimensionality (one latent variable is enough to explain the covariance between responses),
3. the assumption of monotonicity (the probability to give a positive answer to an item is an increasing function of the latent trait) [86].

Figure 9. Item Characteristic Curves of a 7-items scale following a Rasch model. Each curve represents the probability of a positive response to the item as a function of the latent trait. Each item has a higher difficulty parameter value than the one before.



IRT models possess several interesting properties. Indeed, with IRT it is possible to estimate a latent trait with interval scale property, which is rarely achieved with raw scores [110]. Furthermore, regarding the management of missing data, Rasch-based IRT models possess the property of *specific objectivity*, which allows one to obtain consistent estimates of the parameters associated with the latent trait whether or not an item is observed [111].

Consequently, unbiased estimates of the latent trait can be obtained even when some items are missing, in a framework that can be ignorable or not [112].

Specific objectivity is a property restricted to the models of the Rasch family. Indeed, it has to be noted certain authors argue Rasch-family models and other IRT models, although sharing some similarities, are two approaches for modeling responses to items with different epistemologies [113]. To certain authors, IRT models are viewed as more flexible because the possibilities to add parameters in the model are wider. Thus, certain authors consider IRT as an approach from the data to the model: it requires to find a model fitting the data appropriately even at the potential expense of adding too much complexity leading to computational and convergence issues. Contrariwise, Rasch models are often simpler in terms of modeling parameters, but stricter in their fundamental assumptions. Here, the approach is from the model to the data: it requires to design a PRO in order to comply to the assumptions of a Rasch model, and if it is not successful, it is the PRO that is modified (in order to get a PRO with specific Rasch-based model properties like specific objectivity). *In fine*, certain authors argue Rasch Measurement Theory and IRT are two different kinds of psychometric theories [113].

Whatever, although IRT models have been recently extended to accommodate the use of multidimensional constructs, these models are quite computationally intensive to fit with still some convergence issues and therefore may not be the best suited for modeling multidomain concepts.

3. The issue of the longitudinal assessment of a subjective construct

PRO are frequently used to assess the evolution of the level of a subjective construct over time [114]. Often, they are used for the assessment of the effect of the occurrence of a salient medical event (e.g. diagnosis of a severe disease, initiation of a severe treatment like a chemotherapy…), with at least one evaluation before the event and one evaluation after (and eventually a comparison with an appropriate control group) [114]. In this context, this is usually the difference in scores over time that is taken to be an appropriate assessment of the effect of the medical event [114].

However, this relies on the assumption the meaning of the construct one wants to measure is stable in individuals' mind over time [29,34]. As aforementioned, certain empirical results of the end of the 1980s and the beginning of the 1990s have challenged this assumption (see Part 1, chapter 1). These results have led to the conclusion the meaning of a concept is

susceptible to change in individuals' mind over time, especially after the occurrence of a salient event [115]. More importantly, this change in the meaning of a targeted construct over time can have an effect on the difference in scores observed. This leads to the question: when a difference in scores over time is observed, to what can it be attributed (i.e. to an actual change in the targeted construct, and/or to a change in the meaning of the assessed construct)?

This issue has led to the development of what is now known as *response shift* theory (with the term response shift referring to the effect on scores via the change in meaning of a targeted construct) [30]. Since the end of the 1990s, research on the detection, measurement and integration of response shift phenomenon into PRO data analyses has become an important issue in psychometrics, especially within the field of the assessment of HRQL in health-related research. It has led to the development of specific methods, some of them heavily relying on statistical modeling. Nonetheless, as research on response shift phenomenon is still a young field, some issues both on a methodological and statistical level have to be tackled. Thus, certain specific considerations on these methodological and statistical issues constitutes the core of this thesis work. Work on response shift has been prolific since the beginning of the 2000s and as it is not a field with heavily stabilized certainties there are still debate and controversies on the meaning of response shift phenomenon within the academic community. In addition, as it is still a young concept, it is not necessarily an "easy to grasp" concept at a first level. Thus, before the introduction of the original works on statistical and methodological considerations, it seems relevant to firstly propose a state of the art of works conducted on response shift theory at an international level, which is the purpose of the next part of this manuscript.

Part 3: A state of the art on the international works conducted on response shift

“Those who cannot remember the past are condemned to repeat it” – George Santayana, philosopher, essayist, poet and novelist

This part will be a state of the art of the international works conducted on response shift, mostly in the field of health-related research. The core idea of this part was to broadly develop various aspects related to response shift theory: like an historical perspective showing how the occurrence of the response shift notion in psychology was further translated in health-related research, the evolution of the theoretical models linking response shift to perceived HRQL, the various debates and controversies about the meaning of the occurrence of response shift phenomenon, as well as the different methods developed to detect response shift and a synthesis of major empirical results. This part was redacted to be included as a chapter of a book about the issues regarding the measurement of change in perceived health in patients living with a chronic disease.

Book Chapter: The complexity of interpreting changes observed over time in Health-Related Quality of Life: a short overview of 15 years of research on response shift theory

Antoine Vanier[1,2,3], Bruno Falissard[4,5,6], Véronique Sébille[1,7], Jean-Benoit Hardouin[1,7]

Accepted. To be published in Guillemin F, Leplège A, Briançon S, Spitz E, Coste J. (2016). *Perceived health and adaptation in chronic disease. Stakes and future challenge*. New-York-NY. Taylor and Francis

1 Bretagne-Loire University, University of Nantes, EA4275-SPHERE "MethodS for Patients-centered outcomes and HEalth REsearch", Nantes, France

2 Sorbonne Universités, UPMC Univ. Paris 06, Department of Biostatistics, Paris, France

3 AP-HP, University Hospitals Pitié-Salpêtrière Charles-Foix, Department of Biostatistics Public Health and Medical Informatics, Paris, France

4 INSERM, U1178 "Mental Health and Public Health", Paris, France

5 Univ. Paris-Sud and Univ. Paris-Descartes, Paris, France

6 AP-HP, University Hospitals Paris-Sud, Department of Public Health, Le Kremlin-Bicêtre, France

7 University Hospital of Nantes, Unit of Biostatistics and Methodology, Nantes, France

1. Introduction

Patient-Reported Outcomes (PRO) are now widely used in health-related research to assess for instance Health-Related Quality of Life (HRQL), often with self-administered questionnaires [18]. The emphasis on assessing HRQL alongside clinical outcomes is related, in part, to the rise in the prevalence of chronically-ill patients suffering from diseases that cannot be cured. In medical areas such as oncology and palliative care, HRQL is measured over time to add relevant information on patients' subjective experiences in the course of treatment, to counterbalance objective data such as survival time [116]. The increasing interest in collecting data on individuals' HRQL also highlights the recognition that patients should have a say in the choice of their therapy [18]. To enable this, it is necessary to assess other outcomes than health status or symptom levels, since their improvement or deterioration is not always correlated with patients' subjective experiences [14].

Usually, HRQL measures are based on the assumption the meaning of concepts and measurement scales remains stable in individuals' minds over time [34]. Thus, HRQL scores are assumed to be directly comparable for a given individual over time [29]. However, a growing body of literature developed since the mid-1990s has pointed out these assumptions can be over-simplistic, especially when a person experiences a salient health-related event, or has to adapt in living with a chronic disease. Since, numerous studies have suggested the occurrence of a salient event can have an effect on one's representation of the concepts being measured (i.e. HRQL…), which can have an impact on changes observed [115]. Thus, HRQL is now more viewed as a dynamic concept in nature [26]. To put it simply, the abovementioned findings were interpreted as evidences that respondents understand the same questions differently over time [30,117]. This whole process, from the occurrence of the event, to the effect on HRQL scores (via the changes in meaning of the concepts being measured in individuals' minds) is now known as the *response shift theory* [118].

The objective of this chapter is to provide an overview of the works conducted since response shift is investigated in health-related research.

2. First occurrences of the notion of response shift in psychology and health-related research

2.1. Within the field of educational training

Historically, the term "*response shift*" was introduced by HOWARD et al. in 1979 in the field of educational training [51]. Their aim was to experimentally assess the efficacy of various training interventions proposed by psychologists in college and work environments, like improving leadership and performance appraisal or reducing dogmatism.

The typical approach to assess training interventions involves collecting pretest and posttest data on subjects exposed to the intervention, and comparing them with an appropriate control group. In line with usual assumptions on self-report instruments, HOWARD et al. have posited that for pretest and posttest scores to be comparable, a common metric has to exist between the two sets of scores [119]. Thus, in the case of self-report, researchers assume subjects have an internalized perception of their level of functioning and this internalized standard would not change from one testing to the next [51].

Nonetheless, HOWARD et al. published what they called "*a somewhat paradoxical finding*" [120]. In a study many subjects self-reported a higher level of dogmatism after the training despite clients' and therapists' perception that training had been beneficial [120]. They interpreted these results as evidences that the intervention had the ability to improve one's insight or awareness of his/her own level of dogmatism [120]. This led to a change of one's internal standard of measurement over time [120]. Therefore, HOWARD et al. hypothesized that whenever such a shift occurs, conventional pretest/posttest self-reporting is unable to accurately gauge treatment effect [120]. So, they claimed the pretest measurement was inaccurate [120]. Thus, they developed a new experimental design, incorporating a retrospective self-assessment of pretest level (also called "*then-test"*) immediately after posttest assessment [120]. The *difference between the then-test and the posttest* assessment was supposed to assess more accurately *changes induced by training* regarding the concept of interest. The *difference between pretest and then-test* self-report ratings has been referred to as *response shift* [51].

2.2 Within the field of management sciences

In 1976, in the field of organizational changes, GOLEMBIEVSKI et al. introduced a typology of within-individuals changes over time related to self-reports, in a paper about the measurement of change and persistence in human affairs [121].

They defined three types of changes [121]:

1. "*alpha change involves a variation in the level of some existential state, given a constantly calibrated measuring instrument related to a constant conceptual domain;*
2. *beta change involves a variation in the level of some existential state, complicated by the fact that some intervals of the measurement continuum associated with a constant conceptual domain have been recalibrated;*
3. *gamma change involves a redefinition or reconceptualization of some domain, a major change in the perspective or frame of reference within which phenomena are perceived and classified, in what is taken to be relevant in some slice of reality*".

2.3. Within the field of health-related research

One of the first occurrences of the concept of response shift in health-related research can be found in a paper published in 1991 by BREETVELT and VAN DAM [122]. The introduction of the notion was an attempt to explain what they called "*underreporting*" regarding HRQL assessment in cancer patients (i.e. despite frequently reporting a high level of physical complaints, some results suggested no differences in terms of psychological complaints or overall HRQL between cancer patients and healthy people) [122]. In respect to HOWARD et al. definition, response shift, viewed as a change in internal standard of measurement was one of the proposed explanations [122].

BREETVELT and VAN DAM concluded that self-reported HRQL outcomes should be approached with caution, and proposed a then-test design to assess more carefully changes in HRQL related to the occurrence of cancer. In 1996; SPRANGERS proposed one of the first applications of the then-test on two measurement occasions HRQL data in cancer patients [123].

3. Definition and theoretical model

3.1. Current definition of response shift

In 1999, SPRANGERS and SCHWARTZ translated the notion of response shift into the field of HRQL [30]. They proposed the introduction of the response shift concept could be relevant to interpret some "*paradoxical and counter-intuitive*" findings [30] like reporting of stable

HRQL by patients with a life-threatening disease [31], or discrepancies between clinical measures of health and patients' own evaluations of their health [32].

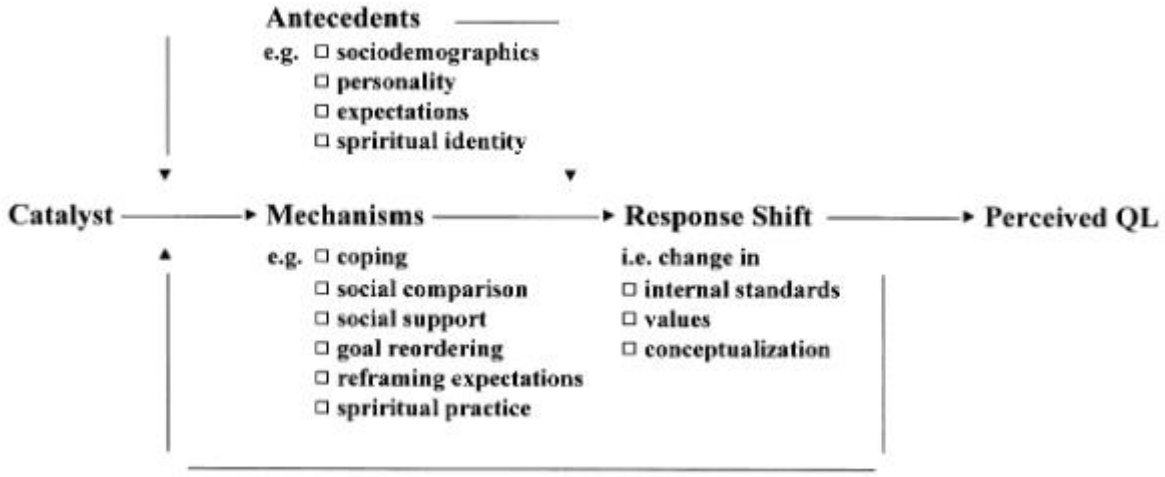
Thus, response shift has been defined as *"a change in the meaning of one's self evaluation of a target construct over time*" [30]. Grounded in the typology of changes introduced by GOLEMBIEVSKI et al. [121], response shift was subdivided in three forms [30]:

1. *recalibration*, which is a change in the respondent's internal standards of measurement (e.g. a person suffering from chronic pain and rating it on a pain scale as 7/10 will later rate it as 5/10 after experiencing acute pain, despite the chronic pain being the same as before), it corresponds to GOLEMBIEVSKI et al. definition of *beta change*;
2. *reprioritization*, which is a change in the respondent's values (i.e. the relative importance of component domains in the target construct - e.g. an athletic person who considers physical functioning as an important part of his/her HRQL may later place emphasis on social functioning after sustaining permanent physical injury), it corresponds, in part, to GOLEMBIEVSKI et al. definition of *gamma change*;
3. *reconceptualization*, which is the redefinition of a target construct (e.g. an item of a multidomain questionnaire initially assessing the domain of mental health, will be later understood by the respondent as assessing another domain, like social functioning), it also corresponds, in part, to GOLEMBIEVSKI et al. definition of *gamma change*.

3.2. First theoretical model proposed (Figure 10)

The first theoretical model designed to address how response shift may have an effect on HRQL scores after changes in health status was published by SPRANGERS and SCHWARTZ in 1999 [30].

Figure 10. First theoretical model of response shift and perceived HRQL proposed by SPRANGERS and SCHWARTZ *(Source: SPRANGERS and SCHWARTZ, Social Science and Medicine, 1999 [30])*



The model was designed to address relationships between five major components. Indeed, it includes:

1. the catalyst: a salient event leading to a change in the respondent's health status;
2. antecedents: stable or dispositional characteristics of the individual. They have both direct and indirect effects on potentiating response shift. They affect the kind of mechanisms engaged, the magnitude and type of response shift;
3. mechanisms: behavioral, cognitive and affective processes to accommodate the catalyst, like the use of coping strategies or upward or downward social comparisons;
4. response shift,
5. perceived HRQL: a multidomain concept with frequently at least 3 broad domains (e.g. physical, psychological and social functioning) [30].

The feedback loop included in the model illustrates the process is thought as iterative and dynamic: perceiving a suboptimal HRQL may lead the individual to reinitiate established or new mechanisms [30].

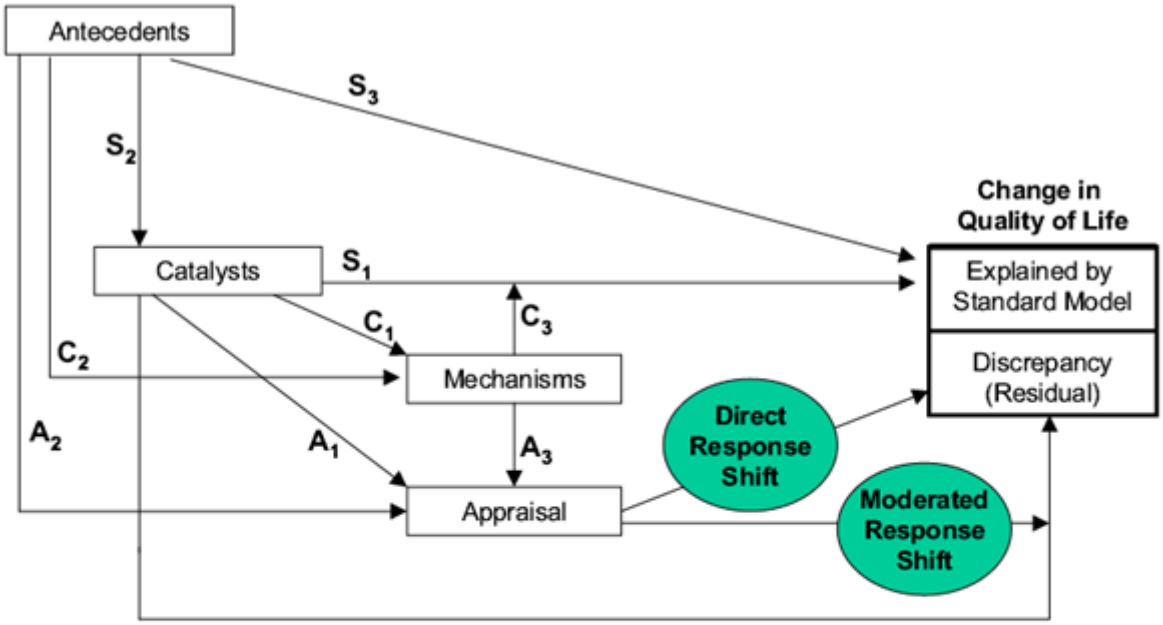
According to SPRANGERS and SCHWARTZ, response shift was isolated both from mechanisms and perceived HRQL as it conceptualizes aspects that are likely to help understanding changes observed in HRQL over time [30]. The isolation of response shift was therefore conceived as a pragmatic approach to explain some aspects of changes in HRQL. As such, response shift was not thought as a replacement to other theories like adaptation theories

[124], discrepancy theories [67], uncertainty in illness theories [125], or stress-coping theories [126], but rather as a concept that can be incorporated in such existing theories [30].

3.3. Update of the theoretical model (Figure 11)

In 2004, RAPKIN and SCHWARTZ extended the first theoretical model [118]. One of the main reasons was that RAPKIN and SCHWARTZ noticed the formulation of the model could be criticized, as it presented some issues of circular reasoning. In the first model, they pointed out response shift was not sufficiently differentiated from both mechanisms and outcomes [118]. Therefore, the concept of response shift overlapped with the psychological mechanisms leading to response shift effect and the outcome affected by response shift effect. So, there was a need to clearly distinguish mechanisms and outcomes from response shift.

Figure 11. Updated theoretical model of response shift and changes in HRQL proposed by RAPKIN and SCHWARTZ. Accounting for changes in standard influences (S), coping processes (C) and appraisal variables (A). *(Source: RAPKIN and SHWARTZ, Health and Quality of Life Outcomes, 2004 [118])*



Note: In this model, three families of hypothetical relationships which can be considered in the explanation of HRQL are postulated. The first family is what we can call *"standard influences"* on HRQL, such as the direct effect of a catalyst (S_1 pathway) or antecedents (S_3), or the indirect effect of antecedents through mediation by the catalyst (S_2). The second family of relationships involves coping mechanisms. First, catalyst is hypothesized to encourage or disrupt coping mechanisms (C_1). There are also hypothesized differences in coping associated with background variables (C_2). Mechanisms of problem-focused coping that reduce the impact of catalyst on HRQL are included as moderator or buffering effect (C_3). The third family of relationships concerns appraisal processes. Coping mechanisms can lead to changes in the appraisal of HRQL (A_3). Catalyst or antecedents variables can also influence appraisal (A_1 and A_2). A change in appraisal ultimately leads to RS effect, hence affecting one's perceived HRQL.

The main new idea behind this updated model is that response shift is now thought as a phenomenon occurring when there is a change in appraisal [118]. Indeed, RAPKIN and SCHWARTZ supposed that any response to a HRQL item can be understood as a function of an appraisal process. RAPKIN and SCHWARTZ posited at least four cognitive processes to account in an individual HRQL assessment.

These four processes are:

1. a *frame of reference* comprising one or more subsets. These subsets can be understood as categories of experiences or events that individuals consider relevant to a particular HRQL scale at the time of the assessment;
2. a *sampling strategy* to extract from categories within a frame of reference specific experiences used to make a HRQL assessment (e.g. in assessing pain an individual can sample "recent instances of pain" or "times when pain interfered with my activities");
3. a *standard of comparison* as a reference point to evaluate the specific experiences sampled (e.g. pain experiences may be compared to "worst pain I've ever had" or to "what my doctor told me to expect");
4. a *combinatory algorithm* used to edit the responses by combining the evaluations into a summary of appraisal of HRQL. It describes how people can use different subjective weights to increase or decrease the relative importance of different experiences [118].

Thus, response shift can be equated neither to mechanisms, nor to the outcome (observed scores). Rather, it is an effect triggered by psychological mechanisms, mediated through a change in appraisal, leading into changes in observed HRQL scores that cannot be explained by standard influences (direct effect of the catalyst, and direct and indirect effect of antecedents: S pathways in Figure 11) [118].

The three forms of response shift proposed by SPRANGERS and SCHWARTZ were related to the aforementioned appraisal processes:

1. change(s) in the *frame of reference* relate to *reconceptualization*,
2. change(s) in the *sampling strategy* or the *combinatory algorithm* relate to *reprioritization*;
3. change(s) in the *standard of comparison* relate to *recalibration* [118].

4. Debate and controversies about the concept of response shift

Initially, in the field of educational training, response shift - equated to recalibration - was interpreted as a measurement bias obscuring the appropriate assessment of the efficacy of an intervention [51].

However, in the model proposed by SPRANGERS and SCHWARTZ the response shift effect is a consequence of the initiation of psychological mechanisms triggered by the catalyst, and it is viewed as one of the path explaining changes in perceived HRQL [30].

Thus, response shift is a concept that can be related to issues in psychology (explaining how people deal with life changes), but also to issues in experimental designs and psychometrics. This has resulted in an ongoing debate about the meaning of response shift [29,36,52,127–131]. This debate often opposes two views on response shift:

1. a view where response shift is more interpreted as a measurement characteristic and therefore a bias that needs to be corrected;
2. a view where response shift is more interpreted as a subject characteristic and therefore seen as a phenomenon leading to meaningful changes worth investigating in their own [52].

This debate has been emphasized in a commentary by UBEL et al. which proposed to abandon the term response shift [36]. They argued the use of the term response shift doesn't help to disentangle two different phenomena. To them, the occurrence of recalibration is a threat to the validity of self-reports (i.e. a measurement bias), obscuring an appropriate assessment of *"true change"* (i.e. a change in the latent targeted construct itself, which corresponds to GOLEMBIEVSKI et al. *alpha change*). However, they view reprioritization and reconceptualization as processes by which people emotionally adapt to circumstances, leading to true change in HRQL. Thus, to them, the indistinct use of the term response shift has led researchers to think of response shift as mainly an issue of measurement bias. Therefore, they proposed to dichotomize response shift by using the term "*scale recalibration*" when referring to recalibration response shift, and using a term like "*emotional adaptation*" or "*hedonic adaptation*" when referring to reprioritization and reconceptualization response shift [132–134].

Nonetheless, this commentary has led to a rebuttal by SPRANGERS and SCHWARTZ [127]. If they have agreed with UBEL et al. the response shift concept is lumping together different phenomena, they have argued the dichotomy proposed by UBEL et al. could be too restrictive.

Indeed, SPRANGERS and SCHWARTZ proposed that scale recalibration can also be a process leading to adaptation to illness; while reprioritization and reconceptualization can also be a psychometric issue (i.e. adding complexity to the measure of true mean change of an invariant construct) but only when HRQL scores are compared, either within individuals over time, or between individuals who have different perspectives on HRQL. Thus, according to SPRANGERS and SCHWARTZ, adaptation is a psychological mechanism and the different types of response shift are consequences of this mechanism.

Thus, if recent studies investigating changes in HRQL put more emphasis on response shift as one of the outcomes of interest [135], the different hermeneutics that are subsumed under the response shift theory are still debated [129]. In that regard, some authors have proposed it is needed to be more accurate about the purpose of an empirical study and the use of the response shift concept when reporting results [128,136].

Nonetheless, debating on what response shift is has helped to define what response shift is not. Indeed, response shift cannot be called anytime there is a difficulty in interpreting changes over time in observed HRQL scores [52]. Thus, response shift cannot be equated to measurement error [37]. It also cannot be equated to other types of processes which may induce changes in observed HRQL scores, like response tendencies such as social desirability or acquiescence, effort justification or cognitive dissonance reduction [30].

5. Methodological approaches to address response shift (Table 1)

Since the occurrence of the first theoretical model of response shift in health-related research, numerous methods have been designed and used to detect response shift effect. These methods can be partitioned into two groups [135]:

1. methods based on specific study design: these approaches are based on a specific design or on the use of specific measurement tools. Therefore, they are used when response shift is anticipated as one of the main outcome of interest;
2. statistical methods: these approaches are based on the use of statistical tools to search for evidences of response shift on datasets. Therefore, they can be used without the need of a specific design.

Table 1 briefly synthetizes the key characteristics of the different presented methods. Each method will be further introduced individually.

Table 1. A summary of some of the key characteristics of the different methods developed to detect response shift

Method	Level at which response shift is analyzed	Forms of response shift investigated	Quantitative assessment of response shift	Number of time points investigated	True change estimate	Note
Methods based on specific study design						
Then-test [120,123]	Group-analysis	Recalibration	Yes	2	Response shift adjusted change	Most used but criticized since
PGI – SEIQoL [137,138]	Individual and/or Groupe analysis	Reprioritization, Reconceptualization	Possible, but complex	At least 2	Possible, but complex	Each person can define what is HRQL for him/her
HEI-Q [139]	Group-analysisl	Recalibration	Yes	2	Response shift adjusted change	Suppose awareness of the response shift process
Vignette ratings [140]	Group-analysis	Reprioritization	No	At least 2	No	
QOLAP [118]	Group-analysis	All forms	Not directly	At least 2	Not directly	Provide an in-depth analysis of changes in appraisal processes
Qualitative interviews [141,142]	Individual and/or group analysisl	All forms	No	At least 2	No	Provide an in-depth analysis of changes observed over time in HRQL
Statistical methods						
SEM (Schmitt's Technique) [143,144]	Group-analysis	All forms	Yes	2, more possible but complex	No	Rely on GOLEMBIEVSKI et al. typology of changes
SEM (OP) [37,39,145]	Group analysis	All forms	Yes	2, more possible but complex	Yes	Can assess response shift from a measurement and a conceptual perspective Currently used at domain level mostly
Latent trajectory analysis of	Identify subgroups with similar	Cannot distinguish which forms are detected	Possible	At least 3	Not directly	Can identify subgroups with similar patterns of response shift along with the group

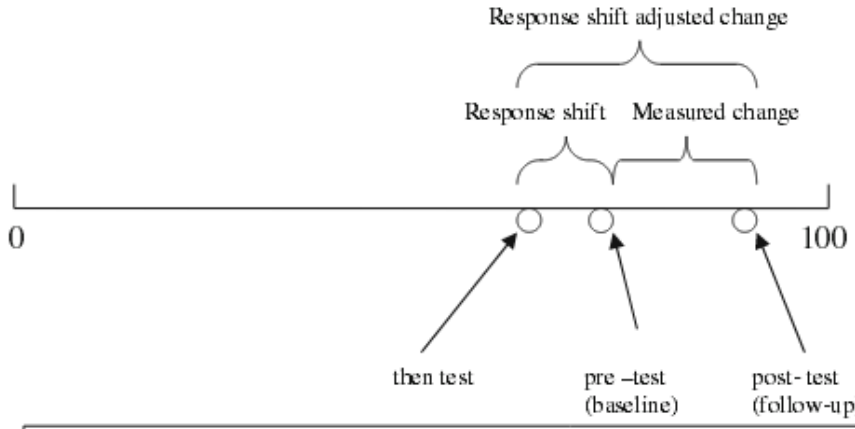
residuals [146]	patterns of response shift					which did not experienced response shift
Relative importance analysis [147]	Group-analysis	Reprioritization	Possible	Multiple-times point	Not directly	Relative analysis: one group is compared to another
CART [148]	Group-analysis	All forms	No	2	No	Data mining technique: infer complex patterns of response shift over time between different investigated groups
CART + Random Forest [149]	Group-analysis	Reprioritization	No	At least 2	No	Assess complex evolution of reprioritization response shift over multiple-times point
IRT (LLRA) [150]	Group-analysis	Recalibration	Yes	2	No	Model retrospective assessment data, assumes absence of true change
IRT (ROSALI) [151]	Group-analysis	Recalibration, Reprioritization	Yes	2	Yes	Could be an interesting choice for detection of response shift at item-level Reconceptualization would require multidimensional IRT modeling

5.1. Methods based on specific study design

5.1.a. Retrospective rating or the then-test approach (Figure 12)

The then-test approach is based on the same design used by HOWARD et al. in the field of educational training [120]. Usually, changes in PRO observed scores, regarding the occurrence of a salient event (e.g. diagnosis of a disease, initiation of a therapy…), are estimated by the difference between a baseline assessment (pretest, before the event) and a follow-up assessment (posttest, after the event). The then-test is a retrospective assessment of the baseline assessment, performed at the same time as the posttest [152]. Therefore, it is assumed that posttest and then-test are sharing the same internal standards of HRQL within individuals, thus accounting for recalibration response shift [135]. The difference between the then-test and the pretest assessment is assumed to be an estimate of the magnitude of the recalibration effect. The difference between the posttest and the then-test is used as a measure of recalibration-adjusted change [135].

Figure 12. The then-test approach to detect response shift *(Source: BARCLAY-GODDARD et al. Quality of Life Research, 2009, [135])*



This figure represents a measure of HRQL on a 0-100 scale with a traditional pre-test at baseline, post-test at follow-up, and a then test of baseline health occurring also at follow-up.

Example of a typical then test question: Please remember to when you last filled out this questionnaire 1 month ago. Please provide a new judgment about your quality of life at the time you last filled out this questionnaire 1 month ago. Rather than try to recall your prior responses, please offer a renewed judgment of your quality of life then, given what your perspective is today.

The then-test has been the most used method for assessing recalibration effect [115]. Since 1996, it has been used with a variety of PRO instruments and patient populations.

However, it has been criticized since [153]. First of all, the basic premise the then-test and posttest shares the same internal standards has been questioned [154]. Then, as the then-test asks respondents to provide a retrospective assessment, it implies individuals can access the evaluated previous state in their minds appropriately without recall bias, although the effect of this bias has been shown in empirical studies [155]. Third, there is a potential contamination due to other response biases, such as social desirability effect of effort justification [51]. Lastly, the then-test implies the retrospective assessment is achieved by directly extracting appropriate data about the health state being assessed and rating it accordingly. However, some authors have argued the actual cognitive mechanism engaged can be more understood as an "*implicit theory of changes"*, where people think about past times starting from the present and reconstruct their baseline state inferring what they think they were at that time [156]. If true, it would threaten the possibility of a retrospective assessment. Thus, the use of the then-test as an adequate method to detect recalibration response shift is currently heavily questioned [153].

5.1.b. Specific questionnaire assessing the magnitude of recalibration for a specific intervention

The Health Education Impact Questionnaire (HEI-Q) was developed to identify recalibration among individuals participating in an arthritis self-management course [139]. Each of the nine items is rated using a seven-point scale to estimate if recalibration has occurred and at which magnitude. Recalibration is described as negative when people realize that they were worse than they thought at a prior point, positive when they were better than thought at a previous point, or absent [139]. This type of questionnaire has been designed to be used to assess the efficacy of a specific intervention whose purpose is to induce recalibration. Nonetheless, it seems the heuristic of the questionnaire is based on the assumption people are aware their internal standard of measurement has changed over time, which imply recalibration is thought here as a conscious process only.

5.1.c. Questionnaires assessing how HRQL is conceptualized by individuals

Using these questionnaires, an individual is asked to select and rate the value of different domains of HRQL. Two PRO instruments have been used to detect reprioritization and reconceptualization response shift: the Patient Generated Index (PGI) and the Schedule for the Evaluation of Individual Quality of Life (SEIQoL) [137,138,157]. Both instruments, if different in presentation, are close in process. First, a person is asked to select five life areas that are

supposed to match his/her individual definition of HRQL. Then, this person rates his/her ability in these areas, and is asked to provide weights reflecting the relative importance according to each of the five areas. Over time, reprioritization can be inferred by a change in the weights of the different domains, and reconceptualization by a change in the choice of the domains [137,138,157].

Those methods have the advantages to be very individualized procedures: each individual can choose what domains best represent HRQL for him/her. However, it is hard to provide a numerical value estimating the magnitude of the reprioritization and/or reconceptualization effect [135].

5.1.d. Vignette ratings

When this design is used, patients are asked to read brief vignettes describing in a few sentences different hypothetical health states. For example, each vignette can briefly describe a side-effect of prostate cancer treatment [140]. Before and after treatment (i.e. surgical removal of the cancer), patients rate how each state described by the vignettes appears to be detrimental by their own values. Reprioritization response shift can be detected when the rating of a vignette is significantly different before and after the treatment [140].

5.1.e. Use of the Quality of Life Appraisal Profile (QOLAP)

This questionnaire was designed to assess the four aforementioned cognitive appraisal processes engaged when respondents are asked to take a survey [118] (see Part3, chapter 3.3). The purpose of the questionnaire is to get a precise description of each of the processes (frame of reference, sampling strategy, standard of comparison and combinatory algorithm) for each respondent when asked about their level of HRQL using a validated PRO instrument [118]. Therefore, the occurrence of response shift effect can be inferred when there are changes in these appraisal processes over time for a given respondent. Each form of response shift can be inferred as each appraisal processes have been linked to different forms of response shift [118].

5.1.f. Qualitative interviews

Qualitative interviews can be conducted to generate an in-depth knowledge of how a person is experiencing changes in HRQL over time. Content analyses using specific tools can

then be used to analyze the verbatim of the interviews and help eliciting occurrences of response shift [141,142].

5.2. Statistical methods

5.2.a. Using Structural Equation Modeling (SEM): The Schmitt's technique

Historically, detecting and taking into account various types of within-individuals changes over time of a targeted construct using factor analysis has been proposed even before response shift was defined in health-related research. Indeed, a method designed to detect the aforementioned types of changes introduced by GOLEMBIEVSKI et al. (see Part 3, chapter 2.2) has been developed by SCHMITT in 1982 [143].

The Schmitt's Technique relies on an operationalization of GOLEMBIEVSKI et al. typology of changes as change(s) in the value of SEM parameters between two times of measurement.

This method has been used several times in the field of health-related research to detect response shift when measuring changes observed over time in HRQL scores [41,48,144].

Nonetheless, as the typology of changes on which it relies is not strictly equivalent to the typology of response shift proposed by SPRANGERS and SCHWARTZ, there was, since the late 2000s a shift in use in favor of another method for detecting response shift using SEM: the Oort's Procedure.

5.2.b. Using Structural Equation Modeling: The Oort's Procedure (OP)

The Oort's Procedure was first proposed in 2005. It relies on an operationalization of the different forms of response shift (as proposed by SPRANGERS and SCHWARTZ) as change(s) in the value of SEM parameters between two times of measurement [37].

As OP proposes certain interesting features, it has been used to detect response shift on several clinical datasets [39–50]. When used to model observed HRQL scores of a multidomain questionnaire (response shift detection at "domain-level"), it allows assessing for each domain whether it is affected by one or several specific form(s) of response shift (and if so, by which magnitude), together with estimating true change in HRQL after taking into account response shift. It also allows indicating the respective contribution of true change and response shift in explaining changes in each observed HRQL domain scores [39].

The development of the OP has also been helpful as a framework to help clarify a formal definition of what is response shift (in terms of models and psychometrics in the broadest sense, irrespective of the method of estimation) [158]. Indeed, in 2009, OORT et al proposed to formally define response shift from two different perspectives (a *"measurement perspective"* and a *"conceptual perspective"*) [145] (Table 2 and Table 3). According to OORT et al., response shift from a measurement perspective is a special case of measurement bias, where changes over time of the level of a latent attribute of interest A (e.g. true change in HRQL itself) cannot fully determine changes of test scores X (e.g. observed HRQL scores) [145]. However, response shift from a conceptual perspective is a special case of explanation bias where a change over time of an attribute of interest A cannot be fully explained by observed variables E representing standard influences on HRQL, but also by other variables V representing mechanisms (e.g. coping mechanisms, social comparisons…) [145] (Figure 13).

Table 2. Measurement perspective, conceptual perspective, measurement bias, explanation bias in HRQL, according to OORT et al. *(Source: OORT et al. Journal of Clinical Epidemiology, 2009 [145])*

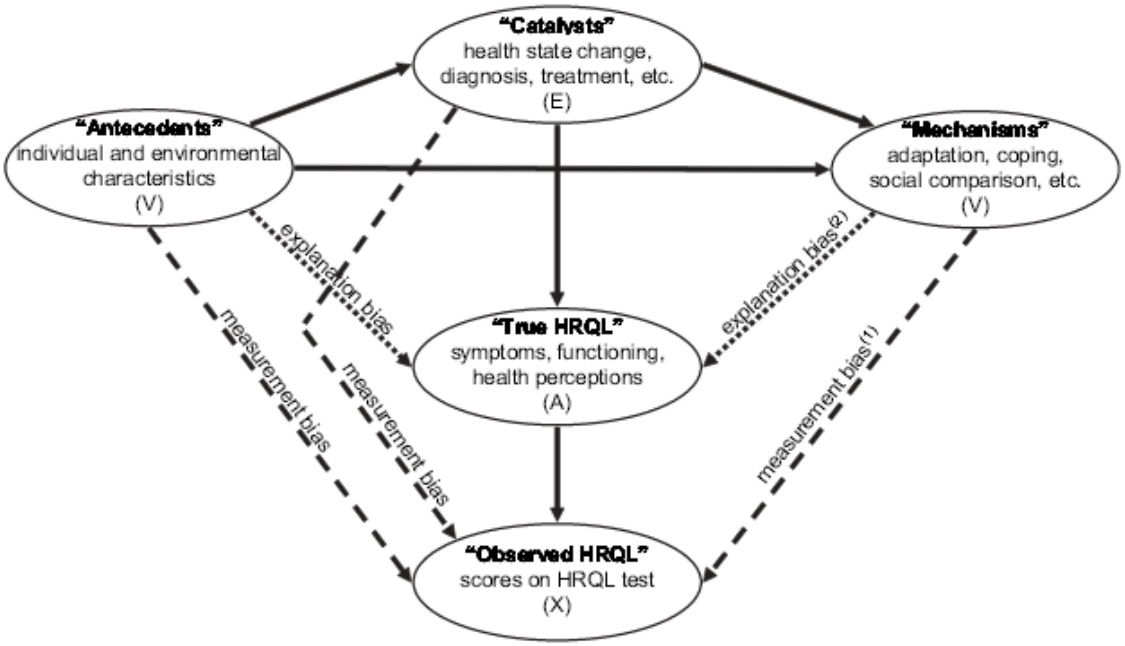
Bias	Measurement perspective	Conceptual perspective
Bias introduction	Measurement bias is bias in the measurement of the target construct (i.e., the attribute of interest). Differences between respondents in observed test scores cannot be fully explained by true differences between respondents in the attribute of interest. Variance in test scores does not fully represent true variance in the attribute.	Explanation bias is bias in the explanation (or prediction) of the target construct (i.e., the attribute of interest). In addition to the acknowledged explanatory variables, there are other variables that also explain part of the variance in the attribute of interest, possibly confounding the effects of the acknowledged explanatory variables.
Bias definition	Measurement bias is formally defined as a violation of measurement invariance, $f(X\|A = a, V = v) = g(X\|A = a)$, for all values of a and v, where X are observed variables (measurements of A), A are the attributes of interest, and V are possible violators of invariance. Function f is the distribution function of X given a and v, and function g is the distribution function of X given just a	Explanation bias can be formally defined as a violation of explanation invariance, $f(A\|E = e, V = v) = g(A\|E = e)$, for all values of e and v, where A are the attributes of interest, E are acknowledged explanatory variables (i.e., causes or predictors) of A, and V are possibly confounding variables. Function f is the distribution function of A given e and v, and function g is the distribution function of A given just e
Bias in HRQL research	In HRQL research, measurements X are observed scores on a HRQL test, attributes A are HRQL itself and a are true HRQL values, and V are all other variables in HRQL studies. With measurement bias in a HRQL test, differences between patients in observed test scores cannot be fully explained by differences in their true HRQL. As a result, test scores also represent something else besides HRQL.	In HRQL research, A are HRQL values, E are causes or predictors of HRQL, such as biological and physiological factors, physical and psychological symptoms, and V are all other variables that may also affect HRQL but are not taken into account. The explanation of patients' HRQL may be confounded by variables other than the acknowledged predictors or causes of HRQL.
Note	In the measurement perspective, all variables other than A are considered potential violators of measurement invariance. We do not distinguish between acknowledged explanatory variables and other, possibly confounding variables; all such variables are subsumed under V.	In the conceptual perspective, we do distinguish between acknowledged explanatory variables E and possibly confounding variables V, but we do not distinguish between the attribute of interest itself A and its operationalization X; that is, A and X are assumed to coincide.

Table 3. Response shift in measurement and conceptual perspective according to OORT et al. *(Source: OORT et al. Journal of Clinical Epidemiology, 2009 [145])*

Response shift	Measurement perspective	Conceptual perspective
Response shift introduction	In measurement perspective, response shift can be understood as bias in the measurement of change in attribute *A*: it explains observed change (in *X*) that cannot be fully explained by true change (in *A*). Hence, response shift is measurement bias that varies with time of measurement.	In conceptual perspective, response shift can be understood as bias in the explanation of change in attribute *A*: it explains (true) change in *A* that cannot be fully explained by the acknowledged explanatory variables *E*. Hence, response shift is explanation bias that varies with time of measurement.
Response shift definition	Response shift can be formally defined as measurement bias, with *X* representing the respondents' observed scores on repeatedly administered tests, *A* representing the true attribute values of the respondents at the times of measurement, and *V* being the time of measurement itself or other variables with effects on *X* that vary with time of measurement.	Response shift can be formally defined as explanation bias, with *A* representing (true) attribute values on repeated test administrations, *E* representing values for acknowledged explanatory variables (causes or predictors) of *A*, and *V* being the time of measurement itself or other variables with effects on *A* that vary with time of measurement.
Response shift in HRQL research	In HRQL research, *X* are observed scores on repeatedly administered HRQL tests, *A* are true HRQL values at the times of measurement, and *V* is (a coding for) the time of measurement itself, e.g., before or after a health-state change, or *V* are individual characteristics that interact with the time of measurement, or occur or change between the measurement occasions (e.g., adaptation, coping, social comparison). With response shift, observed changes in test scores cannot be fully explained by true changes in HRQL.	In HRQL research, *A* are HRQL values at different measurement occasions, *E* are causes or predictors of changes in HRQL (e.g., health-state change, medical diagnosis, medical treatment), and *V* is (a coding for) the time of the measurement occasion, or individual characteristics (other than *E*) that interact with the time of measurement, or occur or change between occasions of measurement (e.g., adaptation, coping, social comparison). With response shift, variables other than the acknowledged explanatory variables also affect changes in HRQL.
Response shift in terms of Sprangers and Schwartz (1999)	*X* are test scores and *A* are true values for HRQL or other patient-reported "outcomes" at different measurement occasions. *V* can be "catalysts," e.g., health-state change, diagnosis, medical treatment, "mechanisms," e.g., adaptation, coping, social comparison, "antecedents," i.e., (other) individual and environmental characteristics, or any other variable that might be relevant. The "mechanisms" affect the observed test scores through recalibration, reprioritization, or reconceptualization of test items and response scales (but not true HRQL). As a result, observed change does not reflect true change, and follow-up scores cannot be compared with previous (baseline) scores.	*A* are HRQL variables or other patient-reported "outcomes" at different measurement occasions. *E* are "catalysts," e.g., health-state change, diagnosis, medical treatment, and possibly "antecedents," i.e., any individual and environmental characteristics that are commonly acknowledged as explanatory of *A*. *V* are "mechanisms," e.g., adaptation, coping, social comparison. The "mechanisms" affect true HRQL through recalibration, reprioritization, or reconceptualization of true HRQL. As a result, observed change still reflects true change, but effects of "catalysts" on HRQL are possibly confounded by co-occurring effects of "mechanisms."

Thus, OORT et al. proposed in 2009 an extended revision of the OP which allows detecting and accounting for response shift from the two aforementioned perspectives. The measurement part of the SEM is used to detect response shift from a measurement perspective, and the structural part of the SEM, including different variables (i.e. variables representing standard influences on HRQL, but also psychological mechanisms…) susceptible to explain changes in HRQL itself, is used to detect response shift from a conceptual perspective [145,159] (Figure 13).

Figure 13. Measurement bias and explanation bias in HRQL according to OORT et al. In a longitudinal design with repeated measurements of HRQL, this measurement bias (1) can be considered as response shift in the measurement of change; and this explanation bias (2) can be considered as response shift in the explanation of change *(Source: OORT et al. Journal of Clinical Epidemiology, 2009 [145])*



The main drawback of OP is the fact this procedure implies analyses at group-level [37]. Therefore, when performing detection of response shift using OP, it is assumed a substantial part of the sample has experienced response shift of a similar form(s), direction and magnitude. In some study settings, this is probably a strong assumption [146].

5.2.c. Using longitudinal regression modeling and latent trajectory analysis

This method has been proposed to overcome the fact OP is group-level analysis. Therefore; this method is suited when the purpose of the analyses is to identify subgroups of subjects exhibiting different patterns of response shift over time or a subgroup having not experienced response shift over time [45,146,160].

Subgroups of patients exhibiting similar fluctuations of centered residuals over time (from a longitudinal regression model predicting changes of HRQL over time using various predictors representing health status, background characteristics and levels of symptoms) are supposed to be individuals exhibiting a similar pattern of response shift over time. Thus, subgroups of individuals exhibiting different patterns of negative and/or positive response shift

can be identified along with the group of individuals supposed to have not experienced response shift (the group with centered residuals close to zero over time) [146]. The timing of the occurrence of response shift can also be hypothesized if appropriately spaced time points are modeled. Nonetheless, this method cannot identify the forms of response shift and does not strictly provide a measure of true change.

5.2.d. Relative Importance analysis using discriminant analysis and/or logistic regression

This method has been proposed to detect specifically reprioritization response shift and therefore is suited when a researcher wants to test if reprioritization response shift has occurred when HRQL is measured with a large number of domain scores [147,161]. It relies on the use of discriminant analysis and/or logistic regression to predict group membership (e.g. active versus inactive disease over a specific period of time) by the difference over time of various HRQL domains scores. The occurrence of reprioritization over time is inferred in one group of individuals by comparison to another group [147]. Thus, the method is useful when one hypothesizes a different evolution between two groups of individuals.

5.2.e. Using machine learning technique such as recursive partitioning tree analysis

This method has been proposed for inferring complex and non-linear patterns of response shift between two-times point in a dataset. It relies on a data-mining technique (i.e. Classification and Regression Trees (CART)) [45,148]. A graphical model, in the form of a decision tree, explaining the difference in HRQL score is constructed by recursively splitting the data set into two groups on the basis of various predictors in such a way the heterogeneity of the obtained subsamples is minimized regarding HRQL score. A model is fitted on various groups of the study based on disease trajectory. The occurrence of each form of response shift can be hypothesized by searching for differences in the way the various predictors explain the differences in HRQL score in each group.

A variant of this method has also been proposed specifically for detecting patterns of reprioritization response shift over multiple-time points by adding the use of random forest method to CART [149,162]. This addition allows estimating the importance of each HRQL domains in explaining overall HRQL score over time. Complex patterns of evolution of

reprioritization response shift can therefore be inferred by observing changes in importance of the HRQL domains over time.

5.2.f. Using Item Response Theory models (IRT)

The use of IRT models to detect response shift has sometimes been evoked, as they possess some interesting measurement properties [135]. Thus, the use of the Linear Logistic model with Relaxed Assumptions (LLRA) model has been tested once, for detecting recalibration response shift [150]. However, this method imposes specifying the model on retrospective assessment data and assumes an absence of true change. Nonetheless, recently, a method specifically designed to detect different forms of response shift and estimating true change using IRT model was proposed: the RespOnse Shift ALgorithm in Item response theory (ROSALI) [163].

ROSALI relies on an operationalization of non-uniform and uniform recalibration, and reprioritization as change(s) in the value of polytomous IRT models between two times of measurement [163].

The use of IRT could be an interesting alternative for response shift detection compared to OP, as it could beneficiate of some properties of IRT models, in particular the possibility to estimate a latent trait with interval scale property [110]. In addition, as IRT directly models responses to items as a function of a latent trait, it could be a method of choice for response shift detection at “item-level” (i.e. with categorical responses to items as variables used to estimate a unidimensional concept).

However, ROSALI is currently based on unidimensional IRT models and does not yet include the possibility of reconceptualization detection which would require multidimensional IRT modeling [163].

6. A brief overview of some results from studies investigating the occurrence of response shift effect

Since the introduction of the first theoretical model of response shift effect on perceived HRQL in 1999, a growing body of studies has investigated the occurrence of response shift, in a variety of clinical settings, using various methods (although, most of the times the then-test approach was used [115]). Thus, evidences of response shift effect (recalibration and/or

reprioritization and/or reconceptualization) has been suggested in a wide range of diseases: including acute events like occurrence of cancer [123,164–166], stroke [43,167,168], coronary artery disease [48], hearing loss [169,170], different types of surgeries [137,161,171–173], or chronic diseases like multiple sclerosis [44,160], chronic back pain [46], diabetes [174], or end-renal stage disease [142]. The occurrence of response shift has been investigated, not only as a consequence of the beginning of a salient disease, but also when patients are recovering from a health-related event susceptible to lead to chronic disabilities (e.g. being handicapped after a stroke [144]), or after the beginning of a major treatment course (e.g. initiation of radiotherapy to treat cancer [123]). Evidences of response shift effect have also been suggested after the initiation of a self-management course to help overcome difficulties related to a particular disease [41,139]. Lastly, the occurrence of response shift effect has been documented in various populations, from children to elderly people [49,170].

One of the many questions that were investigated in studies about response shift was the magnitude of such an effect on changes in scores. In terms of magnitude, response shift effect on changes in observed scores is usually reported as a small effect. Nonetheless, the issue of the magnitude of the response shift effect on HRQL observed scores is related to the issue of the clinical importance of taking into account response shift effect when assessing changes in "true" HRQL. Indeed, the possibility of missing a clinically important change in HRQL if response shift is not appropriately taken into account has been pointed out [135]. In line with this point of view, studies have shown that changes in HRQL can be underestimated when response shift was not taken into account [39,137,167,175].

As stated before, some statistical methods (e.g. OP) that can be used to detect response shift make the assumption that a substantial part of the sample has demonstrated response shift in the same direction and magnitude [37]. However, the results of some studies have challenged this assumption. Indeed, estimates of the prevalence of response shift vary greatly across studies. Within individuals at 6 months post stroke, using a variety of methods to detect response shift, it has been estimated that 28% to 78% of the individuals has demonstrated response shift [157,167,168]. In contrast, in two studies (on patients post-stroke and on patients with multiple sclerosis respectively), it was suggested that most of the individuals did not exhibit response shift [146,160].

Response shift, as a psychological phenomenon, is theoretically often viewed as the result of a process of adaptation: i.e. as a phenomenon triggered by psychological mechanisms helping people to adapt to negative circumstances and helping them to feel themselves as good

as possible despite experiencing a deteriorating health [30,135]. Nonetheless, it has been hypothesized that response shift could be, in some circumstances, a maladaptive process [30]. Thus, it can be noted some studies have pointed out that within a sample of individuals experiencing the same health-related event, response shift can be detected as having an effect on HRQL scores in opposite directions across subgroups [139,146].

The occurrence of response shift effect is often investigated in explaining changes in observed HRQL pre and post occurrence of an acute event, or in explaining changes related to dealing with a chronic disease. Nonetheless, the possibility response shift can be a phenomenon occurring only with the mere passage of time has been postulated [52]. Currently, results have been reported supporting the idea response shift can occur without strong evidences of the occurrence of an acute health-related event [49]. But, in a recent study investigating changes in HRQL within individuals living with chronic diseases, but with reported stable health, little evidences for response shift effect was found over a year of follow-up [50].

Lastly, it has been proposed response shift can be anticipated as a positive effect of interventions such as self-management and psycho-social programs, rehabilitation, and palliative care. Facilitating adaptation can enable the occurrence of response shift as a desired outcome of such interventions. These interventions can be proposed to people suffering from a chronic disease, where symptoms and functions may not be dramatically improved, but better HRQL is desired. Some studies have shown evidences interventions designed to facilitate adaptation (through the measure of response shift occurrence) can help to improve HRQL over time [41,139,176,177].

7. Conclusion

The development of the response shift theory has been helpful in highlighting that when individuals are asked to rate their level of HRQL, their appraisal of the construct being measured can change over time, leading into changes in perceived HRQL. So, interpreting observed changes in HRQL scores over time, especially when a salient health-related event has occurred can be more complex that initially thought. Thus, response shift theory has provided a pragmatic theoretical framework, which was translated into different methods designed to help assessing changes over time of a targeted construct. It was used in various empirical settings.

Nonetheless, the research field about response shift in health-related research is still a young field and there is a lot of room for improvement.

On a theoretical level, if the first theoretical model has been empirically tested once [178], the updated theoretical model proposed by RAPKIN and SCHWARTZ has never been tested on data with statistical methods designed to test linear structural relationships [179]. Such studies should be conducted. In addition, there is still the need of pursuing the debate on the meaning of response shift effect, to clarify what is the hermeneutic of response shift in various study settings. A needed related discussion can also be the debate on interpreting the response shift effect as a phenomenon linked to subjects' characteristics as opposed as a phenomenon linked to the characteristics of a PRO instrument [52]. Indeed, if response shift is currently often interpreted as an effect highlighting how individuals adapt to life changes, some authors have also pointed out that the characteristics of a questionnaire can participate in enhancing response shift. For example, it has been postulated response shift can be more susceptible to occur when a concept is measured by means of evaluation-based items (i.e. rating experiences compared to an internal standard, e.g. how difficult is it to walk up a flight of stairs?), as opposed to performance-based items (e.g. time to walk up a flight of stairs) or perception-based items (e.g. how often do you walk upstairs?) [29,52,53]. It has also been postulated response shift could occur because HRQL scales assumed that people are weak evaluators (i.e. people would rate their level of HRQL in regards to contingent circumstances, like current physical condition), although people could probably be strong evaluators instead (i.e. people would rate their level of HRQL in regards to how a condition have an impact on higher-level individuals' motivations or purposes) [53]. In addition, it has been hypothesized recently the occurrence of response shift could be linked with the semantic complexity of concepts and items [133].

On a methodological level, if various methods have been developed to detect and take into account for response shift, there are still issues. Currently, each method has drawbacks. The then-test approach has been the most used, but it is currently less recommended [153]. OP allows detecting all forms of response shift along with an estimate of true change, but it currently implies group-analysis, which can be a strong assumption [37,145]. Latent trajectory analysis allows identifying subgroups of subjects with different response shift patterns over time, but it cannot assess the form of response shift and does not provide directly an estimate of true change [146]. Relative importance analysis can be useful when focusing on differences between two-groups, but it can only detect reprioritization in one group compared to another [147]. CART and random forest method can be useful for inferring complex patterns of

response shift over time, but they don't provide a quantitative assessment of response shift [148,149]. ROSALI shares the advantages of OP along with some interesting measurement properties because of the use of IRT models, but it cannot be currently used with multidimensional constructs [151]. In addition, there is the need to clarify what is response shift at "item-level" as opposed to response shift at "domain-level" [50,133]. To provide a method, or a combination of methods, which could allow detecting subgroups of subjects with different patterns of response shift, detecting each form of response shift in each group, on multiple-times point, both for item and/or domain-level, along with a direct estimate of true change will be a challenge.

Lastly, when investigating the occurrence of response shift as a phenomenon of interest in empirical settings, there are still a lot of questions that need future developments or clarifications. For example, little is currently known about the subjects' characteristics that are susceptible to explain why some individuals will experience response shift after the occurrence of a catalyst, while others will not [135]. There are still questions about the timing response shift, or about the interrelationships between each of the components (from the catalyst to perceived HRQL) when explaining occurrence, direction, and magnitude of response shift.

Future researches conducted altogether in the three aforementioned areas (i.e. theoretical level, methodological level and empirical level) will be needed to assess if response shift theory will be one of the most heuristic framework in explaining observed changes in HRQL, useful either in studies dealing with the assessment of interventions, or in large epidemiological settings.

Part 4: A pilot simulation study on Oort's procedure performances

"81%...82%...83%..." – Freeza, a fictional evil anthropoid alien life form

This fourth part will be mainly focused on an original work on OP performances at item-level via a pilot simulation study. Indeed, if OP has been used on several clinical datasets, little is known regarding its statistical performances, especially its capacity to detect actual response shift. Moreover, certain methodological choices of the OP algorithm can be discussed and investigated. But before the main core of the simulation study manuscript, a first chapter will be dedicated to introduce the reader to the rationale behind simulation (or "Monte-Carlo") studies in statistics. Then, comprehensive formal details about how the data of the study were simulated will be presented. Also, how the OP algorithm was automatized in order to be performed on a large number of datasets will also be exposed.

1. A simulation study?

1.1. Why a simulation study?

Simulation studies are empirical designs where a large quantity of data are artificially generated in order to investigate the behavior or performances of a complex statistical technique. They are sometimes referred to as "*Monte-Carlo studies*" [180]. They can be used to provide guidance for applied researchers about the best conditions where a statistical technique is supposed to perform the most adequately. Thus, they can provide help in the planning of empirical studies using real data. They can also be useful when comparing the performances of several statistical strategies.

Simulation studies must be used when the performances of a statistical technique cannot be assessed analytically because of complexity (e.g. a simulation study would be useless to study the sampling variability of a sample mean because theories like the Central Limit Theorem provides the necessary solution analytically) [180]. The fundamental principle of a simulation study is the fact the researcher identifies *a priori* the values of *population parameters, probability distributions and models* that will be used to generate a large number of random samples [180]. A *combination of different characteristics* of the sample can be investigated (e.g. variability in sample size, population parameters…). Therefore, a *large number of random samples* have to be generated for each combinations of characteristics assessed. Then, the investigated statistical technique is performed on each generated dataset and the performances of the technique can be estimated by averaging the results for each of the conditions assessed. Thus, simulation studies can help to investigate the effect of the

aforementioned characteristics of the sample on indicators such as bias, power, Type-I-Error, amount of correct detection…

The main advantage of simulation studies comparing to investigate the performance of a statistical technique on an empirical real dataset is the fact population parameters are *fully determined a priori* by the researcher and therefore *known* [180]. Thus, *the laws* which have generated the data are *fully known*. Contrariwise, when using data from an empirical real sample, only sample estimates of the population parameters can be obtained, but "the truth" behind the data remains unknown. Therefore, when performing a simulation study, the performances of a statistical technique can be compared to fully known population parameters, probability distributions, and models.

Nonetheless, the main limit of simulation studies is the fact real-world-data are usually extremely complex (i.e. including hierarchical nested structures, existence of heterogeneous subgroups, missing data with a non-ignorable mechanism…) and simulation studies can quickly be computationally intensive to perform. Therefore, there usually exists a difficult balance to achieve between simulating data with an appropriate simulation model and with sufficient combinations of characteristics of the sample in order to investigate situations which can be representative of empirical conditions, and feasibility of the simulation study. To summarize, a balance must be found between the generalizability of the findings of a simulation study (i.e. quality and quantity of the conditions assessed), the precision of the findings (i.e. number of samples generated for each conditions assessed) and practicality [180].

1.2. How the data of this study were simulated?

1.2.a. Generalities, parameters, distributions and model used

As it will be further mentioned in the core manuscript of this simulation study (see Part 4, chapter 4.1), the structure of the data simulated corresponded to responses to five binary items, at two times of measurement *(t_0* and *t_1)*.

Four types of sample characteristics could vary according to different fixed levels:

1. *n* (sample size) could be fixed at 100, 200 or 300;
2. α (changes in latent trait mean level between the two times or "true change") could be fixed at 0 (no "true change") or -0.2 (a decrease in latent trait mean level between t_0 and t_1);

3. *r* (correlation between latent traits between the two times) could be fixed at 0.4 (moderate correlation) or 0.9 (very strong correlation);
4. *ur* (occurrence of uniform recalibration) could be fixed at 0 item, 1 item (on the third item), or 2 items (on the second and fourth items).

Thus, 36 combinations of the levels of the sample characteristics could be investigated. A thousand datasets have been simulated for each combination.

In terms of models and population parameters, the responses to binary items were simulated using a longitudinal Rasch model. It models the probability for an arbitrary subject *i* (*i* = 1 to *n*) to give a positive answer to item *j* (*j* = 1 to *p*) (which is coded as 1) at time *t* (*t* = 0 or 1) as:

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

where $\theta_i^{(t)} \sim N\left(\mu_\theta^{(t)}, \sigma_\theta^2\right)$ is the level of the subject *i* on the latent trait at time *t* and $\delta_j^{(t)}$ is the difficulty parameter of the item *j* at time *t*.

At t_0: $\theta_i^{(t_0)} \sim N(0,1)$, and at t_1: $\theta_i^{(t_1)} \sim N\left(\mu_\theta^{(t_1)}, 1\right)$ with $\mu_\theta^{(t_1)} = 0$ when no true change was simulated (α = 0) and $\mu_\theta^{(t_1)} = -0.2$ where a true change was simulated (α = -0.2), with $cor\left(\theta_i^{(t_0)}, \theta_i^{(t_1)}\right) = r = 0.4\ or\ 0.9$.

At t_0, the vector of difficulty parameters for the five items *j = 1 to 5* was $\delta_{j=1\ to\ 5}^{(t_0)} = (-2,-1,0,1,2)$. At t_1 it was $\delta_{j=1\ to\ 5}^{(t_1)} = (-2,-1,0,1,2)$ when no uniform recalibration response shift was simulated (*ur* = 0), $\delta_{j=1\ to\ 5}^{(t_1)} = (-2,-1,-1,1,2)$ when uniform recalibration response shift was simulated on one item (*ur* = 1) and $\delta_{j=1\ to\ 5}^{(t_1)} = (-2,-2,0,0,2)$ when uniform recalibration response shift was simulated on two items (*ur* = 2). Thus, a one unit decrease in item difficulty was chosen to simulate uniform recalibration.

1.2.b. Simulation procedure

For an arbitrary individual *i*, the first step was to generate latent trait levels at t_0 and t_1 ($\theta_i^{(t_0)}$ and $\theta_i^{(t_1)}$). First, a value for two variables Y_1 and Y_2 were randomly drawn from a standard normal distribution ($N(0,1)$).

Then $\theta_i^{(t_0)}$ and $\theta_i^{(t_1)}$ levels were determined as: $\theta_i^{(t_0)} = x_1 \times 1 + 0$ and $\theta_i^{(t_1)} = (rx_1 + \sqrt{(1-r^2)}x_2) \times 1 + \mu_\theta^{(t_1)}$, with appropriate *r* and $\mu_\theta^{(t_1)}$ values depending on the characteristics of the sample generated.

The second step was to compute for each item *j = 1 to 5* at t_0 and t_1 the probability of giving a positive answer to the item by solving the longitudinal Rasch model with appropriate $\theta_i^{(t_0)}$ and $\theta_i^{(t_1)}$ levels, and $\delta_{j=1\ to\ 5}^{(t_0)}$ and $\delta_{j=1\ to\ 5}^{(t_1)}$ levels depending on the characteristics of the sample generated.

The third step was to get responses to each item *j = 1 to 5* at t_0 and t_1 using the aforementioned probabilities generated. For each item *j* at each time of measurement, a value for an Y_3 variable was drawn from a uniform distribution (*unif(0,1)*) and if $y_3 < P\left(X_{ij}^{(t)} = 1 \middle| \theta_i^{(t)}, \delta_j^{(t)}\right)$, then the response to the item was 1, either it was 0.

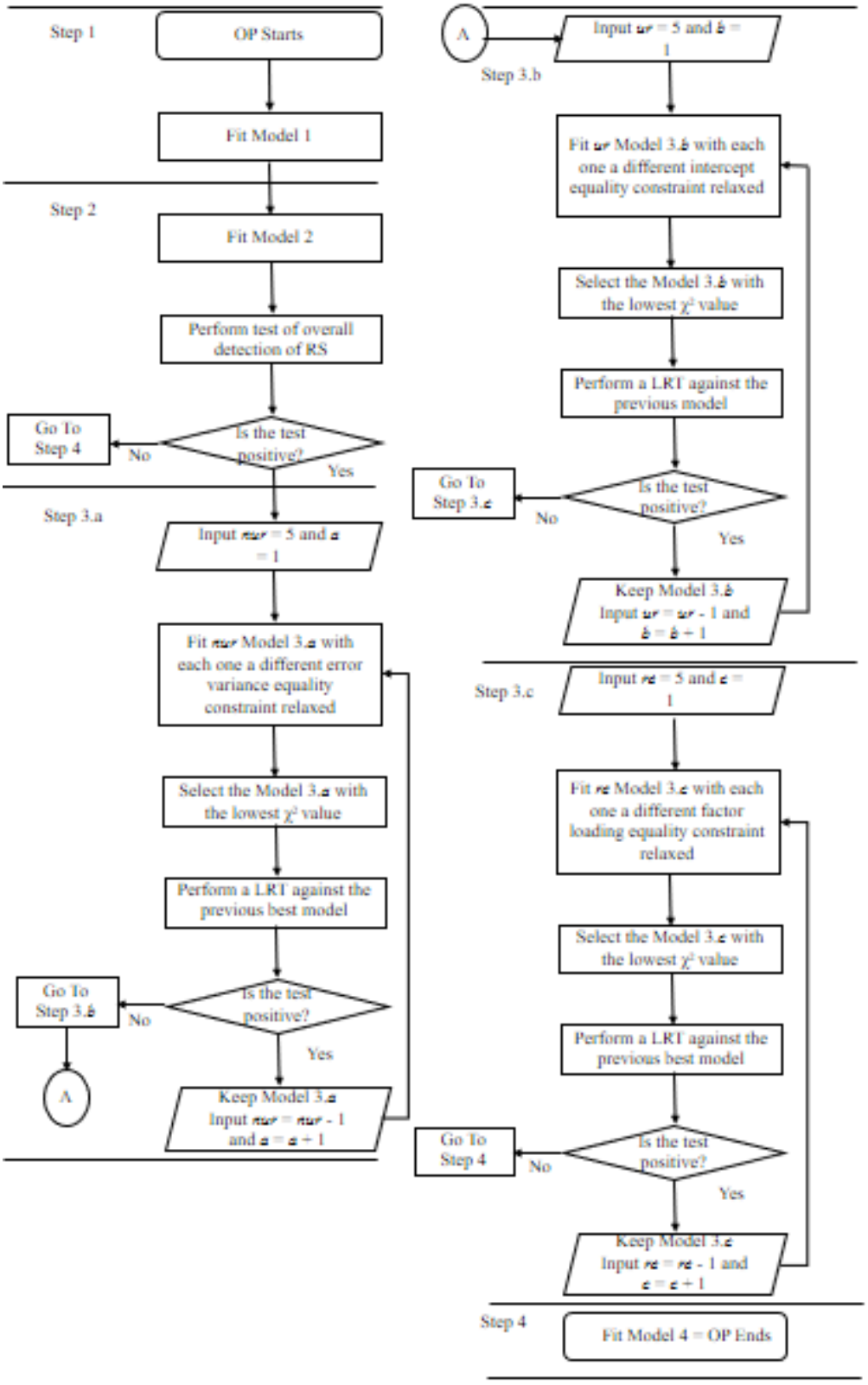
This procedure was then repeated for each subject *1 to n* of a sample and for each of the 36 000 samples generated.

2. How the OP was automatized?

OP is an algorithm composed of four major steps [37]. In short, the first step is dedicated to get a measurement model with adequate structural validity assessed, the second step is dedicated to perform an omnibus test for overall response shift detection, the third step is an iterative procedure dedicated to detect each type(s) of potential response shift on each item(s), and the fourth step is dedicated to fit a final model in order to estimate true change after taking into account response shift detection. OP will be further developed in the core of the manuscript of this study (see Part 4, chapter 3) and comprehensive details about the procedure can be found in the seminal paper [37].

Nonetheless, in this simulation study, OP was used to be performed on 36 000 datasets. So, it was necessary to automatize the procedure using a programming language (R was used [181]). Thus, it seems relevant to display a graphical representation of the automated algorithm that was programmed (Figure 14). It has to be noted the version that was programmed was slightly different of the original OP (a hierarchy in testing the different types of response shift was introduced, for more details see Part 4 chapter 4.2 and Part 6 chapter 1.1).

Figure 14. A graphical representation of how OP was automatized in this simulation study



Original Work: Overall performance of Oort's procedure for response shift detection at item-level: a pilot simulation study

Antoine Vanier[1,2,3], Véronique Sébille[1,4], Myriam Blanchin[1], Alice Guilleux[1], Jean-Benoit Hardouin[1,4]

Published in *Quality of Life Research*. 2015 Aug;24(8):1799-807

1 Bretagne-Loire University, University of Nantes, EA4275-SPHERE "MethodS for Patients-centered outcomes and HEalth REsearch", Nantes, France

2 Sorbonne Universités, UPMC Univ. Paris 06, Department of Biostatistics, Paris, France

3 AP-HP, University Hospitals Pitié-Salpêtrière Charles-Foix, Department of Biostatistics Public Health and Medical Informatics, Paris, France

4 University Hospital of Nantes, Unit of Biostatistics and Methodology, Nantes, France

Abstract

Objective. This simulation study was designed to provide data on the performance of Oort's procedure (OP) for response shift detection (regarding Type-I-Error, power and overall performance), according to sample characteristics, at item-level. A specific objective was to assess the impact of using different information criteria (IC), as alternatives to LRT (Likelihood-Ratio Test), for global assessment of response shift occurrence.

Methods. Responses to 5 binary items at two times of measurement were simulated. Thirty-six combinations of sample characteristics (sample size (*n*), "true change", correlation between the two latent variables and presence/absence of uniform recalibration response shift (*ur*)) were considered. A thousand datasets were generated for each combination. Response shift detection was performed on each dataset following OP. Type-I-Error and power of the global assessment of response shift occurrence, as well as overall performance of the OP was assessed.

Results. The estimated Type-I-Error was close to 5% for the LRT and lower than 5% for the IC. The estimated power was higher for the LRT as compared to the AIC, which was the highest among the other IC. For the LRT, the estimated power for n = 100 and for the combination of n = 200 and ur = 1 item was below 80%. Otherwise, for other combinations of sample characteristics, the estimated power was above 90%.

Conclusion. For the LRT, higher values of power were estimated compared to IC with appropriate values of Type-I-Error. These results were consistent with Oort's proposal to use the LRT as the criterion to assess global response shift occurrence.

3. Introduction

When assessing changes observed over time on a score resulting from a Patient-Reported Outcomes (PRO) instrument, the need to detect potential response shift effects (i.e. a change in the meaning of one's self-evaluation of a target construct over time [30]) that may obfuscate "true change" assessment is well established [15,115]. To do so, various methods have been developed since the late 1990s [135,152]. One of the most attractive methods to detect response shift is Oort's procedure (OP) [37]. OP is based on Structural Equation Modeling (SEM), a statistical modeling technique for testing and estimating different types of causal relations using a combination of quantitative data (i.e. covariance ± mean structures) and qualitative hypotheses [38]. One of the strengths of SEM is the ability to construct latent variables (i.e. variables that are not observed directly, but inferred from several measured variables) [38].

OP allows detection of all forms of response shift (non-uniform and uniform recalibration, reprioritization, reconceptualization) without the need of a specific design [37]. Nonetheless, it implies analyses at group level [37].

OP relies on an operationalization of the different forms of response shift as change(s) in the value of SEM parameters between two times of measurement. These change(s) are the value of error variances for non-uniform recalibration, intercepts for uniform recalibration, and factor loadings for reprioritization [37]. Reonceptualization corresponds to a change in the pattern of factor loadings [37].

OP is an algorithm including four major steps [37]. Each of these steps is associated with a particular longitudinal Confirmatory Factor Analysis (CFA) model. The first step consists in establishing an appropriate measurement model (Model 1) of observed scores at two times of measurement. The second step is a global assessment of response shift occurrence. To do so, a model verifying the hypothesis of no response shift (Model 2) is constructed, and its fit is compared with Model 1 by testing if the difference between the χ^2 values of the two models is statistically significant (χ^2 difference test, also known as Likelihood-Ratio Test (LRT)). If the abovementioned LRT is significant, the fit of Model 2 is worse than Model 1, which is interpreted as a global presence of response shift, and the procedure continues. The third step is performed using an iterative process (by relaxing one constraint at a time) starting from Model 2. It is dedicated to detect all forms of response shift on all potentially affected items

(Models 3). A final model is estimated, in which differences in factor means is indicative of "true change" after accounting for response shift (Model 4).

Since its publication, OP has been successfully used to detect response shift on several clinical datasets, usually at domain-level (i.e. with continuous scores as observed variables) [39,40,43,44,46–48]. However, the performance of the algorithm, regarding the Type-I-Error and statistical power of the global assessment of response shift occurrence, or the overall behavior of the procedure (its ability to detect only truly existing response shift), remain quite unknown. If the performance of SEM to detect measurement bias has already been investigated in previous studies [182–184], the procedures assessed in these studies, although sharing some similarities with OP, are not strictly equivalent. In addition, nothing is known about the performance of OP in the context of detecting response shift at item-level (i.e. with categorical responses as observed variables). Lastly, some methodological choices, like the use of the LRT as a global assessment of response shift occurrence, can be questioned. Indeed, global assessment of response shift occurrence could be achieved using information criteria (IC) instead. IC are designed to help model selection, by summarizing in one numeric value a balance between the information explained by a model and its complexity (parsimony principle). The lowest the value of an IC is, the more parsimonious the model is [185–187]. Therefore, a global presence of response shift would be reflected by an increase in the value of an IC in Model 2 compared to Model 1. Assessing the probabilistic performance of a statistical procedure can be approached by estimating the results that it produces on a large number of simulated datasets, as the values of the parameters (i.e. the values of the sample characteristics) used to generate these datasets are fully determined, and therefore known.

Thus, the main objective of this study was to provide for the first time data on the performance of OP (regarding Type-I-Error, power and overall behavior), at item-level with binary items, via a simulation study. A specific objective was to assess the impact of using different IC, as alternatives to LRT, for global assessment of response shift occurrence.

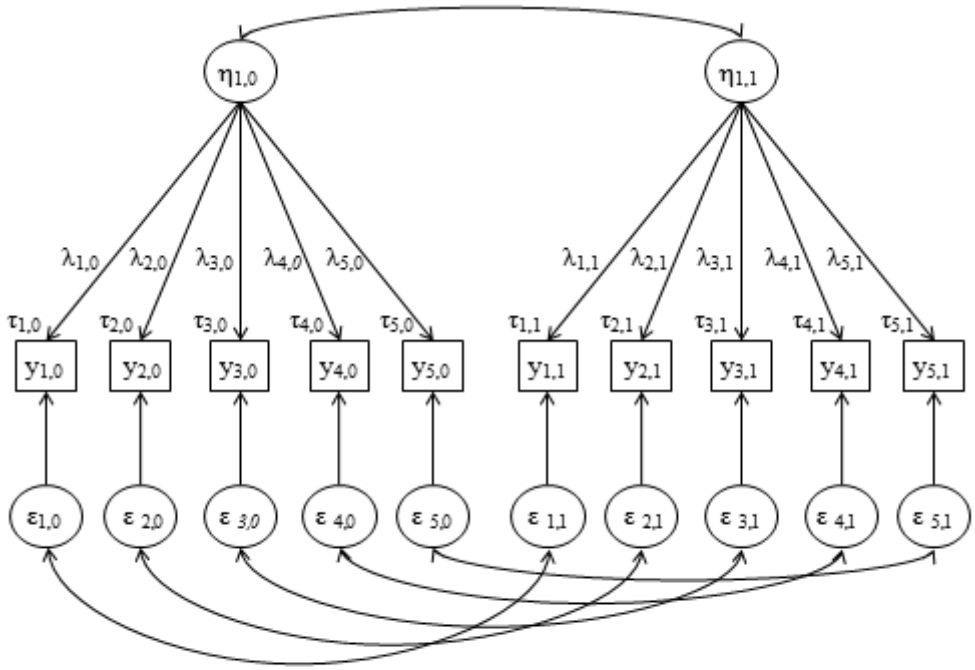
4. Materials and methods

4.1 Simulated datasets

Responses to 5 binary items, at two times of measurement (t_0 and t_1), were simulated. As we chose to investigate the OP at item-level, it appeared to be suited to simulate these responses via a model related to Item Response Theory. So, these responses were generated, as

a function of a latent trait and items difficulties (for each times of measurement), using a longitudinal Rasch model (which has good measurement properties and is commonly used when modeling responses to dichotomous items using the IRT framework) [188]. Thus, the general form of the longitudinal CFA measurement model which was defined to fit Model 1 is of 5 binary items loading on one latent variable at two times of measurement (Figure 15)

. Figure 15 Graphical representation of the general form of the measurement model (Model 1) fitted on the data



Notes: Circles represent latent variables and squares represent observed variables. The measurement model for the observed scores (responses to items) of an arbitrary subject i at time t may be given by: $y_{i,t} = \tau + \Lambda\eta_{i,t} + \varepsilon_{i,t}$, where y are a vector of observed scores, η a vector of unobserved common factor scores and ε a vector of unobserved residual factor scores. Matrix Λ contains factor loadings ($\lambda_{j,t}$ with j the j^{th} item at time t), and vector τ contains intercepts ($\tau_{j,t}$). In the figure, the indices for y, λ, τ and ε are of the form j,t with j the j^{th} item at time t. For η, the indices are of the form k,t, with k the k^{th} common factor score at time t.

As this study was a pilot simulation study and as we chose to simulate data with a Rasch model, when response shift on an item was simulated, it was uniform recalibration only, operationalized as a one unit decrease in item difficulty between t_0 and t_1.

Four types of sample characteristics could vary according to different fixed levels:

5. *n* (sample size) could be fixed at 100, 200 or 300;
6. α (changes in latent trait mean level between the two times or "true change") could be fixed at 0 (no "true change") or -0.2 (a decrease in latent trait mean level between t_0 and t_1);

7. *r* (correlation between latent traits between the two times) could be fixed at 0.4 (moderate correlation) or 0.9 (very strong correlation);
8. *ur* (occurrence of uniform recalibration) could be fixed at 0 item, 1 item (on the third item), or 2 items (on the second and fourth items).

The sample size values were chosen in accordance with sizes usually reported in studies investigating response shift [115]. A small negative effect of the catalyst on latent trait mean level (-0.2) was chosen to reflect plausible effect sizes frequently observed in clinical research. As we hypothesized that the correlation between latent traits between the two times would have a negligible impact on response shift detection, we chose a moderate (0.4) and an extreme value (0.9) to test this hypothesis. A one unit decrease in item difficulty was chosen to simulate uniform recalibration, because we had previously showed in another simulation study (aiming at studying the power of the test of group effect in a Rasch model) that the degree of uncertainty of the item difficulty parameters had to be high (a one unit difference), to observe a moderate impact on power [189].

Thirty-six combinations of the levels of the sample characteristics were investigated. A thousand datasets have been simulated for each combination.

4.2 Response shift detection

Response shift detection was performed on each datasets following the 4 steps of OP [37]. SEM models were fitted using robust maximum-likelihood estimator with a Satorra-Bentler correction (MLM) [100], with lavaan package 0.5-13 [190] for R software 3.0.1 [181].

A Root Mean Square Error of Approximation (RMSEA) close to 0.05 (p of close fit > 0.05) and Comparative Fit Index (CFI) ≥ 0.95 were used as indicators of good fit for Model 1 and 4 [191]. Both of these fit indices were computed using Satorra-Bentler corrected $\chi 2$ values.

Global assessment of response shift occurrence (step 2) was performed with 2 different strategies:

1. a Satorra-Bentler scaled χ^2 difference test (which will be thereafter referred as LRT for simplicity) between Model 2 and Model 1 considered significant if the estimated p-value was below 0.05 [37];
2. an increase in the value of an IC in Model 2 compared to Model 1 (three common IC were investigated in this study: Akaike Information Criterion (AIC),

Bayesian Information Criterion (BIC) and Sample size Adjusted BIC (SABIC)) [90].

If there was global evidence of response shift, untenable constraints on response shift parameters were relaxed one at a time, starting from Model 2 (step 3). Relaxing constraints on error variances (non-uniform recalibration) was performed first, followed by intercepts (uniform recalibration) and factor loadings (reprioritization), thus following a hierarchy in testing the different forms of response shift proposed in two previous studies [41,154]. At each time in step 3, the constraint that was proposed to be relaxed was the one leading to a model with the lowest corrected χ^2 value. Each time, the relevance of relaxing a constraint was tested using a LRT, which was considered significant if the estimated p-value was below 0.05 [192]. Step 3 was performed until relaxing a proposed constraint led to a non-significant LRT.

4.3 Statistical analyses

The Type-I-Error regarding the global assessment of response shift occurrence was estimated as the proportions of datasets where global response shift was evidenced among datasets *where no response shift was simulated*.

Power of the global assessment of response shift occurrence was estimated as the proportions of datasets where global response shift was evidenced among datasets *where response shift was simulated*.

Overall behavior of the procedure was estimated by means of two indicators:

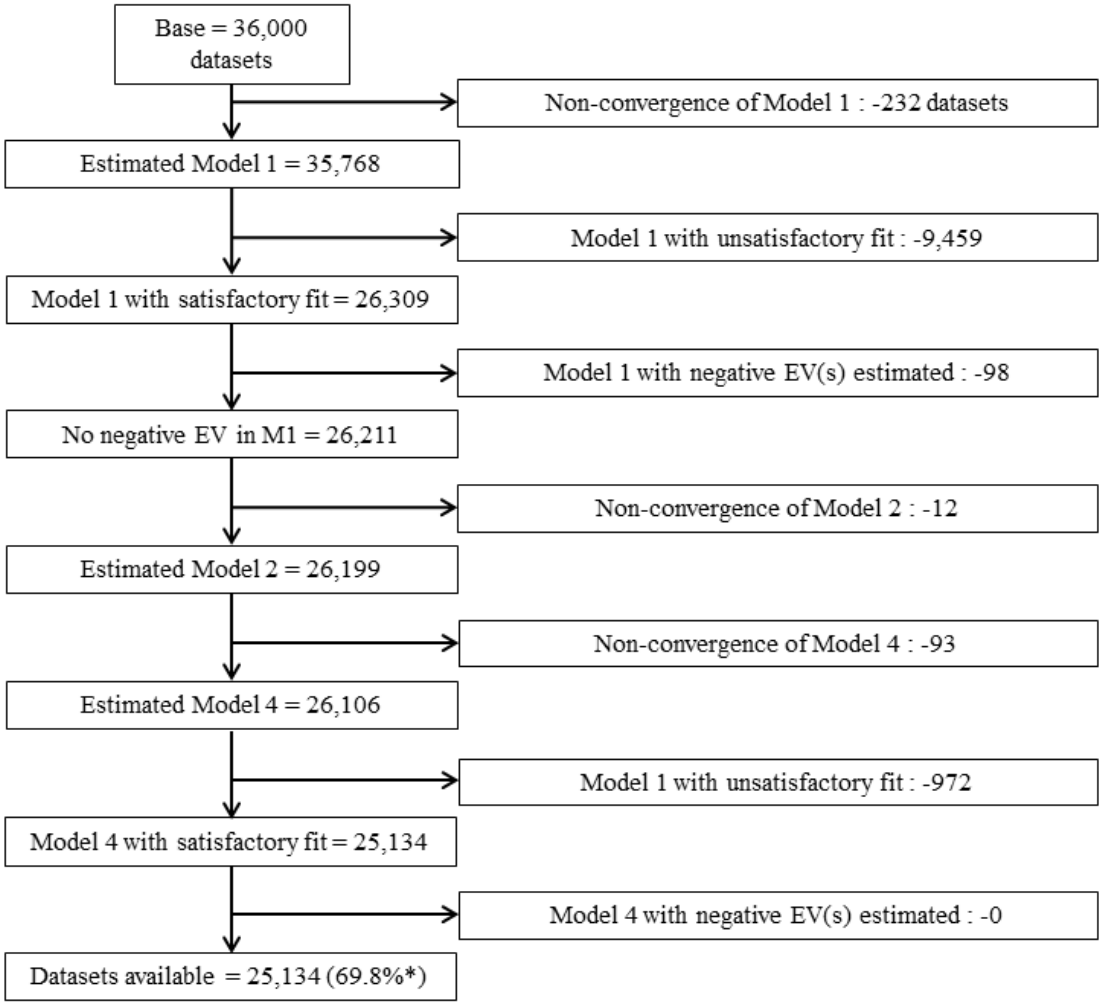
1. *Overall Behavior Indicator 1 (OBI1)*: the assessment of the proportion of datasets for which the whole OP *had properly detected uniform recalibration response shift on only truly affected item(s)* (after a significant LRT ascertaining global response shift occurrence), *disregarding any false detections of response shift on these or one of the other items,* and considering only datasets *where response shift was simulated;*
2. *Overall Behavior Indicator 2 (OBI2)*: this indicator was nearly identical as OBI1, but with an additional requirement of *no false detections of response shift on any item(s).*

Confidence Intervals at a 95% level ($CI_{95\%}$) were estimated for all the aforementioned proportions.

5. Results

5.1 Number of analyzed datasets

Figure 16 . Flow chart of datasets discarded from final statistical analyses



Notes: EV: error variance, M1: Model 1, *denominator = 36,000

As illustrated by Figure 16, analyses were restricted to 25,134 (69.8%) of the 36,000 datasets initially generated. Three main reasons could cause a dataset to be discarded from analyses: (1) the non-convergence of the estimation algorithm when fitting any model of the whole OP; (2) Model 1 or 4 estimated with poor fitting criterion; (3) Model 1 or 4 estimated with any odd parameter(s) (negative error variance) (Figure 16). Most of the 10,866 datasets discarded from analyses were excluded because Model 1 fit (87.0% of these 10,866 datasets), or Model 4 fit (8.9%), wasn't satisfactory.

5.2. Type-I-Error of the global assessment of response shift occurrence (Model 2 vs Model 1)

Table 4 shows estimated Type-I-Error using different strategies for global assessment of response shift occurrence.

Table 4. Estimated Type-I-Error for different strategies for global assessment for RS occurrence (Model 2 vs Model 1)

n	α	r	ur	LRT (p < 0.05)		AIC2 > AIC1		SABIC2 > SABIC1		BIC2 > BIC1	
				%	$CI_{95\%}$	%	$CI_{95\%}$	%	$CI_{95\%}$	%	$CI_{95\%}$
100	**0**	**0.4**	**0**	4.5	[3.1 - 6.7]	0.8	[0.3 - 1.9]	5.1	[3.5 - 7.3]	0.0	[0.0 - 0.7]
		0.9	**0**	3.7	[2.4 - 5.5]	0.2	[0.0 - 0.9]	4.0	[2.7 - 5.9]	0.0	[0.0 - 0.6]
	-0.2	**0.4**	**0**	5.9	[4.1 - 8.2]	0.4	[0.1 - 1.4]	7.0	[5.1 - 9.6]	0.0	[0.0 - 0.7]
		0.9	**0**	5.9	[4.2 - 8.1]	0.7	[0.3 - 1.8]	6.6	[4.8 - 8.9]	0.0	[0.0 - 0.7]
200	**0**	**0.4**	**0**	4.0	[2.8 - 5.8]	0.4	[0.2 - 1.3]	0.1	[0.0 - 0.8]	0.0	[0.0 - 0.6]
		0.9	**0**	3.6	[2.5 - 5.2]	0.8	[0.4 - 1.7]	0.4	[0.1 - 1.1]	0.0	[0.0 - 0.5]
	-0.2	**0.4**	**0**	4.0	[2.8 - 5.7]	0.3	[0.1 - 1.0]	0.3	[0.1 - 1.0]	0.0	[0.0 - 0.5]
		0.9	**0**	5.4	[4.1 - 7.3]	1.4	[0.8 - 2.5]	1.2	[0.6 - 2.2]	0.0	[0.0 - 0.5]
300	**0**	**0.4**	**0**	3.0	[2.0 - 4.4]	0.0	[0.0 - 0.5]	0.0	[0.0 - 0.5]	0.0	[0.0 - 0.5]
		0.9	**0**	4.5	[3.3 - 6.1]	0.5	[0.2 - 1.2]	0.0	[0.0 - 0.4]	0.0	[0.0 - 0.4]
	-0.2	**0.4**	**0**	5.8	[4.4 - 7.7]	0.9	[0.4 - 1.8]	0.0	[0.0 - 0.5]	0.0	[0.0 - 0.5]
		0.9	**0**	4.4	[3.3 - 6.0]	0.5	[0.2 - 1.2]	0.0	[0.0 - 0.4]	0.0	[0.0 - 0.4]

Overall, regardless of the value of *n*, α, or *r*, estimated Type-I-Error for the LRT was close to 5% (5% was included in every $CI_{95\%}$, except for one combination ($n = 300$, $\alpha = 0$, $r = 0.4$)). At $n = 100$, Type-I-Error estimated for SABIC was close to that estimated for LRT. Otherwise, for all the ICs (AIC, BIC and SABIC) and combinations of sample characteristics, Type-I-Error estimated for IC ranged from 0.0 to 1.4%.

5.3 Power of the global assessment of response shift occurrence (Model 2 vs Model 1)

Table 5 shows estimated power using different strategies for global assessment of response shift occurrence.

Table 5. Estimated power for different strategies for global assessment for RS occurrence (Model 2 vs Model 1)

n	α	r	ur	LRT (p < 0.05)		AIC2 > AIC1		SABIC2 > SABIC1		BIC2 > BIC1	
				%	CI 95%	%	CI 95%	%	CI 95%	%	CI 95%
100	**0**	**0.4**	**1**	36.4	[32.3 - 40.7]	12.7	[10.1 - 15.9]	39.1	[35.0 - 43.4]	0.0	[0.0 - 0.7]
			2	57.7	[53.1 - 62.2]	28.8	[24.8 - 33.1]	60.4	[55.8 - 64.8]	0.2	[0.0 - 1.2]
		0.9	**1**	37.8	[34.0 - 41.8]	12.1	[9.7 - 15.0]	39.3	[35.5 - 43.3]	0.0	[0.0 - 0.6]
			2	61.2	[57.1 - 65.1]	33.6	[29.8 - 37.6]	64.4	[60.4 - 68.3]	0.2	[0.0 - 1.0]
	-0.2	**0.4**	**1**	36.6	[32.7 - 40.8]	15.6	[12.8 - 18.8]	38.8	[34.8 - 43.0]	0.0	[0.0 - 0.7]
			2	60.0	[55.6 - 64.2]	28.0	[24.2 - 32.1]	63.6	[59.3 - 67.7]	0.0	[0.0 - 0.8]
		0.9	**1**	41.0	[37.1 - 45.1]	15.0	[12.3 - 18.1]	43.0	[39.0 - 47.0]	0.0	[0.0 - 0.7]
			2	61.1	[57.1 - 65.1]	30.2	[26.6 - 34.1]	63.6	[59.6 - 67.5]	0.2	[0.0 - 1.0]
200	**0**	**0.4**	**1**	72.8	[69.4 - 76.0]	45.0	[41.3 - 48.7]	39.5	[35.9 - 43.1]	0.1	[0.0 - 0.8]
			2	94.6	[92.6 - 96.1]	77.8	[74.5 - 80.8]	71.5	[68.0 - 74.8]	0.3	[0.1 - 1.1]
		0.9	**1**	75.1	[71.8 - 78.0]	44.4	[40.9 - 48.0]	38.3	[34.9 - 41.8]	0.1	[0.0 - 0.8]
			2	94.9	[93.1 - 96.3]	79.9	[76.9 - 82.6]	74.5	[71.2 - 77.5]	0.5	[0.2 - 1.4]
	-0.2	**0.4**	**1**	75.0	[71.7 - 78.1]	44.3	[40.6 - 48.0]	37.6	[34.1 - 41.3]	0.1	[0.0 - 0.8]
			2	92.2	[89.9 - 94.0]	74.4	[71.0 - 77.5]	68.5	[64.9 - 71.9]	0.7	[0.3 - 1.7]
		0.9	**1**	76.2	[73.1 - 79.2]	46.4	[42.9 - 50.0]	41.1	[37.6 - 44.6]	0.0	[0.0 - 0.5]
			2	94.3	[92.5 - 95.8]	80.7	[77.7 - 83.3]	75.6	[72.4 - 78.5]	1.1	[0.5 - 2.1]
300	**0**	**0.4**	**1**	92.3	[90.3 - 94.0]	75.4	[72.3 - 78.2]	50.4	[47.0 - 53.9]	0.1	[0.0 - 0.7]
			2	99.6	[98.9 - 99.9]	95.5	[93.8 - 96.7]	85.8	[83.2 - 88.1]	3.1	[2.1 - 4.6]
		0.9	**1**	94.1	[92.3 - 95.5]	78.3	[75.5 - 81.0]	52.5	[49.2 - 55.8]	0.1	[0.0 - 0.7]
			2	99.6	[98.9 - 99.9]	95.8	[94.2 - 97.0]	86.9	[84.5 - 89.1]	4.6	[3.3 - 6.2]
	-0.2	**0.4**	**1**	92.2	[90.1 - 93.8]	74.7	[71.6 - 77.6]	49.8	[46.3 - 53.3]	0.0	[0.0 - 0.5]
			2	99.5	[98.7 - 99.8]	96.8	[95.3 - 97.8]	86.3	[83.7 - 88.5]	3.9	[2.7 - 5.4]
		0.9	**1**	92.3	[90.3 - 93.9]	73.3	[70.3 - 76.1]	47.1	[43.8 - 50.4]	0.2	[0.1 - 0.8]
			2	99.8	[99.2 - 99.9]	96.6	[95.1 - 97.6]	88.5	[86.3 - 90.5]	4.2	[3.1 - 5.8]

For n = 100, estimated power for SABIC was slightly higher than that estimated for LRT. Otherwise, regardless of the value of α, r or ur, estimated power was higher for LRT than that estimated for AIC, which was the highest among the other IC.

Two sample characteristics were associated with a substantial increase in estimated power, regardless of the assessed criteria (LRT or IC): an increase in sample size (n), and an increase in the number of items affected by uniform recalibration (ur). For LRT, an increase in r was associated with a slight increase in estimated power, especially for n = 100.

For all assessed criteria, estimated power for n = 100 and the combination of n = 200 and ur = 1 was below 80%. Otherwise, for other combinations of sample characteristics, estimated power for LRT was above 90%. Estimated power for BIC was always below 5% and for most of the sample characteristics combinations close to 0%.

5.4 Overall performance of the OP

5.4.a. OBI1

Table 6 shows estimated OBI1 (with LRT as the only strategy investigated for global assessment of response shift occurrence) according to the different combinations of sample characteristics.

Table 6. Estimated OBI1 and OBI2 (see Materials and Methods for definition) as a function of sample characteristics

n	α	r	ur	OBI1		OBI2	
				%	CI 95%	%	CI 95%
100	**0**	**0.4**	**1**	26.2	[22.6 - 30.2]	4.5	[3.0 - 6.7]
		0.9	**1**	27.1	[23.6 - 30.8]	6.6	[4.8 - 8.8]
	-0.2	**0.4**	**1**	27.5	[23.9 - 31.4]	6.6	[4.8 - 9.0]
		0.9	**1**	28.3	[24.8 - 32.2]	6.6	[4.9 - 8.9]
200	**0**	**0.4**	**1**	58.4	[54.7 - 62.0]	11.2	[9.1 - 13.7]
		0.9	**1**	59.7	[56.2 - 63.2]	13.9	[11.6 - 16.5]
	-0.2	**0.4**	**1**	60.5	[56.8 - 64.1]	16.7	[14.1 - 19.6]
		0.9	**1**	60.9	[57.4 - 64.4]	18.5	[15.9 - 21.5]
300	**0**	**0.4**	**1**	75.5	[72.4 - 78.4]	11.0	[9.0 - 13.3]
		0.9	**1**	74.5	[71.5 - 77.3]	13.6	[11.5 - 16.1]
	-0.2	**0.4**	**1**	75.2	[72.1 - 78.1]	16.1	[13.7 - 18.8]
		0.9	**1**	75.5	[72.5 - 78.2]	18.3	[15.9 - 21.0]
100	**0**	**0.4**	**2**	21.9	[18.3 - 25.9]	2.4	[1.4 - 4.3]
		0.9	**2**	27.2	[23.7 - 31.0]	3.7	[2.5 - 5.6]
	-0.2	**0.4**	**2**	28.2	[24.4 - 32.3]	4.0	[2.6 - 6.1]
		0.9	**2**	28.8	[25.2 - 32.7]	3.4	[2.2 - 5.2]
200	**0**	**0.4**	**2**	52.6	[48.8 - 56.4]	1.3	[0.7 - 2.5]
		0.9	**2**	57.7	[54.2 - 61.2]	2.1	[1.3 - 3.4]
	-0.2	**0.4**	**2**	57.4	[53.6 - 61.0]	6.5	[4.9 - 8.6]
		0.9	**2**	64.9	[61.5 - 68.2]	4.9	[3.5 - 6.6]
300	**0**	**0.4**	**2**	62.5	[59.0 - 65.8]	0.6	[0.3 - 1.5]
		0.9	**2**	69.8	[66.6 - 72.8]	0.8	[0.4 - 1.7]
	-0.2	**0.4**	**2**	69.2	[65.9 - 72.3]	2.7	[1.8 - 4.1]
		0.9	**2**	72.5	[69.5 - 75.4]	2.5	[1.7 - 3.8]

Note: OBI1: the proportion of datasets for which the whole OP *had properly detected uniform recalibration response shift on only truly affected item(s), disregarding any false detections of response shift on these or one of the other items*, OBI2: this indicator was nearly identical as OBI1, but with an additional requirement of *no false detections of response shift on any item(s)).*Global assessment of response shift occurrence (Model 2 versus Model 1) was performed using a Satorra-Bentler scaled χ^2 difference test.

The estimated proportion of datasets for which the whole OP had properly detected uniform recalibration, either on only the third item (*ur* = 1), or only on the second and fourth item (*ur* = 2), ranged from 21.9 to 75.5%, mostly according to sample size (*n*). Indeed, that proportion increased as sample size increased for both *ur* = 1 and *ur* = 2. The increase in estimated OBI1 was moderately lower when *ur* = 2 and *n* = 200 or 300 compared to *ur* = 1 and *n* = 200 or 300.

5.4.b. OBI2

Table 6 shows estimated OBI2 according to the different combinations of sample characteristics.

Overall, the estimated proportion of datasets for which the whole OP had properly detected uniform recalibration on affected item(s), and had appropriately not indicated occurrence of whatever other form(s) of response shift on any item(s), ranged from 0.6 to 18.3%. That estimated proportion was substantially lower than that estimated via OBI1 indicator. Estimated proportion via OBI2 indicator decreased as the number of simulated items affected by uniform recalibration (ur) increased.

6. Discussion

Regarding global assessment of response shift occurrence, the main results of this study were:

- overall, estimated Type-I-Error for the LRT was close to 5% but substantially lower for IC (except for SABIC at *n* = 100 for which estimated Type-I-Error was close to that estimated for LRT);
- estimated power for LRT was below 80% for *n* = 100 and for the combination of *n* = 200 and *ur* = 1, otherwise power was above 90%;
- overall, estimated power for LRT was higher than for IC (except for SABIC at *n* = 100, for which estimated power was moderately higher).

Regarding the overall performance of the procedure, the main results of this study were:

- the whole OP properly detected uniform recalibration on only affected item(s) (OBI1) on 21.9 to 75.5% of the datasets, that proportion increased mostly according to sample size (*n*);

– overall, the whole OP properly detected uniform recalibration only on appropriate item(s), and did not indicate occurrence of whatever other form(s) of response shift on any item(s) (OBI2), on 0.6% to 18.3% of the datasets.

6.1 Number of analyzed datasets

In this study, 29.2% of the datasets were discarded from analyses because of unsatisfactory fit, or occurrence of negative error variance(s). Nonetheless, there are solutions in practical setting, to deal with these issues when analyzing a real dataset (these solutions weren't implemented in our simulation framework, due to the huge number of datasets to analyze). For example, adding correlations paths between some residual factors, if, for instance, the hypothesis of local independence does not hold, can greatly improve model fit. In addition, dealing with negative error variance(s) can be done by choosing different starting values.

6.2 Global assessment of response shift occurrence

In this study, estimated Type-I-Error for the LRT was close to 5%. With normally distributed continuous variables, the test statistic of a LRT between two nested SEM models is assumed to follow a χ^2 distribution under the null hypothesis, with a number of degrees of freedom (*df*) equal to the difference in freely estimated parameters between the two models [191]. Here, we have worked with binary items, but we have corrected the test statistic according to SATORRA-BENTLER proposal [100]. If this correction was adequate, as the LRT was considered significant if the estimated p-value was below 0.05, it was expected to observe Type-I-Error for the LRT close to 5% [193]. The results have matched this expectation.

Estimated Type-I-Error using IC was lower compared to LRT. Comparison of an IC between two SEM models does not constitute statistical hypothesis testing in a formal way [194]. Therefore, in theory, it wasn't expected that the estimated Type-I-Error using IC had to be around any specific value (and especially 5%). Model 2 is formally a simpler model than Model 1: it is nested in Model 1, and in this study, Model 2 has 13 more *df* than Model 1. As stated before, IC are criteria designed to help evaluating model parsimony [185–187]. When no response shift was simulated, it was consistent that the value of Model 2 IC was lower than for Model 1, for almost every dataset. Indeed, in that case, Model 2 adequately respected the parsimony principle. Estimated Type-I-Error was the lowest for BIC. This result was consistent with the fact that compared to AIC and SABIC, BIC is constructed to penalize complexity the

most [186,194]. Estimated Type-I-Error for SABIC was the highest among the IC for *n* =100. Again, this was consistent with the fact that at *n* = 100, the added penalty for each supplementary freely estimated parameter to the log-likelihood of a SEM model is lower for SABIC than for AIC and BIC [187,194]. However, at *n* = 200 and *n* =300, for SABIC, this aforementioned penalty is between AIC and BIC [187,194].

In this study, estimated power for LRT was below 80% for n = 100, and above 90% when the sample size was at least equal to 200 (with *ur* at least equal to 2 when *n* =200). Except for SABIC when *n* = 100, the aforementioned estimated power for LRT was the highest, compared to IC. This result could reflect a tendency of IC to be too conservative compared to LRT for a global assessment of response shift occurrence via SEM. As Model 2 is the simplest in terms of number of freely estimated parameters, it was more often considered as the most appropriate fitting model when comparing Model 2 and Model 1 using IC as compared to LRT. This seemed to be particularly the case for BIC, with estimated power close to 0% for most of the sample characteristics combinations assessed in the study.

Overall, as estimated Type-I-Error for LRT was indeed close to the theoretically expected 5%, and as estimated power for LRT was the highest, these results are consistent with Oort's proposal to use LRT between Model 2 and Model 1 as the criterion to assess global response shift occurrence [37], rather than IC.

6.3 Overall performance of the OP

In this study, the estimated proportion of datasets for which the whole OP had properly detected uniform recalibration, either on only the third item (*ur* = 1), or only on the second and fourth item (*ur* = 2) (OBI1 indicator), ranged from 21.9 to 75.5%. That estimated proportion increased mostly with sample size. These results seem to indicate as long as the LRT between Model 2 and Model 1 is significant, the procedure correctly detects uniform recalibration on appropriate item(s) in most of the cases. However, when we consider the fact the procedure must not only detect uniform recalibration on appropriate item(s), but it should also avoid detecting other form(s) of response shift on any item(s) (OBI2 indicator), the resulting estimated proportion decreased compared to OBI1 indicator, and ranged from 0.6 to 18.3%. For *ur* = 1, the procedure had detected non-uniform recalibration on only one item in 30.5 to 56.9% of the datasets according to sample characteristics. In most cases, the item detected was the same that the item on which uniform recalibration was simulated. For *ur* = 2, the procedure had detected

non-uniform recalibration on at least 1 item in 51.6 to 92.5% of the datasets. Again, in most cases, those item(s) (was) were the one(s) on which uniform recalibration were simulated.

A first explanation to this phenomenon can be linked to the simulation process of uniform recalibration coupled with the fact this was a work on binary items. Indeed, uniform recalibration was simulated using a longitudinal Rasch model by a change in the value of difficulties across time. On a binary item, this can be equated with a change in the proportion of positive responses (P). In SEM models, non-uniform recalibration is detected by a change in the value of error variance. Error variances are linked with the variance of the item, which is represented for binary items by P(1 – P). Therefore, when uniform recalibration was simulated on binary items, it seems plausible non-uniform recalibration might have been simulated too.

Nonetheless, another explanation to the aforementioned phenomenon can reflect the issue regarding the need to introduce, or not, at step 3 of the procedure, a hierarchy in testing for different forms of response shift [41,154]. In this study, a hierarchy proposed in two previous studies was followed [41,154], which consists in testing non-uniform recalibration first, followed by uniform recalibration and finally reprioritization. This hierarchy was derived from measurement invariance studies [195]. So, in SEM operationalization, we can hypothesize that when an item is affected by uniform recalibration, it also sometimes operationalizes as contingent non-uniform recalibration, which is detected first when the abovementioned hierarchy is followed. If this hypothesis holds, it could advocate against the need to impose a hierarchy. Indeed, if uniform recalibration was allowed to be detected first, then maybe it would correct for the risk of detecting contingent non-uniform recalibration. Thus, if the aforementioned hierarchy had not been imposed, perhaps estimated OBI2 indicator would have been higher.

6.4. Limits

This study suffered from some limits. The main one is the method used to estimate SEM parameters. Theoretically, working with binary items requires estimating matrices of tetrachoric correlations alongside with the use of robust-Diagnonally Weighted Least Squares (DWLS) estimator [99]. However, this method imposes to estimate thresholds instead of intercepts and requires more identifiability constraints (known as delta or theta parameterizations) [196]. Currently, the operationalization of the response shift detection (especially for non-uniform and uniform recalibration) used in the OP is not adapted to work

with DWLS [37]. Thus, we used covariance analyses with robust maximum-likelihood. So, although we performed a Satorra-Bentler correction, which seemed to have corrected for the risk of a biased LRT (as illustrated by the fact the Type-I-Error is consistent with the 5% theoretically expected), SEM parameters are probably somehow biased which could have affected OBI1 and OBI2 values.

The second main limit is the scope of the study. This study was a pilot simulation study, and simulation studies are usually consuming in terms of computational resources. Therefore, we have chosen to restrict our work to binary items, we also have only simulated uniform recalibration response shift and a unique simple structure (5 items loading on one dimension). Thus, if the results give some clues about how OP behaves at item-level with unidimensional model when detecting uniform recalibration response shift, they cannot be easily extrapolated to other settings (polytomous items or continuous scores, other types of response shift). In particular, the results cannot be easily generalized to other practical settings in HRQL measurement, like multidimensional instruments with many items.

In addition, we have simulated, using Rasch models, uniform recalibration always with the same magnitude. Although we have empirical data to support the fact that this value was of a sufficient magnitude to represent a significant uniform recalibration effect [189], it remains an uncertainty about what this value represents in SEM. For instance, if it was too low to simulate such effect, it could have a negative impact on the results.

Lastly, we did not investigate in that study other relevant issues related to the OP: like the aforementioned issue of the need, or not, of a hierarchy in step 3, or the need, or not, to correct for multiple hypothesis testing [42].

7. Conclusion

This study proposed for the first time results on the probabilistic behavior of OP at item-level, in terms of Type-I-Error, power and overall performance via a simulation study. The results were consistent with Oort's proposal to use the LRT as a criterion for global assessment of response shift occurrence. However, several issues about the most efficient way to conduct response shift detection via OP can still be discussed. Moreover, the results of this study are limited by some choices that were made. New simulation studies could be performed to investigate the aforementioned limits. Lastly, that study also emphasizes the need to properly adapt the OP to item-level analyses.

Acknowledgments

This study was supported by the Institut National du Cancer (France), under reference “INCA_6931”.

Part 5: Could response shift be associated to the semantic complexity of PROs and therefore be prevented? A theoretical approach and proposal

"In theory, there are no differences between theory and practice, in practice there are" – Common sense from an unknown source

Original Work: Semantic primes theory may be helpful in designing questionnaires such as to prevent response shift

Antoine Vanier[1,2,3,4], Alain Leplège[5], Jean-Benoit Hardouin[2], Véronique Sébille[2], Bruno Falissard[1,6,7]

Published in *Journal of Clinical Epidemiology*. 2015 Jun;68(6):646-54

1 INSERM, U1178 "Mental Health and Public Health", Paris, France

2 Bretagne-Loire University, University of Nantes, EA4275-SPHERE "MethodS for Patients-centered outcomes and HEalth REsearch", Nantes, France

3 UPMC Univ. Paris 06, Department of Biostatistics, Paris, France

4 AP-HP, University Hospitals Pitié-Salpêtrière Charles-Foix, Department of Biostatistics Public Health and Medical Informatics, Paris, France

5 Univ. Paris Diderot, Sorbonne Paris Cité, Department of History and Philosophy of Sciences, Paris, France

6 Univ. Paris-Sud and Univ. Paris-Descartes, Paris, France

7 AP-HP, University Hospitals Paris-Sud, Department of Public Health, Le Kremlin-Bicêtre, France

Abstract

Objective. The purpose of Randomized Control Trials (RCTs) can be the assessment of the direct effect of treatment on Health-Related Quality of Life (HRQL). Response shift theory considers that a change in HRQL scores observed over time cannot be explained solely by a direct effect of a medical condition, it may also result from a change in the way people appraise their HRQL. The response shift effect is a potential bias that is liable to compromise efficient assessment of the effect of treatment on HRQL.

Study design and setting. We hypothesize a link between the response shift effect on HRQL scores, and the level of complexity of HRQL conceptualization.

Results. We discuss how the impact of reconceptualization on scores depends on the complexity of the linguistic definition of a subjective construct, and how for reprioritization the impact depends on the dimensionality. The linguistic theory of semantic primes is used to help identify how subjective constructs can be classified according to the complexity of their definitions.

Conclusion. Finally, we suggest that the impact of the response shift effect on HRQL scores could be avoided (or lessened) if questionnaires were designed with a rule of "the least semantic and psychometric complexity" in mind.

1. Introduction

1.1. The evaluation of the patient perspective in health-related research

Patient-Reported Outcomes (PRO) are now widely used in health-related research, some of them to assess Health-Related Quality of Life (HRQL), usually via self-administered questionnaires [18]. In many medical areas (e.g. oncology, palliative care…), HRQL is measured over time to add relevant information on patients' subjective experience in the course of treatment, to counterbalance objective data such as survival time [116]. Indeed, improvement or deterioration of outcomes like health status or symptom levels is not always correlated with patients' subjective experience [14].

The current generation of HRQL measures is based on the assumption that the meaning of concepts and measurement scales remains stable in individuals' minds over time and is similar between groups [34]. Thus, HRQL scale scores are assumed to be directly comparable for a given individual over time [29]. As illustrated by what were initially called "paradoxical and counter-intuitive findings" from the 1980s and 1990s [30] (e.g. reports of stable HRQL levels over time by patients with a life-threatening disease [31], reports of better levels of HRQL by patients with advanced stages chronic illness than by others [197]…), these assumptions can be challenged. Indeed, these abovementioned findings were interpreted as evidence that respondents understand the same questions differently over time [30,117], a phenomenon which is now known as *response shift*.

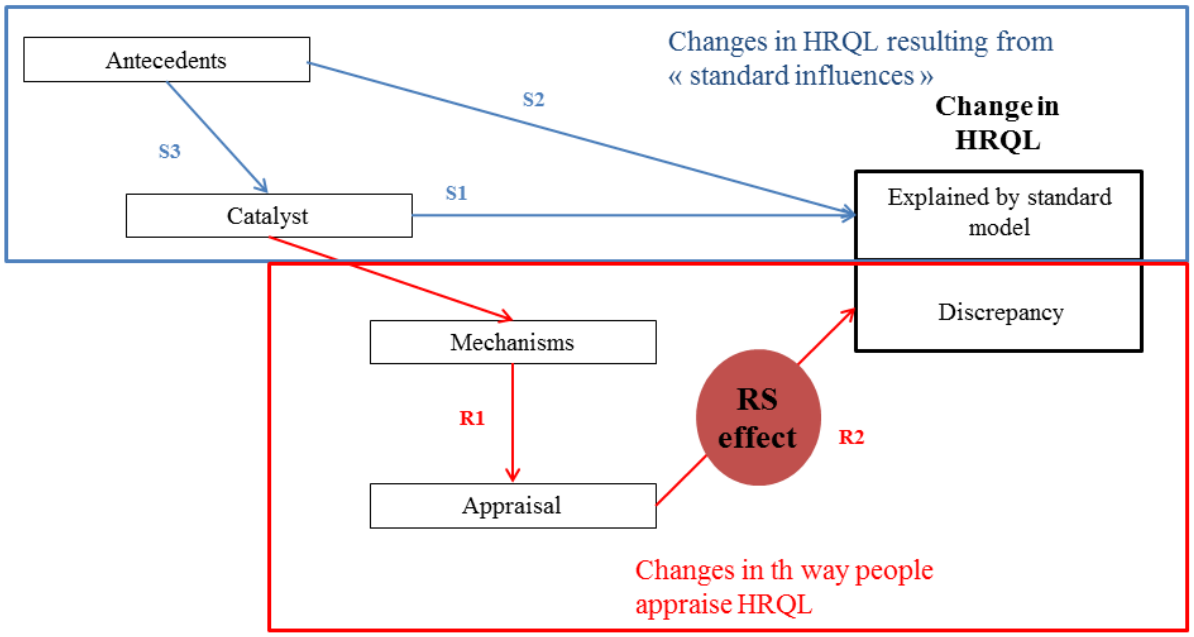
1.2. A brief overview of response shift theory

In health-related research, response shift was defined in 1999 as "*a change in the meaning of one's self-evaluation of a target construct*" [30]. It is operationalized in three forms:

- *recalibration*, which is a change in the respondent's internal standards of measurement (e.g. a person suffering from chronic pain and rating it on a pain scale as 7/10 will later rate it as 5/10 after experiencing acute pain, despite the chronic pain being the same as before);
- *reprioritization*, which is a change in the respondent's values (i.e. the relative importance of component domains in the target construct - e.g. an athletic person who considers physical functioning as an important part of his/her HRQL may later place emphasis on social functioning after sustaining permanent physical injury);

– *reconceptualization*, which is the redefinition of a target construct (e.g. an item of a multidomain questionnaire initially assessing the domain of mental health, will be later understood by the respondent as assessing another domain, like social functioning) [30].

Figure 17. Theoretical model of relationships between "standard influences", response shift and changes in HRQL *(Adapted from RAPKIN and SCHWARTZ [10])*



Response shift effect is assumed to result from psychological mechanisms that individuals use to deal with life changes, triggered by change in health state (a "catalyst") [30,118]. As illustrated by Figure 17, when someone is affected by a catalyst (e.g. occurrence of a chronic disease, initiation of chemotherapy…), this catalyst can have a direct effect on HRQL (S_1 pathway in Figure 17), translating into a change in the person's HRQL assessment. A person's background (e.g. socioeconomic status, personality traits…) can also have a direct effect on HRQL (S_2 pathway) or an effect mediated by the catalyst (S_3 pathway). These effects can be called "standard influences" on HRQL. However, the catalyst can also induce psychological mechanisms (e.g. coping strategies, social comparison…), leading into changes in the way that a person understands and appraises HRQL (R_1 and R_2 pathways), and hence affecting his/her observed scores: response shift has occurred.

Thus, there is a need to disentangle response shift effect from the effects of the "standard influences" on HRQL [135]. Various methods have been used to detect response shift

[37,145,147,148,152,153,157,160,179,198], although there was in the last years a focus on methods based on Structural Equation Modeling. Indeed, some recent works were helpful in showing how response shift effect can affect observed scores or true attributes (i.e. HRQL itself) and how this effect can be modeled in explaining changes in observed scores or HRQL over time [145,159]. Occurrence of response shift has now been documented in a variety of medical conditions [115].

1.3. The particular context of clinical trials

HRQL measurement is increasingly used in the context of Randomized Control Trials (RCTs), as endpoints [199]. In this context, a questionnaire should be designed to allow clear interpretation of the direct effect of treatment on patients' subjective experience. Indeed, it serves as a criterion to enable a decision to be reached between two mutually-exclusive options (i.e. the treatment assessed is effective or not) [59].

The initiation of treatment in the different arms of a RCT can be the catalyst of response shift effect. For example, experiences of extreme levels of fatigue after chemotherapy can induce recalibration response shift when assessing fatigue [123]. Therefore, if response shift has occurred, changes in HRQL scores observed over time cannot be solely explained by a direct effect of the treatment. Moreover, the quality and quantity of response shift effect are dependent of the nature of the catalyst (e.g. as an extreme case, response shift is not likely to occur when using placebo in an open label study). Thus, when assessing different treatment options in a RCT, response shift effect can affect in varying amounts changes in scores in each arms of the trial. As response shift can be triggered by the initiation of treatment (after randomization), it might not be randomly distributed between groups and cannot be equated with measurement error.

Thus, response shift effect in a RCT is liable to confuse interpretation of changes in HRQL scores observed over time, thus making conclusions difficult.

2. Hypothesis

This paper is positioned in the context of the need to measure patients' subjective experience as an endpoint in RCTs. A well-designed questionnaire for use in this context should generate a score that enables simple and clear interpretation of a change over time. It should

therefore not be biased by response shift, as response shift is not usually the effect being investigated.

Our hypothesis is that there is a link between response shift effect on HRQL scores (interfering with the effects of "standard influences"), and the level of complexity of HRQL conceptualization (both in semantic and psychometric terms). Indeed, to allow a difference in interpretability over time or between people, the construct being measured has to be defined itself with a sufficient level of complexity.

More precisely, we will discuss below how the impact of reconceptualization on scores depends mostly on the complexity of the definition of a subjective construct (in semantic terms), and how for reprioritization it depends mostly on construct dimensionality (in psychometric terms).

We will then suggest that the impact of response shift on the measurement of patients' subjective experience could be avoided (or at least lessened) if subjective constructs and questionnaires were conceptualized and designed with these linguistic and psychometric considerations in mind.

3. Polysemy as the original sin of HRQL

3.1. Linguistics and semantic primes

In the general framework of survey development, the impact of the wording and phrasing of the questionnaire in the response process is discussed. TOURANGEAU et al. considers that a person who is responding to a survey has to go through four cognitive processes (comprehension, retrieval, judgment and response) [200]. In this model of survey response, it is stressed out the first process involved is accurate comprehension of the scope of the questionnaire. Thus, certain characteristics of the items, including semantic ones, are acknowledged as leading to difficulties in understanding precisely what is the subjective construct being measured by the questionnaire [200].

This can be linked to the more specific issue of response shift in measuring patients' subjective experience via PRO instruments. Indeed, a PRO instrument can be viewed as a way to communicate from a researcher to a patient [53]. A researcher, through the questionnaire, is aiming to ask patients to rate certain subjective aspects of their lives. response shift theory suggests this communication can be distorted, because the construct being measured can be

interpreted in different ways [53]. Although response shift has mostly been investigated in HRQL [135], it could occur with any subjective construct researchers might want to measure (e.g. it has been investigated in relation to fatigue [201]).

This raises the question of whether any subjective construct can be interpreted in different ways by different people (or by one and the same person over time). Or, at least, the extent to which a subjective construct can be interpreted in different ways: i.e. semantics. In the broadest way, semantics is the question of the relations between signifiers and what they stand for. Of course, this question has been studied extensively over the course of human history. It has penetrated many academic fields, like linguistics, philosophy, even cognitive sciences, and it seems that there is currently no definitive consensus about it [202]. However, one of the main semantic theories based on the notion of *semantic primes* and developed since the 1970s by Anna WIERZBICKA and colleagues [203], seems to be particularly relevant in the context of this work.

According to Anna WIERZBICKA and colleagues, defining every concept used in a language is feasible starting from a core set of primitive concepts: *semantic primes*. Semantic primes are core elements that can be used to coherently represent all complex meanings. They can be viewed as analogous to basic chemical elements from which all chemical compounds can be synthesized [204]. Without a set of primitives, any descriptions of meaning are actually or potentially circular [203].

The search to identify potential semantic primes in languages was based on two main criteria. First, a concept is considered to be a semantic prime if, after extensive trial-and-error lexical-conceptual analysis, the concept cannot be reduced to simpler concepts that define it (reductive paraphrase: it is impossible to define the concept in another way than by paraphrasing it) [203,204]. In addition, a concept can be considered as a semantic prime if, after extensive cross-cultural studies in a wide range of languages, the semantic prime exists as a linguistic exponent (word or word-like element) in all languages [203,204]. Thus semantic primes represent innate concepts, with a meaning assumed to be universal (i.e. shared by every culture and language) [203].

To date, 64 concepts have been confirmed to be semantic primes (listed in Table 7) [205].

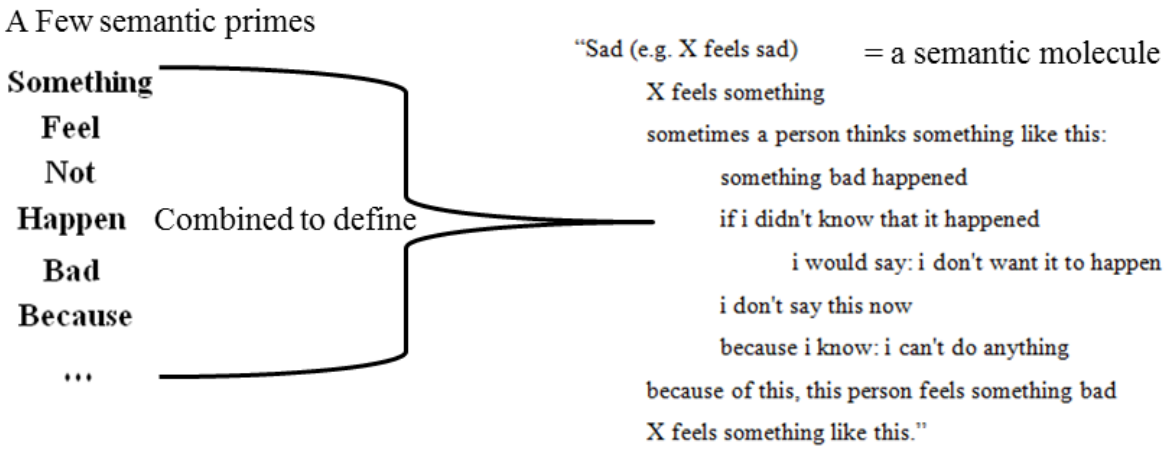
Table 7. List of semantic primes (English exponents), grouped into related categories, *according to GODDARD* **[205]**

Semantic primes	Categories
I, You, Someone, Something~Thing, People, Body	substantives
Kind, Part	relational substantives
This, The same, Other~Else	determiners
One, Two, Much~Many, Little~Few, Some, All	quantifiers
Good, Bad	evaluators
Big, Small	descriptors
Know, Think, Want, Feel, See, Hear	mental predicates
Say, Words, True	speech
Do, Happen, Move, Touch	actions, events, movement, contact
Be (Somewhere), There is, Have (Something), Be (Someone/Something)	location, existence, possession, specification
Live, Die	life and death
When~Time, Now, Before, After, A long time, A short time, For some time, Moment	time
Where~Place, Here, Above, Below, Far, Near, Side, Inside	space
Not, Maybe, Can, Because, If	logical concepts
Very, More	intensifier, augmentor
Like~As~Way	similarity

Looking at Table 7, it seems none of these semantic primes are strictly correlated with a subjective construct used in health-related research. We cannot therefore postulate that there are subjective constructs relevant to biomedical research that have a universal and univocal meaning. Nevertheless, semantic primes theory describes different levels of complexity in the definition of concepts.

A very low level of complexity, just above the notion of semantic prime, is represented by concepts that can be described using a set of only a few semantic primes. These concepts can be understood as *semantic molecules* [203]. According to Anna WIERZBICKA and colleagues, certain concepts related to emotions or feelings can be defined as semantic molecules. As shown in Figure 18, "sadness" can be defined using a few semantic primes combined by means of a universal basic syntax (the "Natural Semantic Metalanguage" [203]).

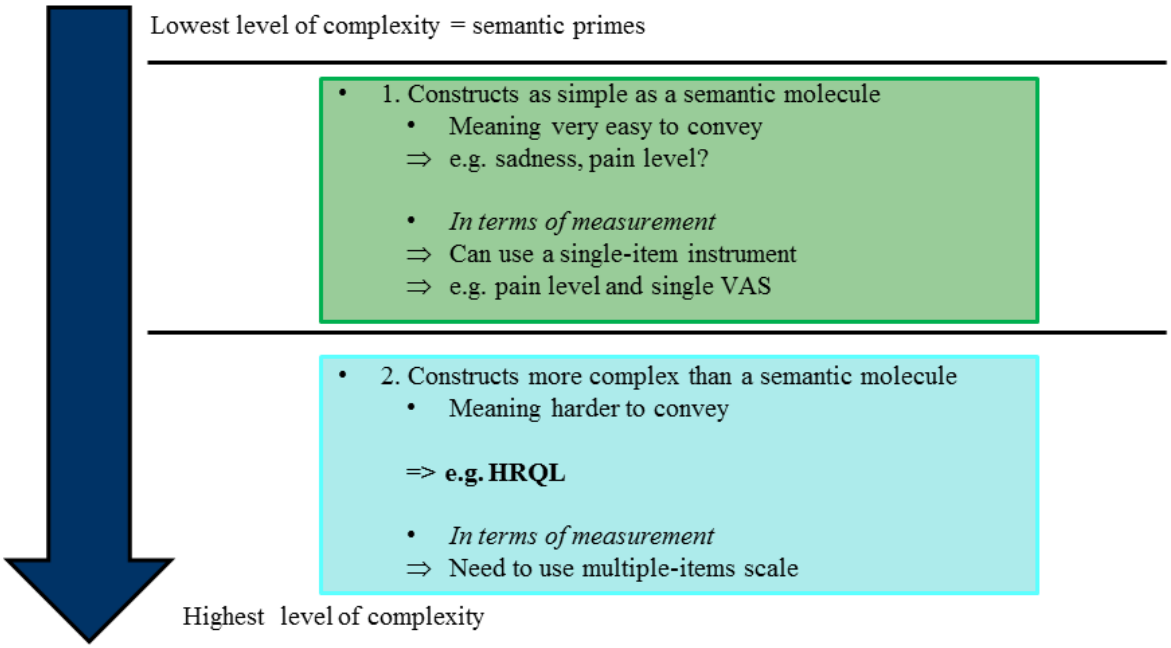
Figure 18 . Example of how a few semantic primes are combined to define the semantic molecule "sad", according to semantic primes theory *(Adapted from WIERZBICKA [31])*



A higher level of complexity in the definition of concepts is reached when it is at least necessary to define a concept step-by-step, i.e. when first a semantic molecule needs to be formed to define a more complex concept. For instance, to define "face", first the definition of the semantic molecule "head" is required. "Face" will then be defined using the semantic molecule "head" in its definition [203]. Others levels of complexity are then also described [203].

We hypothesize there is *a set of subjective constructs* useful in health-related research that could be understood as *semantic molecules*. The complexity of their definition is low; therefore, their meaning is easy to convey (Figure 19). We used "sadness" as an example, but it might be plausible to define the concept of "pain intensity" (we are referring here to the phenomenological experience of pain) as a semantic molecule. The low level of complexity of this definition could be related, in part, to the fact that some subjective constructs can be measured using a single-item instrument (Figure 19). For instance, the measurement of pain intensity is frequently achieved using a simple Visual Analogue Scale (VAS).

Figure 19. Categorization of two sets of subjective constructs useful in health-related research according to the complexity of their linguistic definition



On the other hand, the definition of *other subjective constructs* is of a *higher level of complexity* (Figure 19). At least, the definition of semantic molecule(s) is first required to describe the concept. In that case, this high level of complexity of definition could be related, in part, to the need for multiple-item scales to measure a targeted construct. According to current views in the international scientific community, HRQL is highly complex to conceptualize [18,116,206,207]. Prior to the development of the response shift theory, there was already some research focusing on the dynamic nature of the HRQL construct [26]. It was then already acknowledged the meaning of that construct fluctuates across individuals and time [26]. Thus it seems HRQL is a concept that falls into the category of a subjective construct with a definition entailing a high level of complexity (Figure 19).

To briefly synthesize, we postulate a typology of subjective constructs used in health-related research based on the complexity of their definition:

- there is *a set of subjective constructs* that can be understood as *semantic molecules* (i.e. their *definition exhibits a low level of complexity*);
- *other subjective constructs exhibit a higher level of complexity* (e.g. HRQL).

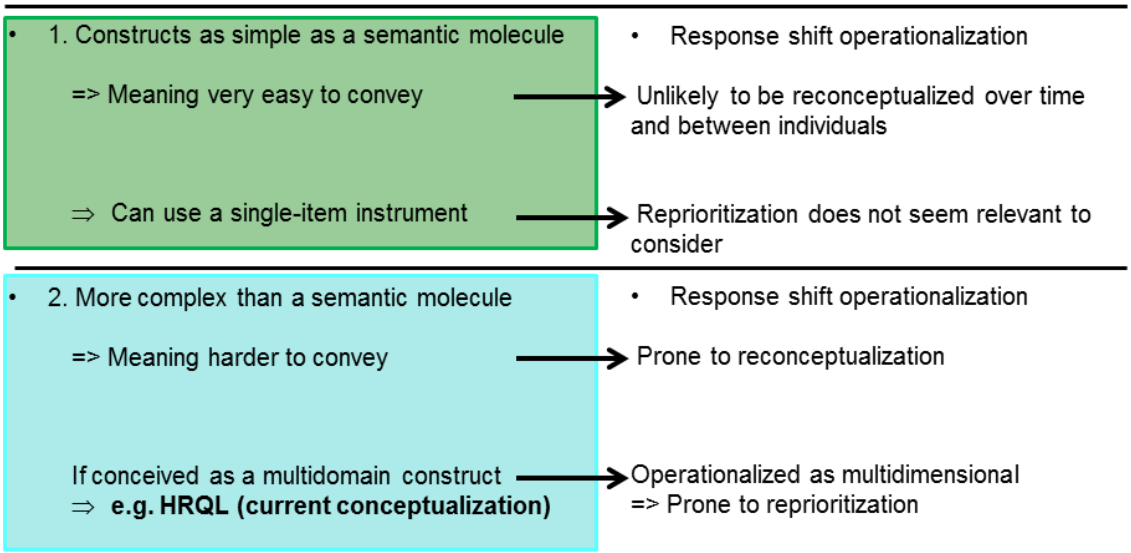
3.2. Relationship between the level of complexity of a construct and reconceptualization response shift

We previously hypothesized that there is a set of subjective constructs useful in health-related research that can be defined as *semantic molecules*. As such, the meaning of these constructs, because it is closely related to the meaning of the few semantic primes used to describe them, if not strictly universal and unambiguous, exhibits a very low level of complexity, and therefore is easy to communicate. If we consider that "pain intensity" referred to earlier is one of these constructs, the frequent use of a single-item scale to measure it could be related to this low level of complexity. An investigator only needs one (or few) item(s) and few instructions to capture most of the relevant information, and to be understood by the respondent.

On the other hand, the meaning of *constructs with a highly complex definition*, such as HRQL, is harder to convey. A multi-item instrument is needed to capture sufficient information and cover all the facets of these complex concepts.

In terms of the operationalization of reconceptualization response shift, we postulate the impact of reconceptualization on a measure is not the same according to the level of complexity of the definition of the constructs explored (Figure 20).

Figure 20. Relationships between complexity of linguistic definition, dimensionality and impact of response shift effects on changes in HRQL scores observed



Indeed, for *constructs that can be equated with semantic molecules*, we postulate that they are unlikely to be reconceptualized by given individuals over time, and that if they are presented with the appropriate wording and response scale [200], they will be understood in a very similar way between individuals (Figure 20).

Conversely, *constructs with a highly complex definition*, such as HRQL, are prone to reconceptualization response shift, which can affect HRQL scores (Figure 20).

4. Multidimensionality as the next sin of HRQL

4.1. The current conceptualization of HRQL

HRQL is also often viewed as highly complex in psychometric terms. Indeed, although there is no current consensus, whether on the number of domains involved, or on the content of those domains [18,65], most researchers consider HRQL to be a *multidomain* concept - often including physical, social, psychological and spiritual domains - [18,65]. In addition, a broad variety of outcomes, from symptoms and functioning assessment to more subjective appreciations, are frequently subsumed under the umbrella term HRQL, overlapping with the term PRO [59,207].

It is tempting here to make a jump from concept to psychometrics by associating a *multidomain concept* with a *multidimensional construct*. In fact, when researchers are assessing the factorial validity of a PRO instrument using data from a validation sample, it is expected an instrument designed to measure a multidomain concept will prove to be multidimensional after factor analysis [18,90]. Thus, as HRQL is often conceptualized as a multidomain concept, it is often operationalized to be measured as a *multidimensional construct* (in psychometric terms). One example is the factor structure of the SF-36 Health Survey [74,208].

4.2. Relationship between the dimensionality of a construct and reprioritization response shift at domain-level

In line with what we previously postulated on the relationship between reconceptualization and the complexity of a construct (in linguistics terms), we now postulate the impact of reprioritization on measures will vary according to the dimensionality of a construct (in psychometric terms).

As highlighted by SPRANGERS and SCWHARTZ in their definition of response shift [30], reprioritization implies a change in the relative importance in the respondent's mind between at least two domains constituting the targeted construct. We have also previously seen that a correspondence between domain (in terms of conceptualization) and dimensionality (in terms of operationalization) is often determined by assessing the factor validity of a construct.

Therefore, a subjective construct operationalized as markedly unidimensional may not be a candidate for reprioritization: reprioritization may only occur if a construct is at least bidimensional (Figure 20).

5.Theoretical proposals to achieve a meaningful score measuring patients' subjective experience for clinical trials

To summarize, while there is clearly a need to measure patients' subjective experience in health-related research, in the context of a RCT, a measure of this type needs to be simple and clear to interpret. It should not be biased by response shift, as it is usually not the effect of interest. However, HRQL is currently conceptualized in highly complex manner, both in terms of semantic definition (definitions with a high level of complexity according to the semantic primes theory) and in terms of factorial structure (often operationalized as multidimensional), making HRQL scale scores prone to reconceptualization and reprioritization effects.

Thus, we would like to propose that this issue could be efficiently dealt with in the design phase of a PRO instrument, with a rule of "the least linguistic and psychometric complexity" in mind.

Indeed, if a construct leading to a PRO instrument assessing patients' subjective experience of the course of treatment was soundly define (i.e. with the least semantic complexity), then this construct could be broken down into items that can be equated as closely as possible with semantic molecules; and if it was operationalized in unidimensional form, it would be possible to obtain:

- a score that could not be readily biased by response shift (reconceptualization and reprioritization);
- a score that could be used as an easily interpretable endpoint in RCTs (because it provides a single score, not a profile [59]).

6. Discussion

In this paper, we discussed links between complexity of a subjective construct and reconceptualization and reprioritization, but not recalibration response shift. Indeed, recalibration is related to the metric standard associated with response categories. In TOURANGEAU et al. psychological model of survey response, setting the metric standard and mapping a judgment onto response categories are part of the judgment and response processes [200]. Therefore, it seems recalibration can be thought as mainly independent of the issue of semantic and psychometric complexity. However, we would like to propose that the problem of recalibration could be dealt with using a suitable response format. Indeed, recalibration can occur because the metric standard is implicit and internal when using a VAS or Likert-scale. We therefore hypothesize that by using brief case vignettes (as in some instruments designed to assess global impression, like the improved Clinical Global Impression Scale (iCGI) [20]), illustrating what each response choice is referring to, the standard metric is rendered explicit and external, forcing the respondent to calibrate himself/herself on the metric the researcher expects.

In this paper, we hypothesize a link between reprioritization and dimensionality, in agreement with the theoretical definition of reprioritization, which seems to define reprioritization at domain level [30]. We are nevertheless aware that in SEM-based methods, reprioritization is operationalized as a change in values of factor loadings between two measurement times.[37]. Therefore, in SEM operationalization, a change in the values of factor loadings will be viewed as “reprioritization” (it will be “reprioritization at item-level”), even if the construct is unidimensional. It seems there may be a need here for clarification of the interpretation of changes in parameters of SEM models when performed at item level on a unidimensional construct; or, at least, clarification of the definition of reprioritization.

Response shift theory can be viewed as related to two different conceptual issues in psychometrics [114]. It is a theory designed to highlight the influence of psychological mechanisms called upon when an individual has to deal with life changes [30]. But as it highlights a potential change in the way people appraise their HRQL [118], it can also be viewed as a source of bias when it is not the effect investigated by the HRQL measurement [53,145]. Our proposals are related to the latter conceptualization. However, we claim these two views can be thought to be complementary and context-dependent views: using one or the other depends on the purpose of a study. In the context of a RCT, the direct effect of a treatment on HRQL is usually investigated, so that response shift can be viewed as a source of bias [123].

On the other hand, if the purpose is to investigate the different variables affecting a person's HRQL, then it is more appropriate to view response shift as one of the outcomes of interest. Therefore, the use of a method to explicitly quantify response shift is better suited.

In relation to these last comments, we acknowledge, as has been suggested by some earlier studies [176,177], that psychological interventions can be used to facilitate adaptation, therefore resulting in response shift occurence, as a way to enhance his/her HRQL, especially in the context of palliative care, when curing the disease is no longer an option. If the effectiveness of an intervention of this type were then to be investigated in the context of a RCT, since response shift would be the main effect of interest, a method designed to discriminate the response shift effect would be better suited than an instrument designed according to our proposals.

When designing a PRO instrument, establishing validity of the questionnaire is of paramount importance. In the last century, a great deal of efforts has been devoted to successfully develop techniques to investigate structural validity [18]. Although developments have been made via, for instance, the field of "Cognitive Aspects of Survey Methodology", to reduce the cognitive burden associated with the syntax of items [200,209–212], methodological advances addressing issues regarding face or content validity are still needed to be built. Potential tools derived from semantic primes theory could be developed to help assessing the semantic complexity of construct and items of actual or new HRQL questionnaires. It could also allow for experimental investigation of the theory, by developing items with different semantic complexity and investigate whether it influences the occurrence of response shift effect. These potential tools could be complementary options to help researchers to improve face and content validity of a questionnaire.

7. Conclusion

The development of the response shift theory has been helpful in highlighting the impact of psychological mechanisms on the assessment of individual HRQL over time. It has also shown the response shift effect could be a source of bias in the context of a RCT, when interpreting an observed change over time in HRQL. Hence our proposal as theoretical guidelines for designing an instrument providing a score that is not biased by response shift, with a straightforward interpretation.

We have also highlighted links between response shift and HRQL conceptualization and operationalization. Therefore, debate on what HRQL is seems relevant in the wake of the response shift theory, in order to clarify what content an investigator wants to convey to a respondent when measuring HRQL using a questionnaire.

We have seen that the "simpler" is the definition of constructs and items, the "easier" is the interpretation of the score measuring patients' subjective experience of the course of treatment. Semantic primes theory has been helpful to define what is "simple", in semantic terms. A next step would be to convert it into a tool assessing the level of complexity of the definition of constructs and items when designing PRO instruments.

Acknowledgment

The authors would like to thank Angela VERDIER for her help for the English wording and phrasing of the paper.

Part 6: General discussion and conclusion

"In the end, we are all alone and no one is coming to save you" – John Reese, a fictional skillful person

This last part will be dedicated to discuss some of the main results or proposals of the two original works of this thesis (the fourth and fifth part). Regarding the simulation study, the main limit (the fact an *a priori* inadequate estimator has been used) will be discussed with more depth, along with the introduction of additional analyses performed after the publication of the seminal paper which will help to better apprehend the potential robustness of the original results. Then, some proposals about potential refinements of the OP algorithm will be discussed. Then, some questions about the links between operationalization and interpretation of response shift at item-level using SEM will be proposed. A second chapter will be dedicated to discuss the fifth part of the manuscript (i.e. semantic primes and response shift paper). First, a discussion about the issue of recalibration response shift in this context will be introduced. Finally, this theoretical proposal will be used as an example to illustrate a larger issue which is to find an adequate balance between studying complex phenomena and planning experimental designs.

1. A discussion about the detection of response shift using Oort's procedure

1.1 On the limits and robustness of the results of the simulation study of this thesis work

In the fourth part of this thesis work (it will be further referred as the "simulation study" for simplicity), we have chosen to simulate responses to five binary items using a longitudinal Rasch model. At each times of measurement, the five items measured the same construct (i.e. assumption of unidimensionality). As it was, to our knowledge, the first simulation study assessing OP performances (i.e. it was therefore a pilot simulation study) and as simulation study can be easily quite computationally intensive to perform, the rationale behind the choose of these simulation conditions was to obtain a simple data structure in order to optimize feasibility. In addition, it seemed relevant to dissociate the choice of the simulation model from the choice of the model used to analyze data, hence the use of a longitudinal Rasch model instead of a CFA model, hence the simulation of binary indicators.

Nonetheless, one of the consequences, which was noted in the discussion of the seminal simulation paper, was the datasets were analyzed using CFA with *a priori* an inappropriate estimator (robust Maximum-Likelihood with Satorra-Bentler correction (MLM)) [99]. Indeed,

when the study was performed, operationalization of response shift for categorical items was not proposed yet [213]. If the use of MLM can be an appropriate way to deal with a departure from the assumption of multivariate normality that is required when using ML, it only corrects χ^2 value, fit indices values and standard errors values of the parameters, but not the values of the parameters themselves [99]. Therefore, it can be a simple and effective way of correcting from moderate departure from multivariate normality (e.g. with categorical variables as indicators with a sufficient number of categories, like four to seven) [99]. But, with high departure from normality, there is a risk of estimating biased parameters [99]. In the aforementioned pilot study, as the simulated data were binary items, this risk of estimating biased parameters was a plausible one. Thus, the robustness of the main results of the study can be questioned.

A first indicator of the robustness of the results can be the fact the Type-I-error that was estimated for the omnibus test for overall response shift detection (model 2 versus model 1) when using the LRT was in the 5% range as theoretically expected when using MLM. Thus, it seems MLM indeed adequately corrected from estimating inappropriate χ^2 value and standard errors of the parameters. Therefore, results on Type-I-error and power of the global test for response shift detection could be in some sort considered as robust. Nonetheless, as there was still the risk of biased parameters estimates, relying on the aforementioned argument to conclude on the robustness of the results can be considered as insufficient.

One of the main result of the study was the fact the use of LRT leads to better power for the omnibus test for response shift detection than the use of IC (AIC, BIC or SABIC). But again, as there was still a risk of biased parameters estimates, the robustness of this result can be discussed. To investigate, the analysis cannot be done on the data of the study using the appropriate estimator (DWLS estimator) as it does not rely on the estimation of the value of a likelihood function [196]. Therefore, when using DWLS estimator, IC values are not computed. Nonetheless, another simulation study was performed with the simulation of continuous indicators using a CFA model (with the same combinations of 36 characteristics as in the original study) to produce a situation where ML estimator is the appropriate option to get values of the parameters. If power values were lower than those obtained with MLM (but they must not be compared, as they were obtained using different datasets with different assumptions), these new results confirmed power of the omnibus test is better when using LRT (χ^2 difference test) than IC (Table 8). Thus, this result seems to be a robust one.

Table 8. Estimated power for different strategies for global assessment for RS occurrence (continuous indicators, ML estimator)

n	α	r	ur	LRT (p < 0.05)		AIC2 > AIC1		SABIC2 > SABIC1		BIC2 > BIC1	
				%	CI 95%	%	CI 95%	%	CI 95%	%	CI 95%
100	**0**	**0.4**	**1**	16.1	[14.0 - 18.6]	7.9	[6.4 - 9.8]	35.1	[32.2 - 38.2]	0.0	[0.0 - 0.4]
			2	26.7	[24.1 -29.6]	15.0	[12.9 - 17.4]	45.0	[41.9 - 48.1]	0.0	[0.0 - 0.4]
		0.9	**1**	15.2	[13.1 -17.6]	7.1	[5.7 - 8.9]	31.6	[28.8 - 34.6]	0.0	[0.0 - 0.4]
			2	26.6	[24.0 -29.4]	14.8	[12.7 - 17.2]	46.0	[42.9 - 49.1]	0.0	[0.0 - 0.4]
	0.2	**0.4**	**1**	16.0	[13.9 -18.4]	8.1	[6.6 - 10.0]	32.8	[29.9 - 35.8]	0.0	[0.0 - 0.4]
			2	24.3	[21.7 -27.0]	13.3	[11.3 - 15.6]	41.9	[38.8 - 45.0]	0.0	[0.0 - 0.4]
		0.9	**1**	14.6	[12.5 -16.9]	7.1	[5.7 - 8.9]	30.5	[27.7 - 33.4]	0.0	[0.0 - 0.4]
			2	25.8	[23.2 -28.6]	13.4	[11.4 - 15.7]	44.0	[40.9 - 47.1]	0.0	[0.0 - 0.4]
200	**0**	**0.4**	**1**	30.7	[27.9 -33.7]	16.8	[14.6 - 19.3]	12.8	[10.9 - 15.0]	0.0	[0.0 - 0.4]
			2	53.5	[50.4 -56.6]	35.1	[32.2 - 38.1]	29.5	[26.8 - 32.4]	0.0	[0.0 - 0.4]
		0.9	**1**	30.2	[27.5 -33.1]	16.9	[14.7 - 19.4]	12.7	[10.8 - 14.9]	0.0	[0.0 - 0.4]
			2	52.1	[49.0 -55.2]	36.0	[33.1 - 39.0]	29.3	[26.6 - 32.2]	0.1	[0.0 - 0.6]
	0.2	**0.4**	**1**	28.6	[25.9 -31.5]	15.0	[12.9 - 17.3]	11.6	[9.8 - 13.7]	0.0	[0.0 - 0.4]
			2	51.0	[47.9 -54.1]	35.4	[32.5 - 38.4]	28.8	[26.1 - 31.7]	0.0	[0.0 - 0.4]
		0.9	**1**	30.9	[28.1 -33.9]	14.9	[12.8 - 17.3]	11.4	[9.6 - 13.5]	0.0	[0.0 - 0.4]
			2	49.6	[46.5 -52.7]	34.7	[31.8 - 37.7]	28.1	[25.4 - 31.0]	0.0	[0.0 - 0.4]
300	**0**	**0.4**	**1**	46.6	[43.5 -49.7]	29.6	[26.9 - 32.5]	11.3	[9.5 - 13.4]	0.0	[0.0 - 0.4]
			2	75.5	[72.7 -78.1]	61.0	[57.9 - 64.0]	32.7	[29.9 - 35.7]	0.0	[0.0 - 0.4]
		0.9	**1**	45.3	[42.2 -48.4]	28.9	[26.2 - 31.8]	9.5	[7.8 - 11.5]	0.0	[0.0 - 0.4]
			2	72.5	[69.7 -75.2]	58.7	[55.6 - 61.7]	31.6	[28.8 - 34.5]	0.0	[0.0 - 0.4]
	0.2	**0.4**	**1**	45.5	[42.4 -48.6]	30.4	[27.6 - 33.3]	9.8	[8.1 - 11.8]	0.0	[0.0 - 0.4]
			2	71.7	[68.8 -74.4]	56.5	[53.4 - 59.5]	30.1	[27.3 - 33.0]	0.0	[0.0 - 0.4]
		0.9	**1**	44.9	[41.8 -48.0]	28.1	[25.4 - 31.0]	9.0	[7.4 - 10.9]	0.0	[0.0 - 0.4]
			2	76.3	[73.6 -78.8]	60.5	[57.4 - 63.5]	29.6	[26.9 - 32.5]	0.0	[0.0 - 0.4]

The original datasets of the seminal simulation study were reanalyzed in recent times with the use of an appropriate estimator as proposed by VERDAM et al. (i.e. a DWLS estimator) [213]. However, as the simulated data were binary items, there was only one threshold to estimate by item. Therefore, we couldn't choose the alternative parameterization (it imposes identifiability constraints on at least two thresholds to be usable) to scale the continuous latent indicators, but the delta parameterization instead. With this parameterization, the mean and variance of the latent continuous indicator are fixed to 0 and 1 (see Part 2 chapter 2.2.b.b). Thus, as the metric of the latent continuous indicator is fixed as identifiability constraint, it is a change in the value of the threshold associated to an item over time that can be operationalized as indicative of recalibration response shift. However, it has to be stressed out it is impossible with binary items to dissociate uniform from non-uniform recalibration response shift. Therefore, even if uniform recalibration response shift was the only form simulated in the datasets, the only conclusion after analyses, if a change in threshold value over time was detected, is the detection of recalibration response shift without further distinction.

Regarding the omnibus test of response shift detection, one disturbing result was the fact when running analyses with an a priori adequate estimator (the DWLS estimator) the estimated Type-I-error was below the expected 5% in most of the cases (Table 9). When performing this analysis, we used a mean and variance adjusted test statistic (scaled and shifted) for DWLS estimator as recommended in the recent SEM literature [214]. Nonetheless, it seems to have maybe under-corrected for the risk of an incorrect risk of Type-I-error and have led to lower value than theoretical correct ones.

Table 9. Estimated Type-I-Error for global assessment for RS occurrence with DWLS or MLM estimator

n	α	ρ	rs	DWLS		MLM	
				%	$CI_{95\%}$	%	$CI_{95\%}$
100	**0**	**0.4**	**0**	1.7	[0.9 - 3.2]	4.5	[3.1 - 6.7]
		0.9	**0**	4.5	[3.2 - 6.2]	3.7	[2.4 - 5.5]
	-0.2	**0.4**	**0**	1.9	[1.1 - 3.4]	5.9	[4.1 - 8.2]
		0.9	**0**	3.7	[2.5 - 5.3]	5.9	[4.2 - 8.1]
200	**0**	**0.4**	**0**	2.5	[1.6 - 3.8]	4.0	[2.8 - 5.8]
		0.9	**0**	2.9	[1.9 - 4.2]	3.6	[2.5 - 5.2]
	-0.2	**0.4**	**0**	0.9	[0.4 - 1.8]	4.0	[2.8 - 5.7]
		0.9	**0**	3.7	[2.6 - 5.1]	5.4	[4.1 - 7.3]
300	**0**	**0.4**	**0**	1.1	[0.6 - 2.1]	3.0	[2.0 - 4.4]
		0.9	**0**	2.4	[1.6 - 3.6]	4.5	[3.3 - 6.1]
	-0.2	**0.4**	**0**	1.5	[0.9 - 2.5]	5.8	[4.4 - 7.7]
		0.9	**0**	1.5	[0.9 - 2.4]	4.4	[3.3 - 6.0]

In terms of power, in most of the cases, the estimated power with DWLS was lower than those estimated with MLM (Table 10). Indeed, if power was estimated to be above 90% for a sample size (*n*) of at least 200 and uniform-recalibration (*ur*) simulated on two items when using MLM, power was below 80% for these conditions using DWLS and was around or above 90% with $n = 300$ and $ur = 2$ only (Table 10). Theoretically, DWLS estimator along with the mean and variance adjusted test for χ^2 difference test is the appropriate way to deal with binary indicators [99,214]. Thus, these additional results regarding estimation of power of the omnibus test for response shift detection should be considered as the appropriate ones. However, as results regarding Type-I-error using a DWLS estimator did not match theoretical expected values (contrariwise to MLM estimator), the robustness of these new results can also be discussed. Especially it can be discussed if these estimates of power are underestimated values of the true power of the method. Nonetheless, irrespective of the estimator used, it seems power for a global detection of uniform recalibration is above 80% when $ur = 2$ and the sample size is around 200 to 300 when the structure of the data is five items measuring a single construct.

Table 10. Estimated power for global assessment for RS occurrence (DWLS or MLM estimator)

n	α	ρ	rs	DWLS		MLM	
				%	CI 95%	%	CI 95%
100	**0**	**0.4**	**1**	15.9	[13.1 - 19.3]	36.4	[32.3 - 40.7]
			2	25.5	[21.9 - 29.5]	57.7	[53.1 - 62.2]
		0.9	**1**	19.6	[16.9 - 22.7]	37.8	[34.0 - 41.8]
			2	30.7	[27.5 - 34.2]	61.2	[57.1 - 65.1]
	-0.2	**0.4**	**1**	14.6	[12.0 - 17.7]	36.6	[32.7 - 40.8]
			2	25.3	[21.9 - 29.0]	60.0	[55.6 - 64.2]
		0.9	**1**	20.1	[17.3 - 23.2]	41.0	[37.1 - 45.1]
			2	29.5	[26.3 - 33.0]	61.1	[57.1 - 65.1]
200	**0**	**0.4**	**1**	42.2	[38.8 - 45.7]	72.8	[69.4 - 76.0]
			2	64.2	[60.7 - 67.5]	94.6	[92.6 - 96.1]
		0.9	**1**	42.0	[38.8 - 45.3]	75.1	[71.8 - 78.0]
			2	68.5	[65.3 - 71.5]	94.9	[93.1 - 96.3]
	-0.2	**0.4**	**1**	43.2	[39.7 - 46.7]	75.0	[71.7 - 78.1]
			2	65.1	[61.6 - 68.4]	92.2	[89.9 - 94.0]
		0.9	**1**	44.6	[41.4 - 47.9]	76.2	[73.1 - 79.2]
			2	74.1	[71.2 - 76.9]	94.3	[92.5 - 95.8]
300	**0**	**0.4**	**1**	70.2	[67.1 - 73.1]	92.3	[90.3 - 94.0]
			2	89.3	[87.0 - 91.2]	99.6	[98.9 - 99.9]
		0.9	**1**	72.2	[69.3 - 75.0]	94.1	[92.3 - 95.5]
			2	90.2	[88.1 - 91.9]	99.6	[98.9 - 99.9]
	-0.2	**0.4**	**1**	69.9	[66.8 - 72.9]	92.2	[90.1 - 93.8]
			2	92.6	[90.7 - 94.2]	99.5	[98.7 - 99.8]
		0.9	**1**	69.5	[66.5 - 72.3]	92.3	[90.3 - 93.9]
			2	93.8	[92.1 - 95.1]	99.8	[99.2 - 99.9]

Regarding OBI1 values (i.e. the assessment of the proportion of datasets for which the whole OP has properly detected uniform recalibration response shift on only truly affected item(s) (after a significant LRT ascertaining global response shift occurrence), disregarding any false detections of response shift on these or one of the other items), the estimates were globally lower when using DWLS instead of MLM. This decrease in OBI1 values when using DWLS comparing to MLM can be attributed mostly to the decrease in power of the omnibus test (Table 10, Table 11). Nonetheless, this difference vanished as sample size and the quantity of uniform recalibration simulated increase. Indeed, for *ur* = 2 and *n* = 300, OBI1 values were found to be close between the two estimators (Table 11). *In fine*, even if a decrease in OBI1 values has been globally observed when using DWLS estimator, uniform recalibration simulated is adequately detected in most of the cases when using DWLS estimator conditionally on the fact the omnibus test is positive, which confirms the results of the seminal simulation study.

Regarding OBI2 values (this indicator is nearly identical as OBI1, but with an additional requirement of no false detections of response shift on any item(s)), estimated values were globally higher when using DWLS instead of MLM (Table 11). This result can be attributed to the relationships between the simulation process and the differences in operationalization of recalibration detection between DWLS and MLM. As aforementioned, theoretically, non-uniform and uniform recalibration cannot be distinguished from one another when modeling binary items. In this study, we simulated uniform recalibration only. With DWLS estimator, there is only one parameter to operationalize the detection of uniform or non-uniform recalibration which is a change in the value of the threshold associated to an item (using the delta parameterization). Thus, whenever a significant change in threshold value was detected on correct item(s) only after using OP, it was interpreted as a correct detection of the uniform recalibration simulated and the condition to conclude on the positivity of OBI2 criterion was considered to be met. Contrariwise, when using MLM estimator, there are two parameters which represents the operationalization of recalibration response shift: error variances for non-uniform recalibration and intercepts for uniform recalibration. But, as aforementioned in the pilot study paper, simulating uniform recalibration on binary items could have led to a situation where this recalibration was detected by a change in error variances instead of intercepts and therefore interpreted as non-uniform recalibration (see Part 5, chapter 6.4). When this was the case, the condition to consider OBI2 criterion as positive was not met and therefore OBI2 criterion was not assumed to be fulfilled. *In fine*, this is probably why OBI2 values did not rise

above 18.3% when using MLM estimator but rise to 52.2% when using DWLS estimator (Table 11).

Table 11. Estimated OBI1 and OBI2 values, DWLS and MLM estimator

n	α	ρ	rs	DWLS				MLM			
				OBI1		OBI2		OBI1		OBI2	
				%	CI 95%	%	CI 95%	%	CI 95%	%	CI 95%
100	**0**	**0.4**	**1**	11.5	[9.1 - 14.4]	2.6	[1.6 - 4.3]	26.2	[22.6 - 30.2]	4.5	[3.0 - 6.7]
		0.9	**1**	14.3	[12.0 - 17.1]	3.8	[2.6 - 5.4]	27.1	[23.6 - 30.8]	6.6	[4.8 - 8.8]
	-0.2	**0.4**	**1**	9.5	[7.4 - 12.1]	2.7	[1.7 - 4.4]	27.5	[23.9 - 31.4]	6.6	[4.8 - 9.0]
		0.9	**1**	12.4	[10.2 - 15.1]	2.4	[1.5 - 3.8]	28.3	[24.8 - 32.2]	6.6	[4.9 - 8.9]
200	**0**	**0.4**	**1**	32.2	[29.1 - 35.6]	16.8	[14.4 - 19.6]	58.4	[54.7 - 62.0]	11.2	[9.1 - 13.7]
		0.9	**1**	32.5	[29.4 - 35.6]	16.9	[14.5 - 19.5]	59.7	[56.2 - 63.2]	13.9	[11.6 - 16.5]
	-0.2	**0.4**	**1**	33.8	[30.6 - 37.2]	19.0	[16.4 - 21.9]	60.5	[56.8 - 64.1]	16.7	[14.1 - 19.6]
		0.9	**1**	33.1	[30.1 - 36.3]	17.7	[15.3 - 20.4]	60.9	[57.4 - 64.4]	18.5	[15.9 - 21.5]
300	**0**	**0.4**	**1**	56.4	[53.1 - 59.7]	35.9	[32.8 - 39.2]	75.5	[72.4 - 78.4]	11.0	[9.0 - 13.3]
		0.9	**1**	56.4	[53.3 - 59.5]	37.4	[34.4 - 40.5]	74.5	[71.5 - 77.3]	13.6	[11.5 - 16.1]
	-0.2	**0.4**	**1**	55.8	[52.5 - 59.1]	37.6	[34.4 - 40.9]	75.2	[72.1 - 78.1]	16.1	[13.7 - 18.8]
		0.9	**1**	54.9	[51.7 - 58.0]	36.1	[33.1 - 39.2]	75.5	[72.5 - 78.2]	18.3	[15.9 - 21.0]
100	**0**	**0.4**	**2**	11.8	[9.3 - 14.9]	5.5	[3.8 - 7.8]	21.9	[18.3 - 25.9]	2.4	[1.4 - 4.3]
		0.9	**2**	14.8	[12.4 - 17.5]	7.4	[5.7 - 9.5]	27.2	[23.7 - 31.0]	3.7	[2.5 - 5.6]
	-0.2	**0.4**	**2**	13.0	[10.5 - 16.0]	6.7	[5.0 - 9.1]	28.2	[24.4 - 32.3]	4.0	[2.6 - 6.1]
		0.9	**2**	17.7	[15.1 - 20.7]	7.2	[5.5 - 9.4]	28.8	[25.2 - 32.7]	3.4	[2.2 - 5.2]
200	**0**	**0.4**	**2**	41.6	[38.1 - 45.1]	27.3	[24.2 - 30.6]	52.6	[48.8 - 56.4]	1.3	[0.7 - 2.5]
		0.9	**2**	47.0	[43.7 - 50.4]	30.9	[27.9 - 34.1]	57.7	[54.2 - 61.2]	2.1	[1.3 - 3.4]
	-0.2	**0.4**	**2**	44.9	[41.4 - 48.5]	26.4	[23.3 - 29.6]	57.4	[53.6 - 61.0]	6.5	[4.9 - 8.6]
		0.9	**2**	54.8	[51.5 - 58.0]	35.3	[32.2 - 38.5]	64.9	[61.5 - 68.2]	4.9	[3.5 - 6.6]
300	**0**	**0.4**	**2**	64.7	[61.5 - 67.8]	47.7	[44.4 - 51.1]	62.5	[59.0 - 65.8]	0.6	[0.3 - 1.5]
		0.9	**2**	67.9	[64.8 - 70.8]	48.1	[44.9 - 51.2]	69.8	[66.6 - 72.8]	0.8	[0.4 - 1.7]
	-0.2	**0.4**	**2**	69.3	[66.2 - 72.3]	50.1	[46.7 - 53.4]	69.2	[65.9 - 72.3]	2.7	[1.8 - 4.1]
		0.9	**2**	71.8	[68.8 - 74.5]	52.2	[49.1 - 55.4]	72.5	[69.5 - 75.4]	2.5	[1.7 - 3.8]

To conclude, even if the presence or not of a bias of the estimates (i.e. Type-I-error, power, OBI1, OBI2) of this simulation study can still be discussed, some essential results of the pilot simulation study were confirmed by the subsequent analyses performed after the publication of the seminal paper. Indeed, it seems the LRT (or χ^2 difference test) is the better test to use as an omnibus test for global response shift occurrence. Conditionally on the positivity of the omnibus test, it also seems simulated recalibration was adequately detected in most of the cases when using OP whether MLM or DWLS estimator is used. Finally, when using DWLS estimator, with *ur* = 2 and *n* = 300, it seems OP works perfectly (OBI2 criterion fulfilled) in around half of the cases (with a structure of five binary items loading on one construct).

The last aforementioned result is interesting. Indeed, it seems to indicate the use of OP on binary items, even with an a priori correct estimator, can lead frequently to false positive detection of other forms of response shift. Assuming these subsequent analyses of the simulation study are somehow sound, this result raises issues about the most appropriate way to perform OP as an algorithm.

1.2. On a possible refinement of Oort's procedure algorithm

In the seminal OORT paper which introduced OP, it was presented as a backward and forward approach [37]. The backward component, performed in one step (step 2 of the algorithm, see Part 4, chapter 3), is the omnibus test of overall response shift detection (model 2 versus model 1). The forward component is the iterative procedure of response shift detection one parameter at a time (step 3, see Part 4, chapter 3). Moreover, OP was introduced by OORT without a hierarchy in testing the different types of response shift (i.e. reconceptualization, reprioritization, non-uniform or uniform recalibration) and OORT did not especially advocate for correction for multiple statistical hypothesis testing in step 3 (i.e. the forward component). Thus, OP, as initially proposed by OORT, was an algorithm with: an omnibus test for overall response shift detection, no hierarchy in testing different types of response shift, and no adjustment for multiple statistical hypothesis testing in step 3 [37].

The algorithm used in the simulation study of this work was a slightly modified version of the original OP as it introduced a hierarchy in testing the different types of response shift. Non-uniform recalibration was tested first, followed by recalibration and reprioritization response shift. This hierarchy was introduced in part as a matter of programming feasibility (the

simulation study required to automatize OP as an algorithm and the programming was considered easier at the time to do so with the introduction of a hierarchy), and in part as it was sometimes proposed by some authors [41,154], as it can be performed in that way in measurement invariance studies [195]. The results of the simulation study showed, even with an appropriate estimator (i.e. DWLS), a perfect detection of response shift simulated in around half of the cases top (OBI2 estimates) (see Part 6, chapter 1.1).

Thus, it seems the flow of the OP algorithm, especially regarding certain key characteristics can be questioned in order to optimize its performances. Three characteristics may be questionable: the need or not for the omnibus test for overall response shift detection (step 2), the need or not for a hierarchy in testing the different types of response shift at step 3, and the need or not to correct for multiple statistical hypothesis testing at step 3.

The need or not for a hierarchy in testing the different types of response shift has been already mentioned. Indeed, some authors argue for in order to reproduce what can be performed in measurement invariance studies [41,154]. The need or not to correct for multiple statistical hypothesis testing has also been mentioned in some empirical studies using the OP as the procedure to detect response shift [42,44,159]. Indeed, the iterative LRTs performed in step 3 in order to detect the occurrence of response shift can be thought as a same family of dependent tests. Thus, for n tests with an a priori α level of Type-I-error, there is a probability of $1 - (1 - \alpha)^n$ to have at least one false positive test for response shift detection at step 3.

The need or not to perform the omnibus test for overall response shift detection has, to our knowledge, never been discussed in response shift literature. OP can be thought as an algorithm designed to search for the most parsimonious longitudinal CFA model to explain changes over time of a measurement of a construct [37]. If there is no response shift in the data, then the most parsimonious model is the model with all parameters constrained to be equal at each times of measurement (which is the second model of step 2 used for the omnibus test). If there is response shift everywhere in the data, then the most parsimonious model is the model with no equality constraints at all, which is the model fitted at step 1. This last assumption can be seen as empirically unrealistic. Thus, it can be questioned if the OP could start by fitting the model with all parameters constrained to be equal at each times of measurement and searching for response shift from this model (i.e. restricting the OP to the forward approach only). The rationale behind the fit of the model at step 1 of the current OP algorithm (the model with no constraints at all) is to verify if a measurement model with appropriate structural validity can be fit on the data. However, as aforementioned, it is probably unrealistic to suppose this model

would be the most parsimonious as it assumes all forms of response shift on all items, thus the fit of this model as a test for the possibility of an adequate measurement model can be questioned. Another argument for fitting model 1 can be the fact the omnibus test for response shift detection can be a way of preventing for a too high probability of Type-I-error, especially if the hierarchy in testing different types of response shift is used. Nonetheless, this need to perform the omnibus test could be prevented by an adequate correction for multiple statistical hypothesis testing when performing the iterative search for occurrence of response shift (i.e. step 3).

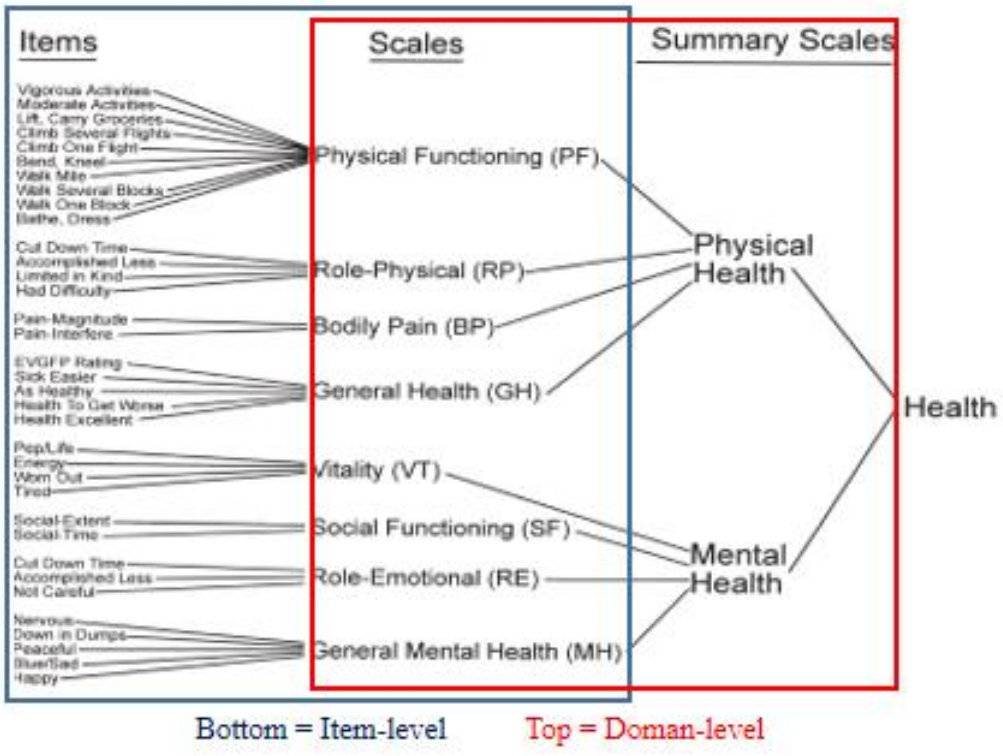
Whatever the theoretical soundness of these different options, the different combinations of these choices regarding the best way to perform OP can be investigated by simulation studies in order to define which strategy leads to the best performances (i.e. highest OBI2 values). In that regard, these different strategies are currently investigated by simulation studies within the research team by Guilleux et al.

1.3. On the operationalization and interpretation of response shift at item-level

Response shift detection at what is called "*item-level*" has become of prime interest in recent years. Indeed, a special section of *Quality of Life Research* was devoted to this issue in 2016 [215]. Initially, especially when using OP, response shift was usually investigated at what is called "*domain-level*" [39].

On a conceptual basis, the distinction between domain-level and item-level response shift is, in part, linked to the structure of the SF-36 which is composed of two levels of measurement [74] (Figure 21). The *top level* of the structure of the SF-36 is the *domain-level*. It describes how two broad latent variables (the Physical Component of HRQL and the Mental Component of HRQL) can be subdivided in four domains each (for a total of eight domains). These eight domains are scales or scores that are measured each by a certain number of items. Thus, the *bottom level* of the structure of the SF-36 describes how each domain (or dimension in the psychometric sense) are measured by items.

Figure 21. The two-levels structure of the SF-36 *(Adapted from: KELLER et al., Journal of Clinical Epidemiology, 1998 [74])*



So, when investigating the occurrence of response shift at domain-level using the SF-36, what is explored and therefore modeled is the relationships between the eight subdomains and the two aforementioned broad latent variables representing HRQL. Thus, what is used as indicators of the two broad latent variables are eight quantitative scores representing different sides of what is supposed to be HRQL [39]. At this level, the operationalization and interpretation of response shift occurrence has been heavily discussed and especially formalized by OORT et al in 2009 [145,159]. But, when investigating the occurrence of response shift at item-level using the SF-36, what is explored and therefore modeled is the relationships between each of the eight subdomains and the items designed to measure them. Thus, what is used as indicators of each subdomain are items.

One issue that had to be tackled to investigate response shift at item-level is the fact items are frequently assessed using Likert-scale and therefore corresponds to categorical variables (from two to six categories in the SF-36) [60]. Therefore, modeling the relationships between categorical items with these number of categories and a domain using CFA cannot be done using ML estimator as it assumes multivariate normality [99]. Thus, one major issue of investigating adequately response shift at item-level was a statistical one (i.e. to use an adequate

estimator in order to obtain unbiased SEM parameters, which implies to redefine adequately the operationalization of response shift at this level). Theoretically, it has to be noted if items indicative of a domain were measured in a quantitative manner (assuming a Gaussian distribution), this would not be a statistical issue as ML estimator would then be available. Therefore, we stressed out the issue of detecting response shift at item-level using SEM cannot always be a statistical issue: it is directly linked to the response format used for measuring items.

One of the appropriate way to accommodate the use of SEM with categorical indicators (or items) has been already described in the theoretical prerequisites of this thesis work: it is the three-stages approach (see Part 2, chapter 2.2.b.b) [101]. In short, first, categorical items are assumed to be representative of latent continuous indicators following a Gaussian distribution that has been truncated and thresholds are estimated. Second, polychoric correlations between the indicators are estimated. Third, these polychoric correlations are used as data to fit a CFA model using a DWLS estimator. This estimator requires more identifiability constraints (i.e. the metric of latent continuous indicators has to be scaled) and usually more parameters to estimate (i.e. for an item with *k* categories, there are *k*-1 thresholds to estimate) than ML estimator.

An operationalization of the detection of response shift at item-level using this aforementioned method of estimation along with an application on empirical data (with HRQL measured using the SF-36) has been proposed by VERDAM et al. in 2016 [213]. Comprehensive details about the method developed can be found in the corresponding paper. In short, the OP for response shift detection was adapted to accommodate the use of categorical items. At this level, it is now decomposed in two-main stages. The first one is dedicated to get a measurement model of the relationships between items and a SF-36 dimension with unbiased estimates of SEM parameters using the aforementioned DWLS estimator. In this proposal, the alternative parametrization proposed by Karl Gustav JÖRESKOG is used as the parameterization to scale the latent continuous indicators [103]. This parameterization imposes identifiability constraints by fixing values of at least two thresholds (to zero and one respectively) in order to be able to estimate the means and error variances of the latent continuous indicators. Once this first stage is accomplished, a second stage is the detection of response shift in itself. As the means and error variances of latent continuous indicators have been estimated, a model can be fitted with the same parameters used for domain-level (i.e. factor loadings, intercepts, error variances) and therefore OP can be accomplished using the same operationalization of response shift as usual.

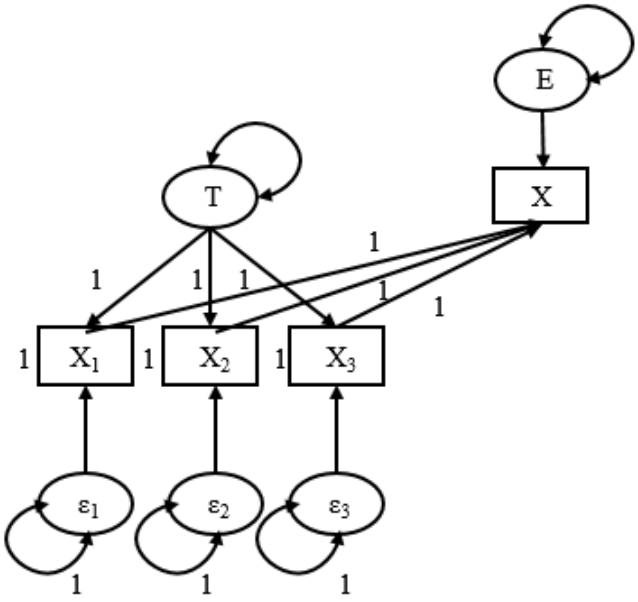
This method has been performed once on an empirical dataset, with each domain of the SF-36 explored one by one due to computational and potential convergence issues [213]. On a statistical level, it indeed used an estimator that has been proved to lead to unbiased SEM parameters values contrariwise to ML estimator when incorporating categorical indicators in a CFA model [99]. Thus, the method is theoretically sound in terms of model estimates and is a valuable addition to the literature on response shift detection. Nonetheless, it seems plausible the operationalization and meaning of response shift detection at item-level can still be discussed on a more conceptual level, especially when confronting these with the fundamental hypotheses of CTT. Especially, a potential issue can be the interpretation of changes in factor loadings over time as reprioritization in the same sense reprioritization was defined as domain-level response shift (i.e. a change in the relative importance of component domains in the target construct [30]).

In CTT, the score (X), which is usually the sum of the values of p items (X_1, X_2, …, X_p), is taken to be an appropriate representation of the true score (T, the latent variable supposed to be measured), measurement error aside (E), which leads to the fundamental equation [80–82]:

$$X = T + E.$$

Each item is taken to be a repeatedly administered test of the measurement of the same construct and the residual factor associated to each item is supposed to be measurement error only. Therefore, with an infinite number of items the expected value of X is supposed to be T. However, this relationship can be assumed to be true in the most restrictive sense only if certain conditions are met regarding the relationships between the indicators and the true score. Ideally, these conditions correspond to what is called in the psychometric literature to a *parallel model*. The parallel model assumes all items are exactly equivalent to one another [216,217]. All items measure the same construct, on the same scale, with the same difficulty, and with the same amount of error (i.e. assumption of unidimensionality plus equal factor loadings, equal intercepts and equal error variances) [218] (Figure 22).

Figure 22. A parallel model as a SEM path diagram displaying the relationships between T, X and E



When applied to the fundamental CTT equation, each item *p* for an individual *i* can be shown as [218]:

$$X_{ip} = T_i + E_i.$$

Least restrictive conditions can be assumed in describing the relationships between the construct and the indicators. In the *tau-equivalent model*, the individual error variances associated to each item can differ from one another [216,217]. All items measure the same construct, on the same scale, with the same difficulty, but with different amounts of measurement error, thus leading to [218]:

$$X_{ip} = T_i + E_{ip}.$$

Then, in the *essentially tau-equivalent model*, the individual difficulties associated to each item can be of different values [216,219]. Here, all items measure the same construct, on the same scale, with different difficulties and different amounts of measurement error. The essentially tau-equivalent model allows each item true score to differ by an additive constant unique to each pair of items [218]:

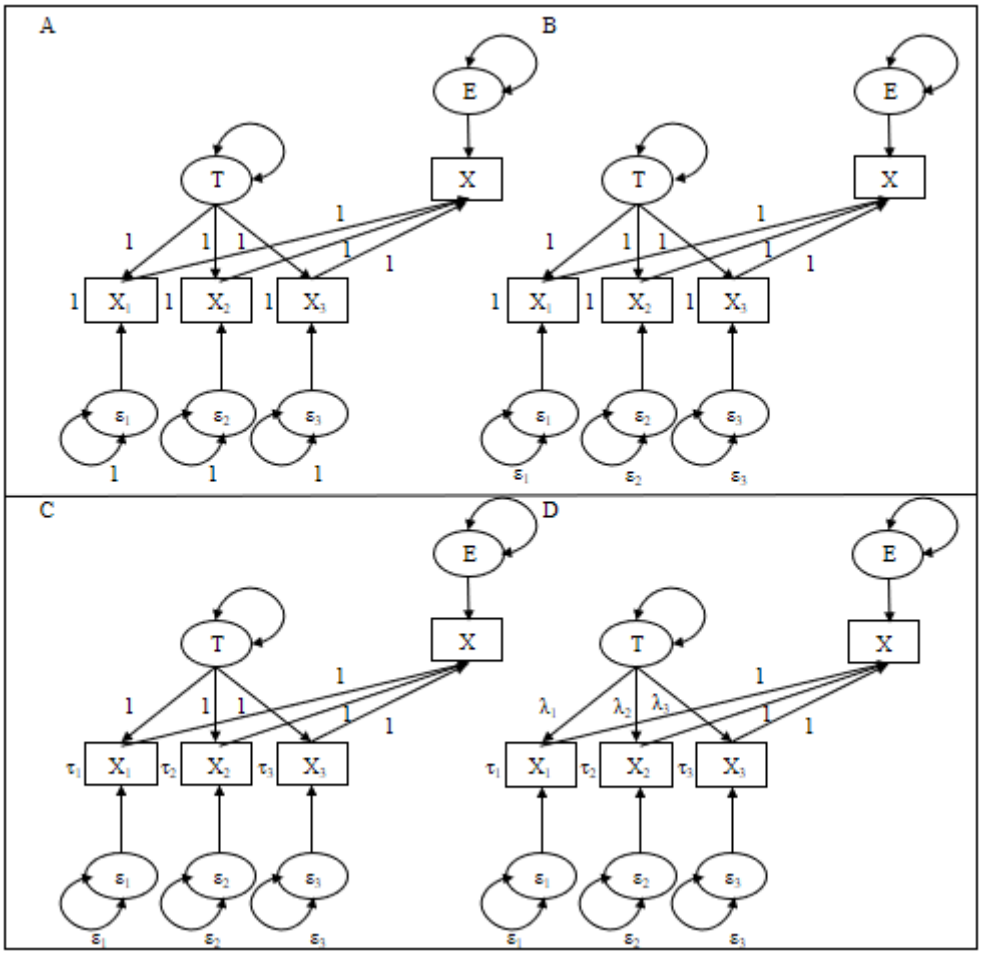
$$X_{ip} = (\alpha_p + T_i) + E_{ip}.$$

Finally, the least restrictive model is the *congeneric model* where the individual factor loadings associated to each items are freed to differ from one another [216]. Here, all items measure the same construct, with different scales, difficulties and amounts of measurement

error. The congeneric model allows each item true score to differ by an additive and multiplicative constant unique to each pair of items [218]:

$$X_{ip} = (\alpha_p + \beta_p(T_i)) + E_{ip}.$$

Figure 23. Path diagrams of A: A parallel model, B: A tau-equivalent model, C: A essentially tau-equivalent model, D: A congeneric model



Thus, when the relationships between a construct and indicators can only be modeled via a congeneric model (i.e. when more restrictive assumptions are untenable), it seems plausible to suppose differences in factor loadings between two items measuring the same construct are linked to differences in scales (i.e. the unit of measurement) between these two-items. As a result, with a congeneric model, if one path from the latent true variable to one of the items is set to 1 (i.e. as identifiability constraint), this indicates the true scores of the other items are expressed in terms of the true score of the fixed item [218]. A higher value of factor

loading is indicative of an item that is measured on a larger scale than another item with a lower value of a factor loading.

Therefore, if this assumption holds and if we extend this reasoning to a longitudinal CFA model with two-times of measurement, it seems plausible to suppose differences in factor loadings over time on the same item measuring the construct (at each time of measurement) are linked to a change in scale over time.

In the VERDAM et al. study, the adaptation of the OP to detect response shift at item-level was used on empirical data with each of the eight domains of the SF-36 modeled one at a time. When a significant change in factor loadings of an item over time was found using the OP, it was labeled as reprioritization response shift [213]. If this change was for example an increase over time, it was interpreted as "*this item has become more indicative of the domain at the second time than the first*", which is factual [213]. Nonetheless, on a theoretical level, reprioritization is defined as a change in the respondent's values (i.e. the relative importance of component domains in the target construct) [30]. But, as aforementioned, changes in factor loadings over time at item-level could be interpreted as changes in the scale of the item designed to measure a construct. So it leads to the question: at item-level, is a change in factor loadings over time can univocally be interpreted as the occurrence of reprioritization response shift with a meaning that can be exactly the same than reprioritization response shift at domain-level? Indeed; if the assumption than changes in factor loadings at item-level are somehow related to changes in scales holds, it could lead to suppose changes in factor loadings are also somehow related with a meaning that is closer to the definition of non-uniform recalibration than reprioritization.

On a more hermeneutical level, it can also lead to the question if item-level response shift can also be interpreted as a phenomenon indicative of psychological adaptation to illness as domain-level response shift, or if it is a phenomenon more directly linked with the violation of some fundamental assumptions of CTT. In that regard, one could easily argue the strictest assumptions of CTT (i.e. the assumption a parallel model holds) are empirically untenable, and therefore modeling adequately responses to items using a latent variable model as SEM or IRT models and searching for changes in parameters over time is a more heuristic approach, as it conveys more information about what have changed in the measurement of a construct over time. Nonetheless, at the very least, it seems plausible the aforementioned arguments could open a discussion about the adequate relationships between changes in SEM parameters over time, the operationalization of these parameters and response shift definition at item-level.

2. Deepen the discussion about links between semantic complexity of a PRO instrument and response shift

2.1 Is recalibration a very critical difference between objective and subjective measures?

In the fifth part of this manuscript (it will be further referred as the "semantic primes study" for simplicity), we hypothesized a link between semantic complexity of constructs and items and reconceptualization, along with a link between multidomain concepts and reprioritization. In short, it has led to a theoretical proposal which can be summarized as if someone could define a unidomain concept measured by items both with the least semantic complexity possible, it could lead to a PRO instrument unaffected by reconceptualization and reprioritization response shift.

Of course, it has to be stressed out it is currently a theoretical proposal only that has not been investigated in any empirical manner. Therefore, the soundness of this proposal remains unknown and unassessed. Nonetheless, assuming it could be a valid proposal, it was mentioned in the discussion the issue of dealing with recalibration response shift (i.e. is a PRO instrument could be unaffected by recalibration?). Indeed, recalibration is related to the metric standard associated with response categories. As aforementioned, setting the metric standard and mapping a judgment onto response categories are part of the judgment and response processes and therefore can appear to be a phenomenon mainly independent of semantic complexity (see Part 6, chapter 6) [200]. Nonetheless, it was proposed in the discussion of the semantic primes paper the use of brief case vignettes illustrating what each response choice is referring to as a suitable response format for an item [20]. The idea was to rendered the metric of the item explicit and external. In some sort, it can be related to the idea of setting the unit of measurement relatively to a standard, as exposed in the Introduction of this manuscript (see Part 1, chapter 1).

However, again, assuming this proposal could be valid in some way, it seems insufficient to effectively deal with the occurrence of recalibration without further development. Indeed, even if brief case vignettes were used to set more objectively the unit of measurement of an item, it could only be a valid option if these vignettes could not be interpreted differently over time or within individuals. Thus, the vignettes should also comply to the requirement of being conceptualized and written with the least semantic complexity possible, which would add another difficulty to the potential practical feasibility of the

theoretical proposals of the semantic primes paper. *In fine*, this remark may be an illustration of the fact the difference between the definition of the unit of measurement of dimensions used to describe objective phenomena (i.e. units defined relatively to a phenomenon unrelated with the internal perspective of a thinking subject) and a PRO instrument (i.e. unit defined relatively to an internal perspective) is perhaps a very critical difference between objective and subjective measures.

2.2. Complexity versus simplicity: expanding the talk about semantic primes theory and response shift to study design and purposes of a study

The semantic primes study was written with a very focused context in mind: the assessment of the direct efficacy of a new therapeutic strategy measured by a subjective construct reflecting patients' experiences as a primary endpoint using RCT design. The relationship between this very focused context and the idea exposed in the study (i.e. could it be possible to design a PRO instrument capturing patients' subjective experience while being unaffected by reconceptualization and reprioritization?) is of paramount importance and need to be discussed with more depth.

Indeed, the occurrence of response shift theory has highlighted the dynamic nature of subjective concepts and measures and the need to better capture psychological adaptation to illness when assessing the longitudinal effect of a salient health-related event on individuals' mind [135]. Humans' mind functioning can be thought as a complex system and response shift theory has forced researchers to disregard an over-simplistic assumption of traditional psychometric theory (i.e. the idea a concept that is assessed is stable in individuals' mind over time) [29]. Thus, response shift theory has led to practical developments. In the recent years, the methodological developments were focused toward complex statistical modeling (see Part 2, chapter 5.2). In any non-experimental setting (e.g. epidemiological study…), this can be a more heuristic approach to try to integrate complex phenomena using complex methods as it probably conveys more information about the intricacies of the numerous processes involved when an individual has to deal with a significant change in health state.

Nonetheless, there is currently a will to use more frequently PRO as primary endpoints in experimental designs such as RCTs [199]. RCT is, in health-related research, the design that is the closest to a traditional design used in experimental physics which is the design that is the closest to the epistemological ideal of the counterfactual account of causality (i.e. if *c* and *e* are two events, *c* causes *e* if (1) *c* and *e* both occur and (2) if *c* had not occurred and all else remained

the same, then *e* would have not occurred) [220,221]. In that regard, RCTs are designed to answer to a single simple question: is a new therapeutic strategy can be proved to have more efficacy than at least one another or not? The answer of a RCT is binary: proof of efficacy or absence of proof [59]. Thus, even if RCTs can be distinguished from designs used in experimental physics (because of the number of cofounders to account for, which is controlled by randomization, and because of the risk of Type-I-error and Type-II-error, which are managed by statistical hypothesis testing theory [11]), it nevertheless corresponds to a traditional simple experimental scientific device used to assess a single and simple causality assumption. Therefore, if the idea of assessing subjective experiences and using them as primary endpoints in RCTs is a sound one, as it recognizes patients as thinking subjects and not just as bodies with a set of biological parameters, admitting the dynamic nature of subjective measures, which is one the purpose of response shift theory, could contradict the simplicity of RCTs design. Indeed, from a methodological standpoint, two key characteristics of designing a RCT are planning the number of subjects needed to be included and concluding if the amount of differences observed regarding the primary endpoint is significant proof of efficacy [11]. Designing a RCT while complying with these two aforementioned methodological characteristics, with the intent of using a subjective concept as primary endpoint, and with the additional requirement of taking into account potential response shift occurrence can therefore appear to be a difficult challenge.

This was this potential challenge which motivated the work on response shift and semantic complexity. Indeed, by investigating if it could be possible to design a PRO unaffected by response shift, it could lead to one of the solution to deal with the difficulty of using subjective concepts as primary endpoints in the context of RCTs. However, as the answer could be the need to design a PRO instrument in a way to assess constructs with the least semantic complexity possible, the downside could be it would require to give up on the idea to explore complex phenomena.

In fine, this balance between complexity and simplicity, and its relationship with study design in health-related research, might be seen as a reflection of an issue that is currently pregnant is various scientific fields: the transition to the investigation of more and more complex objects and therefore the difficulties of managing this complexity [222]. For example, in epidemiology or ecology and sociology, there is an increase will since the end of the 20th century to explore complex functional or causal relationships between multiple agents of wide systems [223], hence a more frequent use of methods derived from graph or network theories

(within which SEM can be considered as a special case) [224,225], statistical learning such as machine learning techniques [226] or artificial intelligence [227] and giving up on traditional simple experimental designs and classical statistical methods.

3. Conclusion

Response shift theory has led researchers in the recent years to admit the dynamic nature of subjective measures and is currently a middle-range theory under which practical developments have been performed to deal with this complex issue. Thus, it has led to substantial theoretical and methodological developments, as well as empirical investigations of the effect of a change in health state on individuals' mind while integrating more complex phenomena regarding psychological adaptation to illness. Nonetheless, the third part of this manuscript (i.e. the state of the art of international works on response shift) has shown it remains a theory with few very stabilized certainties, whether on a theoretical, methodological or empirical level. The simulation study has shown the difficulties to assess the performances of complex statistical modeling techniques, especially the difficulties of translating the results of these types of studies to practical guidelines for empirical studies (i.e. the issue of generalizability). The semantic primes study has led to question the feasibility of incorporating subjective concepts as primary endpoints in RCTs while accounting for response shift occurrence. Moreover, these works have illustrated the fact response shift theory, and in a broader way, psychometrics, are truly transdisciplinary scientific fields. Stabilizing a modern psychometric theory which will fulfill the expected qualities of good measurement devices while integrating the complexity of dealing with humans' psychological processes will remain a thrilling challenge.

References

[1] Godfrey-Smith P. Theory and reality: an introduction to the philosophy of science. Chicago: University of Chicago Press; 2003.

[2] Popper KR. The Logic of scientific discovery. Repr. 2008 (twice). London: Routledge; 2008.

[3] Bernard C, Greene HC, Henderson LJ, Cohen IB. An introduction to the study of experimental medicine. 1. publ. New York: Dover; 1957.

[4] Feinstein AR. Clinical Judgment. Baltimore: Williams and Wilkins; 1967.

[5] Eddy DM. A manual for assessing health practices & designing practice policies: the explicit approach. Philadelphia, Pa: American College of Physicians; 1992.

[6] Evidence-Based Medicine Working Group. Evidence-based medicine. A new approach to teaching the practice of medicine. JAMA 1992;268:2420–5.

[7] Cochrane AL. Effectiveness and efficiency: random reflections on health services. London: Nuffield Provincial Hospitals Trust; 1972.

[8] Haute Autorité de Santé. Guide d'analyse de la littérature et gradation des recommandations 2000.

[9] Rothman KJ, Greenland S, Lash TL, editors. Modern epidemiology. 3., [rev. and updated] ed. Philadelphia, Pa.: Wolters Kluwer, Lippincott Williams & Wilkins; 2008.

[10] Joyner MJ, Paneth N. Seven Questions for Personalized Medicine. JAMA 2015;314:999–1000. doi:10.1001/jama.2015.7725.

[11] Brody T. Clinical trials: study design, endpoints and biomarkers, drug safety, and FDA and ICH guidelines. 2016.

[12] Falissard B. Mesurer la subjectivité en santé: perspective méthodologique et statistique. Issy-les-Moulineaux: Elsevier-Masson; 2008.

[13] Preamble to the Constitution of the World Health Organization as adopted by the International Health Conference, New York, 19-22 June, 1946; signed on 22 July 1946 by the representatives of 61 States (Official Records of the World Health Organization, no. 2, p. 100) and entered into force on 7 April 1948. n.d.

[14] Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. JAMA 1995;273:59–65.

[15] The SAMSI Psychometric Program Longitudinal Assessment of Patient-Reported Outcomes Working Group, Swartz RJ, Schwartz C, Basch E, Cai L, Fairclough DL, et al. The king's foot of patient-reported outcomes: current practices and new developments for the measurement of change. Quality of Life Research 2011;20:1159–67. doi:10.1007/s11136-011-9863-1.

[16] US Food and Drug Administration. Guidance for Industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. 2009.

[17] World Health Organization. The 10 leading causes of death in the world, 2000 and 2012 2014.

[18] Fayers PM, Machin D. Quality of life: the assessment, analysis, and interpretation of patient-reported outcomes. 2nd ed. Chichester ; Hoboken, NJ: J. Wiley; 2007.

[19] Armitage P, Colton T, editors. Encyclopedia of biostatistics. 2nd ed. Chichester, West Sussex, England ; Hoboken, NJ: John Wiley; 2005.

[20] Kadouri A, Corruble E, Falissard B. The improved Clinical Global Impression Scale (iCGI): development and validation in depression. BMC Psychiatry 2007;7:7. doi:10.1186/1471-244X-7-7.

[21] Schwartz MD. Quantum field theory and the standard model. New York: Cambridge University Press; 2014.

[22] van der Lugt A. Imaging tests in determination of brain death. Neuroradiology 2010;52:945–7. doi:10.1007/s00234-010-0765-7.

[23] Alpert JS, Thygesen K, Antman E, Bassand JP. Myocardial infarction redefined--a consensus document of The Joint European Society of Cardiology/American College of Cardiology Committee for the redefinition of myocardial infarction. J Am Coll Cardiol 2000;36:959–69.

[24] World Health Organization. International Statistical Classification of Diseases and Related Health Problems 10th Revision 2016.

[25] American Psychiatric Association, American Psychiatric Association, editors. Diagnostic and statistical manual of mental disorders: DSM-5. 5th ed. Washington, D.C: American Psychiatric Association; 2013.

[26] Allison PJ, Locker D, Feine JS. Quality of life: a dynamic construct. Soc Sci Med 1997;45:221–30.

[27] Bermúdez JL. Cognitive science: an introduction to the science of the mind. Second edition. Cambridge: Cambridge University Press; 2014.

[28] Bureau International des Poids et Mesures, editor. Le système international d'unités (SI) =: The international system of units (SI). 8. éd. Sèvres: BIPM; 2006.

[29] Schwartz CE, Rapkin BD. Reconsidering the psychometrics of quality of life assessment in light of response shift and appraisal. Health and Quality of Life Outcomes 2004;2:16.

[30] Sprangers MAG, Schwartz CE. Integrating response shift into health-related quality of life research: a theoretical model. Social Science & Medicine 1999;48:1507–1515.

[31] Andrykowski M, Brady M, Hunt J. Positive psychosocial adjustment in potential bone narrow transplant recipients: cancer as a psychosocial transition. Psychooncology 1993;2:261–76.

[32] Daltroy LH, Larson MG, Eaton HM, Phillips CB, Liang MH. Discrepancies between self-reported and observed physical function in the elderly: the influence of response shift and other factors. Social Science & Medicine 1999;48:1549–1561.

[33] Merton RK. Social theory and social structure. Enlarged ed., [Nachdr.]. New York, NY: Free Press; 2000.

[34] Ahmed S, Schwartz C, Ring L, Sprangers MAG. Applications of health-related quality of life for guiding health care: advances in response shift research. Journal of Clinical Epidemiology 2009;62:1115–7. doi:10.1016/j.jclinepi.2009.04.006.

[35] Howard GS, Dailey PR, Gulanick NA. The Feasibility of Informed Pretests in Attenuating Response-Shift Bias. Applied Psychological Measurement 1979;3:481–94. doi:10.1177/014662167900300406.

[36] Ubel PA, Peeters Y, Smith D. Abandoning the language of “response shift”: a plea for conceptual clarity in distinguishing scale recalibration from true changes in quality of life. Quality of Life Research 2010;19:465–71. doi:10.1007/s11136-010-9592-x.

[37] Oort FJ. Using structural equation modeling to detect response shifts and true change. Quality of Life Research 2005;14:587–598.

[38] Raykov T. A first course in structural equation modeling. 2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers; 2006.

[39] Oort FJ, Visser MRM, Sprangers MAG. An application of structural equation modeling to detect response shifts and true change in quality of life data from cancer patients undergoing invasive surgery. Quality of Life Research 2005;14:599–609.

[40] Visser MRM, Oort FJ, Sprangers MAG. Methods to detect response shift in quality of life data: a convergent validity study. Qual Life Res 2005;14:629–39.

[41] Ahmed S, Bourbeau J, Maltais F, Mansour A. The Oort structural equation modeling approach detected a response shift after a COPD self-management program not detected by the Schmitt technique. Journal of Clinical Epidemiology 2009;62:1165–72. doi:10.1016/j.jclinepi.2009.03.015.

[42] Barclay-Goddard R, Lix LM, Tate R, Weinberg L, Mayo NE. Response shift was identified over multiple occasions with a structural equation modeling framework. Journal of Clinical Epidemiology 2009;62:1181–8. doi:10.1016/j.jclinepi.2009.03.014.

[43] Barclay-Goddard R, Lix LM, Tate R, Weinberg L, Mayo NE. Health-related quality of life after stroke: does response shift occur in self-perceived physical function? Arch Phys Med Rehabil 2011;92:1762–9. doi:10.1016/j.apmr.2011.06.013.

[44] King-Kallimanis BL, Oort FJ, Nolte S, Schwartz CE, Sprangers MAG. Using structural equation modeling to detect response shift in performance and health-related quality of life scores of multiple sclerosis patients. Quality of Life Research 2011;20:1527–40. doi:10.1007/s11136-010-9844-9.

[45] Schwartz CE, Sprangers MAG, Oort FJ, Ahmed S, Bode R, Li Y, et al. Response shift in patients with multiple sclerosis: an application of three statistical techniques. Quality of Life Research 2011;20:1561–72. doi:10.1007/s11136-011-0056-8.

[46] Nagl M, Farin E. Response shift in quality of life assessment in patients with chronic back pain and chronic ischaemic heart disease. Disability and Rehabilitation 2012;34:671–80. doi:10.3109/09638288.2011.619616.

[47] Fokkema M, Smits N, Kelderman H, Cuijpers P. Response Shifts in Mental Health Interventions: An Illustration of Longitudinal Measurement Invariance. Psychological Assessment 2013;25:520–31. doi:10.1037/a0031669.

[48] Gandhi PK, Ried LD, Huang I-C, Kimberlin CL, Kauf TL. Assessment of response shift using two structural equation modeling techniques. Quality of Life Research 2013;22:461–71. doi:10.1007/s11136-012-0171-1.

[49] Barclay R, Tate RB. Response shift recalibration and reprioritization in health-related quality of life was identified prospectively in older men with and without stroke. Journal of Clinical Epidemiology 2014;67:500–7. doi:10.1016/j.jclinepi.2013.12.003.

[50] Ahmed S, Sawatzky R, Levesque J-F, Ehrmann-Feldman D, Schwartz CE. Minimal evidence of response shift in the absence of a catalyst. Quality of Life Research 2014;23:2421–30. doi:10.1007/s11136-014-0699-3.

[51] Howard GS, Dailey PR. Response-shift bias: A source of contamination of self-report measures. Journal of Applied Psychology 1979;64:144–50. doi:10.1037/0021-9010.64.2.144.

[52] Schwartz CE, Sprangers MA, Fayers PM. Response shift: you know it's there, but how do you capture it? Challenges fo the next phase of research. Assessing quality of life in clinical trials. 2nd edition, Oxford; New-York: Oxford University Press; 2005.

[53] McClimans L, Bickenbach J, Westerman M, Carlson L, Wasserman D, Schwartz C. Philosophical perspectives on response shift. Qual Life Res 2013;22:1871–8. doi:10.1007/s11136-012-0300-x.

[54] Dintiman G, Greenberg J. Health through discovery. Random House. New-York: 1986.

[55] Nordenfelt L. Concepts and measurement of quality of life in health care. 1994.

[56] Frank L, Basch E, Selby JV, Patient-Centered Outcomes Research Institute. The PCORI perspective on patient-centered outcomes research. JAMA 2014;312:1513–4. doi:10.1001/jama.2014.11100.

[57] Methodology Committee of the Patient-Centered Outcomes Research Institute (PCORI). Methodological standards and patient-centeredness in comparative effectiveness research: the PCORI perspective. JAMA 2012;307:1636–40. doi:10.1001/jama.2012.466.

[58] Ziegelstein RC. Personomics. JAMA Intern Med 2015;175:888–9. doi:10.1001/jamainternmed.2015.0861.

[59] McKenna SP. Measuring patient-reported outcomes: moving beyond misplaced common sense to hard science. BMC Medicine 2011;9:86.

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

[61] Rabin R, Charro F de. EQ-SD: a measure of health status from the EuroQol Group. Annals of Medicine 2001;33:337–343.

[62] Zigmond A, Snaith R. The Hospital Anxiety and Depression Scale. Acta Psychiatr Scand 1983;67:361–70.

[63] Mahoney FI, Barthel DW. Functional evaluation: the Barthel Index. Md State Med J 1965;14:61–5.

[64] Galloway S, Bell D, Hamilton C, Scullion A. Well-being and quality of life: measuring the benefits of culture and sport: a litterature review and a thinkpiece. Edinburgh, Scotland: Scottsh Executive Social Research; 2006.

[65] Bakas T, McLennon SM, Carpenter JS, Buelow JM, Otte JL, Hanna KM, et al. Systematic review of health-related quality of life models. Health Qual Life Outcomes 2012;10:134. doi:10.1186/1477-7525-10-134.

[66] Ferrans CE, Zerwic JJ, Wilbur JE, Larson JL. Conceptual Model of Health-Related Quality of Life. Journal of Nursing Scholarship 2005;37:336–342.

[67] Calman K. Quality of life in cancer patients: an hypothesis. J Med Ethics 1984;10:124–7.

[68] O'Boyle C, McGee HM, Hickey A. The Schedule for the Evaluation of Individual Quality of Life (SEIQoL). Administration Manual 1993.

[69] Ruta DA, Garratt AM, Leng M, Russell IT, MacDonald LM. A new approach to the measurement of quality of life. The Patient-Generated Index. Med Care 1994;32:1109–26.

[70] McKenna SP, Doward LC. The Needs-Based Approach to Quality of Life Assessment. Value in Health 2004;7:S1–S3.

[71] Hunt SM, McKenna SP. The QLDS: a scale for the measurement of quality of life in depression. Health Policy 1992;22:307–319.

[72] Bergner M, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. Med Care 1981;19:787–805.

[73] Hunt SM, McKenna SP, McEwen J, Williams J, Papp E. The Nottingham Health Profile: subjective health status and medical consultations. Soc Sci Med A 1981;15:221–9.

[74] Keller SD, Ware JE Jr, Bentler PM, Aaronson NK, Alonso J, Apolone G, et al. Use of structural equation modeling to test the construct validity of the SF-36 Health Survey in ten countries: results from the IQOLA Project. International Quality of Life Assessment. J Clin Epidemiol 1998;51:1179–88.

[75] World Health Organization, editor. International classification of functioning, disability and health: ICF. Geneva: World Health Organization; 2001.

[76] Center for Disease Control and Prevention. Measuring Healthy Days: Population assessment of Health-Related Quality of Life. Atlanta, Goergia: 2000.

[77] Calvert M, Brundage M, Jacobsen PB, Schünemann HJ, Efficace F. The CONSORT Patient-Reported Outcome (PRO) extension: implications for clinical trials and practice. Health and Quality of Life Outcomes 2013;11:184. doi:10.1186/1477-7525-11-184.

[78] Kaplan RM, Saccuzzo DP. Psychological testing: principles, applications, & issues. 8th ed. Belmont, CA: Wadsworth, Cengage Learning; 2013.

[79] Nunnally JC, Bernstein IH. Psychometric theory. 3rd ed. New York: McGraw-Hill; 1994.

[80] Lord FM, Novick MR, Birnbaum A. Statistical theories of mental test scores. Charlotte, NC: Information Age Publ; 2008.

[81] Aitkin MA. Test Theory. D. Magnusson, Reading, Massachusetts: Addison-Wesley Publishing Company, 1966. Journal of the American Statistical Association 1968;63:379–80. doi:10.1080/01621459.1968.11009260.

[82] Gulliksen H. Theory of mental tests. Hillsdale, N.J: L. Erlbaum Associates; 1987.

[83] Novick M. The axioms and principal results of classical test theory. Journal of Mathematical Psychology 1966;3:1–18.

[84] Mellenbergh G. Measurement precision in test score and item response models. Psychological Methods 1996;1:293–9.

[85] Cronbach L. Coefficient alpha and the internal structure of a test. Psychometrika 1951;16:297–334.

[86] Embretson SE. Item response theory for psychologists. Mahwah, N.J: L. Erlbaum Associates; 2000.

[87] Kline T. Psychological testing: a practical approach to design and evaluation. Thousand Oaks, Calif: Sage Publications; 2005.

[88] Borsboom D. Measuring the mind: conceptual issues in modern psychometrics. Cambridge: Cambridge University Press; 2009.

[89] Bollen KA, Hoyle R. Latent Variables in Structural Equation Modeling. Handbook of Structural Equation Modeling, New-York: Guilford Press; 2012, p. 56–67.

[90] Brown TA. Confirmatory factor analysis for applied research. New York: Guilford Press; 2006.

[91] Spearman C. “General Intelligence” Objectively Determined and Measured. American Journal of Psychology 1904;15:201–92.

[92] Thurstone L. Multiple-factor analysis. Journal of Clinical Psychology 1947;4:224–224. doi:10.1002/1097-4679(194804)4:2<224::AID-JCLP2270040225>3.0.CO;2-7.

[93] Edwards JR, Bagozzi RP. On the nature and direction of relationships between constructs and measures. Psychol Methods 2000;5:155–74.

[94] Falissard B. The unidimensionality of a psychiatric scale: A statistical point of view. International Journal of Methods in Psychiatric Research 1999;8:162–167.

[95] Bollen KA. Latent variables in psychology and the social sciences. Annu Rev Psychol 2002;53:605–34. doi:10.1146/annurev.psych.53.100901.135239.

[96] Jöreskog KG. A general approach to confirmatory maximum likelihood factor analysis. Psychometrika 1969;34:183–202. doi:10.1007/BF02289343.

[97] Pui-Wa L, Qiong W. Estimation in Structural Equation Modeling. Handbook of Structural Equation Modeling, New-York: Guilford Press; 2012, p. 164–80.

[98] West S, Taylor A, Wu W. Model Fit and Model Selection in Structural Equation Modeling. Handbook of Structural Equation Modeling, New-York: Guilford Press; 2012, p. 209–31.

[99] Finney SJ, DiStefano C. Nonnormal and categorical data in stuctural equation modeling. Structural Equation Modeling : a second course, Charlotte, NC: IAP, Information Age Publ.; 2013, p. 439–92.

[100] Satorra A, Bentler P. Corrections to test statistics and standards errors in covariance structure analysis. Latent variables analysis: Applications for developmental research. A. von Eye and C.C.Clogg, Thousand Oaks, CA: Sage; 1994, p. 399–419.

[101] Muthén B. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika 1984;49:115–32. doi:10.1007/BF02294210.

[102] Beaujean A. Models with dichotomous indicator variables. Latent variable modeling using R. A step-by-step guide, New-York, NY: Taylor and Francis; 2014, p. 93–113.

[103] Jöreskog KG. Structural equation modeling with ordinal variables using LISREL. 2002.

[104] Hoyle RH, editor. Handbook of structural equation modeling. New York: Guilford Press; 2012.

[105] Wright S. Correlation and Causation. J Agricultural Research 1921;20:557–85.

[106] Wright S. The method of path coefficients. Annals of Mathematical Statistics 1934;5:161–215.

[107] Joreskog KG. A General Method for Analysis of Covariance Structures. Biometrika 1970;57:239. doi:10.2307/2334833.

[108] Ringo Ho M, Stark S, Chernyshenko O. Graphical Representation of Structural Equation Models Using Path Diagrams. Handbook of Structural Equation Modeling, New-York: Guilford Press; 2012, p. 43–55.

[109] De Ayala RJ. The theory and practice of item response theory. New York: Guilford Press; 2009.

[110] Wang W, Chyi I. Gain score in item response theory as an effect size measure. Educational and Psychological Measurement 2004;64:758–80.

[111] Andrich D. Rating scales and Rasch measurement. Expert Rev Pharmacoecon Outcomes Res 2011;11:571–85. doi:10.1586/erp.11.59.

[112] de Bock E, Hardouin J-B, Blanchin M, Le Neel T, Kubis G, Bonnaud-Antignac A, et al. Rasch-family models are more valuable than score-based approaches for analysing longitudinal patient-reported outcomes with missing data. Stat Methods Med Res 2013. doi:10.1177/0962280213515570.

[113] Molenaar I. Some Background for Item Response Theory and the Rasch Model. In G. H. Fisher and I. W. Molenaar. Rasch Models, Foundations, Recent Developments and Applications, New York, NY: Springer-Verlag; 1995, p. 3–14.

[114] Fayers PM, Hays RD. Assessing quality of life in clinical trials : methods and practice. Oxford; New York: Oxford University Press; 2005.

[115] Schwartz CE, Bode R, Repucci N, Becker J, Sprangers MAG, Fayers PM. The clinical significance of adaptation to changing health: a meta-analysis of response shift. Qual Life Res 2006;15:1533–50. doi:10.1007/s11136-006-0025-9.

[116] Leplège A, Hunt S. The problem of quality of life in medicine. JAMA 1997;278:47–50.

[117] McClimans L. A theoretical framework for patient-reported outcome measures. Theoretical Medicine and Bioethics 2010;31:225–40. doi:10.1007/s11017-010-9142-0.

[118] Rapkin BD, Schwartz CE. Toward a theoretical model of quality-of-life appraisal: Implications of findings from studies of response shift. Health Qual Life Outcomes 2004;2:14. doi:10.1186/1477-7525-2-14.

[119] Cronbach L, Furby L. How we should measure “change” - Or should we ? Psychol Bull 1970;74:68–80.

[120] Howard GS, Ralph KM, Gulanick NA, Maxwell SE, Nance DW, Gerber SK. Internal Invalidity in Pretest-Posttest Self-Report Evaluations and a Re-evaluation of Retrospective Pretests. Applied Psychological Measurement 1979;3:1–23. doi:10.1177/014662167900300101.

[121] Golembiewski RT. Measuring Change and Persistence in Human Affairs: Types of Change Generated by OD Designs. The Journal of Applied Behavioral Science 1976;12:133–57. doi:10.1177/002188637601200201.

[122] Breetvelt IS, Van Dam FS. Underreporting by cancer patients: the case of response-shift. Soc Sci Med 1991;32:981–7.

[123] Sprangers MA. Response-shift bias: a challenge to the assessment of patients' quality of life in cancer clinical trials. Cancer Treat Rev 1996;22 Suppl A:55–62.

[124] Helson H. Adaptation Level Theory. New-York: Harper and Row; 1964.

[125] Mishel M. Uncertainty in illness. Journal Nursing Scolarship 1988;20:225–32.

[126] Lazarus R, Folkman S. Stress, Apraisal, and Coping. New-York, NY: Springer; 1984.

[127] Sprangers MAG, Schwartz CE. Do not throw out the baby with the bath water: build on current approaches to realize conceptual clarity. Response to Ubel, Peeters, and Smith. Quality of Life Research 2010;19:477–9. doi:10.1007/s11136-010-9611-y.

[128] Reeve BB. An opportunity to refine our understanding of "response shift" and to educate researchers on designing quality research studies: response to Ubel, Peeters, and Smith. Quality of Life Research 2010;19:473–5. doi:10.1007/s11136-010-9612-x.

[129] Boyer L, Baumstarck K, Michel P, Boucekine M, Anota A, Bonnetain F, et al. Statistical challenges of quality of life and cancer: new avenues for future research. Expert Rev Pharmacoecon Outcomes Res 2014;14:19–22. doi:10.1586/14737167.2014.873704.

[130] Ubel PA, Smith DM. Why should changing the bathwater have to harm the baby? Qual Life Res 2010;19:481–2. doi:10.1007/s11136-010-9613-9.

[131] Eton DT. Why we need response shift: an appeal to functionalism. Quality of Life Research 2010;19:929–30. doi:10.1007/s11136-010-9684-7.

[132] Stanton AL, Revenson TA, Tennen H. Health Psychology: Psychological Adjustment to Chronic Disease. Annual Review of Psychology 2007;58:565–92. doi:10.1146/annurev.psych.58.110405.085615.

[133] Carver CS, Scheier MF. Control theory: a useful conceptual framework for personality, social, clinical and health psychology. Psychological Bulletin 1982;92:111–35.

[134] Carver CS, Scheier MF. Scaling back goals and recalibration of the affect system are processes in normal adaptive self-regulation: understanding “response shift” phenomena. Soc Sci Med 2000;50:1715–22.

[135] Barclay-Goddard R, Epstein JD, Mayo NE. Response shift: a brief overview and proposed research priorities. Quality of Life Research 2009;18:335–46. doi:10.1007/s11136-009-9450-x.

[136] Schwartz CE, Ahmed S, Sawatzky R, Sajobi T, Mayo N, Finkelstein J, et al. Guidelines for secondary analysis in search of response shift. Quality of Life Research 2013;22:2663–73. doi:10.1007/s11136-013-0402-0.

[137] Ring L, Höfer S, Heuston F, Harris D, O’Boyle CA. Response shift masks the treatment impact on patient reported outcomes (PROs): the example of individual quality of life in edentulous patients. Health and Quality of Life Outcomes 2005;3:55.

[138] O’Boyle CA, McGee H, Browne J. Measuring response shift using the schedule for evaluation of individual quality of life. Adaptation to changing health— response shift in quality of life research, American Psychological Association; 2000, p. 123–36.

[139] Osborne RH, Hawkins M, Sprangers MAG. Change of perspective: A measurable and desired outcome of chronic disease self-management intervention programs that violates the premise of preintervention/postintervention assessment. Arthritis & Rheumatism 2006;55:458–65. doi:10.1002/art.21982.

[140] Korfage IJ, de Koning HJ, Essink-Bot M-L. Response shift due to diagnosis and primary treatment of localized prostate cancer: a then-test and a vignette study. Qual Life Res 2007;16:1627–34. doi:10.1007/s11136-007-9265-6.

[141] Korfage IJ, Hak T, de Koning HJ, Essink-Bot M-L. Patients’ perceptions of the side-effects of prostate cancer treatment—A qualitative interview study. Social Science & Medicine 2006;63:911–9. doi:10.1016/j.socscimed.2006.01.027.

[142] Elliott BA, Gessert CE, Larson PM, Russ TE. Shifting responses in quality of life: People living with dialysis. Quality of Life Research 2014;23:1497–504. doi:10.1007/s11136-013-0600-9.

[143] Schmitt N. The Use Of Analysis Of Covariance Structures To Assess Beta And Gamma Change. Multivariate Behavioral Research 1982;17:343–58. doi:10.1207/s15327906mbr1703_3.

[144] Ahmed S, Mayo NE, Corbiere M, Wood-Dauphinee S, Hanley J, Cohen R. Change in quality of life of people with stroke over time: true change or response shift? Qual Life Res 2005;14:611–27.

[145] Oort FJ, Visser MRM, Sprangers MAG. Formal definitions of measurement bias and explanation bias clarify measurement and conceptual perspectives on response shift. Journal of Clinical Epidemiology 2009;62:1126–37. doi:10.1016/j.jclinepi.2009.03.013.

[146] Mayo NE, Scott SC, Dendukuri N, Ahmed S, Wood-Dauphinee S. Identifying response shift statistically at the individual level. Qual Life Res 2008;17:627–39. doi:10.1007/s11136-008-9329-2.

[147] Lix LM, Sajobi TT, Sawatzky R, Liu J, Mayo NE, Huang Y, et al. Relative importance measures for reprioritization response shift. Quality of Life Research 2012;22:695–703. doi:10.1007/s11136-012-0198-3.

[148] Li Y, Schwartz CE. Data mining for response shift patterns in multiple sclerosis patients using recursive partitioning tree analysis. Quality of Life Research 2011;20:1543–53. doi:10.1007/s11136-011-0004-7.

[149] Boucekine M, Loundou A, Baumstarck K, Minaya-Flores P, Pelletier J, Ghattas B, et al. Using the random forest method to detect a response shift in the quality of life of multiple sclerosis patients: a cohort study. BMC Medical Research Methodology 2013;13:20.

[150] Anota A, Bascoul-Mollevi C, Conroy T, Guillemin F, Velten M, Jolly D, et al. Item response theory and factor analysis as a mean to characterize occurrence of response shift in a longitudinal quality of life study in breast cancer patients. Health Qual Life Outcomes 2014;12:32. doi:10.1186/1477-7525-12-32.

[151] Carver CS, Scheier MF. Control theory: a useful conceptual framework for personality-social, clinical, and health psychology. Psychol Bull 1982;92:111–35.

[152] Schwartz CE, Sprangers MAG. Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. Social Science & Medicine 1999;48:1531–1548.

[153] Schwartz CE, Sprangers MAG. Guidelines for improving the stringency of response shift research using the thentest. Quality of Life Research 2010;19:455–64. doi:10.1007/s11136-010-9585-9.

[154] Nolte S, Elsworth GR, Sinclair AJ, Osborne RH. Tests of measurement invariance failed to support the application of the “then-test.” Journal of Clinical Epidemiology 2009;62:1173–80. doi:10.1016/j.jclinepi.2009.01.021.

[155] Schwartz CE, Sprangers MAG, Carey A, Reed G. Exploring response shift in longitudinal data. Psychology & Health 2004;19:51–69. doi:10.1080/0887044031000118456.

[156] Norman G. Hi! How are you? Response shift, implicit theories and differing epistemologies. Qual Life Res 2003;12:239–49.

[157] Ahmed S, Mayo NE, Wood-Dauphinee S, Hanley JA, Cohen SR. Using the patient generated index to evaluate response shift post-stroke. Quality of Life Research 2005;14:2247–2257.

[158] Oort FJ. Towards a Formal Definition of Response Shift (In Reply to G.W. Donaldson). Quality of Life Research 2005;14:2353–5. doi:10.1007/s11136-005-3978-1.

[159] King-Kallimanis BL, Oort FJ, Visser MRM, Sprangers MAG. Structural equation modeling of health-related quality-of-life data illustrates the measurement and conceptual perspectives on response shift. Journal of Clinical Epidemiology 2009;62:1157–64. doi:10.1016/j.jclinepi.2009.04.004.

[160] Ahmed S, Mayo N, Scott S, Kuspinar A, Schwartz C. Using latent trajectory analysis of residuals to detect response shift in general health among patients with multiple sclerosis article. Quality of Life Research 2011;20:1555–60. doi:10.1007/s11136-011-0005-6.

[161] Schwartz CE, Sajobi TT, Lix LM, Quaranto BR, Finkelstein JA. Changing values, changing outcomes: the influence of reprioritization response shift on outcome assessment after spine surgery. Quality of Life Research 2013;22:2255–64. doi:10.1007/s11136-013-0377-x.

[162] Boucekine M, Boyer L, Baumstarck K, Millier A, Ghattas B, Auquier P, et al. Exploring the response shift effect on the quality of life of patients with schizophrenia: an application of the random forest method. Med Decis Making 2015;35:388–97. doi:10.1177/0272989X14559273.

[163] Guilleux A, Blanchin M, Vanier A, Guillemin F, Falissard B, Schwartz CE, et al. RespOnse Shift ALgorithm in Item response theory (ROSALI) for response shift detection with missing data in longitudinal patient-reported outcome studies. Qual Life Res 2015;24:553–64. doi:10.1007/s11136-014-0876-4.

[164] Jansen SJ, Stiggelbout AM, Nooij MA, Noordijk EM, Kievit J. Response shift in quality of life measurement in early-stage breast cancer patients undergoing radiotherapy. Qual Life Res 2000;9:603–15.

[165] Bernhard J, Hürny C, Maibach R, Herrmann R, Laffer U. Quality of life as subjective experience: reframing of perception in patients with colon cancer undergoing radical resection with or without adjuvant chemotherapy. Annals of Oncology 1999;10:775–782.

[166] Hagedoorn M, Sneeuw KCA, Aaronson NK. Changes in physical functioning and quality of life in patients with cancer: response shift and relative evaluation of one's condition. Journal of Clinical Epidemiology 2002;55:176–183.

[167] Ahmed S, Mayo NE, Wood-Dauphinee S, Hanley JA, Cohen SR. Response shift influenced estimates of change in health-related quality of life poststroke. J Clin Epidemiol 2004;57:561–70. doi:10.1016/j.jclinepi.2003.11.003.

[168] Ahmed S, Mayo NE, Wood-Dauphinee S, Hanley JA, Cohen SR. The structural equation modeling technique did not show a response shift, contrary to the results of the then test and the individualized approaches. J Clin Epidemiol 2005;58:1125–33. doi:10.1016/j.jclinepi.2005.03.003.

[169] Joore MA, Potjewijd J, Timmerman AA, Anteunis LJC. Response shift in the measurement of quality of life in hearing impaired adults after hearing aid fitting. Qual Life Res 2002;11:299–307.

[170] Timmerman AA, Anteunis LJC, Meesters CMG. Response-shift bias and parent-reported quality of life in children with otitis media. Arch Otolaryngol Head Neck Surg 2003;129:987–91. doi:10.1001/archotol.129.9.987.

[171] Finkelstein JA, Razmjou H, Schwartz CE. Response shift and outcome assessment in orthopedic surgery: is there a difference between complete and partial treatment? J Clin Epidemiol 2009;62:1189–90. doi:10.1016/j.jclinepi.2009.03.022.

[172] Razmjou H, Schwartz CE, Yee A, Finkelstein JA. Traditional assessment of health outcome following total knee arthroplasty was confounded by response shift phenomenon. J Clin Epidemiol 2009;62:91–6. doi:10.1016/j.jclinepi.2008.08.004.

[173] Razmjou H, Schwartz CE, Holtby R. The impact of response shift on perceived disability two years following rotator cuff surgery. J Bone Joint Surg Am 2010;92:2178–86. doi:10.2106/JBJS.I.00990.

[174] Postulart D, Adang EM. Response shift and adaptation in chronically ill patients. Med Decis Making 2000;20:186–93.

[175] Gandhi PK, Ried LD, Kimberlin CL, Kauf TL, Huang I-C. Influence of explanatory and confounding variables on HRQoL after controlling for measurement bias and response shift in measurement. Expert Rev Pharmacoecon Outcomes Res 2013;13:841–51. doi:10.1586/14737167.2013.852959.

[176] Schwartz CE, Sendor RM. Helping others helps oneself: response shift effects in peer support. Social Science & Medicine 1999;48:1563–1575.

[177] Schwartz CE. Teaching coping skills enhances quality of life more than peer support: results of a randomized trial with multiple sclerosis patients. Health Psychol 1999;18:211–20.

[178] Visser MRM, Oort FJ, Lanschot JJB, Velden J, Kloek JJ, Gouma DJ, et al. The role of recalibration response shift in explaining bodily pain in cancer patients undergoing invasive surgery: an empirical investigation of the Sprangers and Schwartz model. Psycho-Oncology 2012:n/a-n/a. doi:10.1002/pon.2114.

[179] Li Y, Rapkin B. Classification and regression tree uncovered hierarchy of psychosocial determinants underlying quality-of-life response shift in HIV/AIDS. Journal of Clinical Epidemiology 2009;62:1138–47. doi:10.1016/j.jclinepi.2009.03.021.

[180] Bandalos DL, Gagné P. Simulation Methods in Structural Equation Modeling. Handbook of Structural Equation Modeling, New York, NY: Guilford Press; 2012, p. 92–110.

[181] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria: 2015.

[182] Barendse MT, Oort FJ, Werner CS, Ligtvoet R, Schermelleh-Engel K. Measurement Bias Detection Through Factor Analysis. Structural Equation Modeling: A Multidisciplinary Journal 2012;19:561–79. doi:10.1080/10705511.2012.713261.

[183] Barendse MT, Oort FJ, Garst GJA. Using restricted factor analysis with latent moderated structures to detect uniform and nonuniform measurement bias; a simulation study. AStA Advances in Statistical Analysis 2010;94:117–27. doi:10.1007/s10182-010-0126-1.

[184] Woods CM, Grimm KJ. Testing for Nonuniform Differential Item Functioning With Multiple Indicator Multiple Cause Models. Applied Psychological Measurement 2011;35:339–61. doi:10.1177/0146621611405984.

[185] Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control 1974;19:716–23. doi:10.1109/TAC.1974.1100705.

[186] Schwarz G. Estimating the Dimension of a Model. The Annals of Statistics 1978;6:461–4. doi:10.1214/aos/1176344136.

[187] Sclove L. Application of model-selection criteria to some problems in multivariate analysis . Psychometrika 1987:333–43.

[188] Fischer G., Molenaar I. Rasch models: foundation, recent developments, and applications. New-York: Springer; 1995.

[189] Sébille V, Hardouin J-B, Le Neel T, Kubis G, Boyer F, Guillemin F, et al. Methodological issues regarding power of classical test theory and IRT-based approaches for the comparison of Patient-Reported Outcome measures – A simulation study. BMC Med Res Methodol 2010:10–24.

[190] Rosseel Y. lavaan: An R package for structural equation modeling. Journal of Statistical Software 2012;48:1–36.

[191] Schermelleh-Engel K, Moosbrugger H, Müller H. Evaluating the fit of Structural Equation Models: Tests of significance and descriptive goodness of fit measures. Methods of Psychological Research Online 2003;8:23–74.

[192] Bryant FB, Satorra A. Principles and Practice of Scaled Difference Chi-Square Testing. Structural Equation Modeling: A Multidisciplinary Journal 2012;19:372–98. doi:10.1080/10705511.2012.687671.

[193] Lehmann EL. Testing statistical hypotheses. 3rd ed. New York: Springer; 2008.

[194] Hu L, Bentler P. Evaluating model fit. Structural equation modeling. Concepts, issues, and applications, London: Sage; 1995, p. 76–99.

[195] Gregorich SE. Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. Medical Care 2006;44:S78.

[196] Beaujean A. Models with dichotomous indicator variables. Latent variable modeling using R. A step-by-step guide, New-York, NY: Taylor and Francis; 2014, p. 93–113.

[197] Cassileth B, Lusk E, Tenaglia A. A psychological comparison of patients with melanoma and other dermatological disorders. American Academy of Dermatology 1984;7:742–6.

[198] Mayo NE, Scott SC, Ahmed S. Case management poststroke did not induce response shift: the value of residuals. Journal of Clinical Epidemiology 2009;62:1148–56. doi:10.1016/j.jclinepi.2009.03.020.

[199] Sanders C, Egger M, Donovan J, Tallon D, Frankel S. Reporting on quality of life in randomised controlled trials: bibliographic study. BMJ: British Medical Journal 1998;317:1191.

[200] Tourangeau R, Rips LJ, Rasinski KA. The psychology of survey response. Cambridge, U.K.; New York: Cambridge University Press; 2000.

[201] Westerman MJ, The A-M, Sprangers MAG, Groen HJM, van der Wal G, Hak T. Small-cell lung cancer patients are just "a little bit" tired: response shift and self-presentation in the measurement of fatigue. Qual Life Res 2007;16:853–61. doi:10.1007/s11136-007-9178-4.

[202] Ricœur P. Hermeneutics and the human sciences: essays on language, action, and interpretation. Cambridge [Eng.] ; New York : Paris: Cambridge University Press ; Editions de la Maison des sciences de l'homme; 1981.

[203] Wierzbicka A. Semantics: primes and universals. Oxford [England] ; New York: Oxford University Press; 1996.

[204] Goddard C, Wierzbicka A. Semantic primes and cultural scripts in language learning and intercultural communication. Applied Cultural Linguistics: Implications fro second language learning and intercultural communication, Amsterdam: Gary Palmer and Farzad Sharifian; 2007, p. 105–24.

[205] Goddard C. Semantic primes, semantic molecules, semantic templates: Key concepts in the NSM approach to lexical typology. Linguistics 2012;50:711–43. doi:10.1515/ling-2012-0022.

[206] Brow JP, McGee HM, O'Boyle CA. Conceptual approaches to the assessment of quality of life. Psychology & Health 1997;12:737–51. doi:10.1080/08870449708406736.

[207] Leventhal H, Colman S. Quality of life: A process view. Psychology & Health 1997;12:753–67. doi:10.1080/08870449708406737.

[208] Ware JE Jr, Kosinski M, Gandek B, Aaronson NK, Apolone G, Bech P, et al. The factor structure of the SF-36 Health Survey in 10 countries: results from the IQOLA Project. International Quality of Life Assessment. J Clin Epidemiol 1998;51:1159–65.

[209] McColl E, Meadows K, Barofsky I. Cognitive aspects of survey methodology and quality of life assessment. Qual Life Res 2003;12:217–8.

[210] Jobe JB. Cognitive psychology and self-reports: Models and methods. Quality of Life Research 2003;12:219–227.

[211] Collins D. Pretesting survey instruments: an overview of cognitive methods. Quality of Life Research 2003;12:229–238.

[212] Bjorner JB, Ware JE, Kosinski M. The potential synergy between cognitive models and modern psychometric models. Quality of Life Research 2003;12:261–274.

[213] Verdam MGE, Oort FJ, Sprangers MAG. Using structural equation modeling to detect response shifts and true change in discrete variables: an application to the items of the SF-36. Qual Life Res 2016;25:1361–83. doi:10.1007/s11136-015-1195-0.

[214] Muhén L., Muthén B. Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén; 2012.

[215] Schwartz CE. Introduction to special section on response shift at the item level. Qual Life Res 2016;25:1323–5. doi:10.1007/s11136-016-1299-1.

[216] Raykov T. Estimation of Composite Reliability for Congeneric Measures. Applied Psychological Measurement 1997;21:173–84. doi:10.1177/01466216970212006.

[217] Raykov T. Scale Reliability, Cronbach's Coefficient Alpha, and Violations of Essential Tau-Equivalence with Fixed Congeneric Components. Multivariate Behavioral Research 1997;32:329–53. doi:10.1207/s15327906mbr3204_2.

[218] Graham JM. Congeneric and (Essentially) Tau-Equivalent Estimates of Score Reliability: What They Are and How to Use Them. Educational and Psychological Measurement 2006;66:930–44. doi:10.1177/0013164406288165.

[219] Miller MB. Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. Structural Equation Modeling: A Multidisciplinary Journal 1995;2:255–73. doi:10.1080/10705519509540013.

[220] Lewis D. Causation. Journal of Philosophy 1973;70:556–67.

[221] Morgan SL, Winship C. Counterfactuals and causal inference: methods and principles for social research. New York: Cambridge University Press; 2007.

[222] Pearl J. Causality: models, reasoning, and inference. Cambridge, U.K. ; New York: Cambridge University Press; 2000.

[223] Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology 1999;10:37–48.

[224] Glymour M, Greenland S. Causal diagrams. Modern Epidemiology. Third Edition. Lippincott Williams and Wilkins, Philadelphia, Pa.: 2008, p. 183–209.

[225] Grace JB, Schoolmaster DR, Guntenspergen GR, Little AM, Mitchell BR, Miller KM, et al. Guidelines for a graph-theoretic implementation of structural equation modeling. Ecosphere 2012;3:art73. doi:10.1890/ES12-00048.1.

[226] James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning. vol. 103. New York, NY: Springer New York; 2013.

[227] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. New York, NY: Springer New York; 2009.

UNIVERSITE BRETAGNE LOIRE

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Antoine VANIER

The concept, measurement and integration of response shift phenomenon in Patient-Reported Outcomes data analyses. On certain methodological and statistical considerations

Abstract

Patient-Reported Outcomes are increasingly used in health-related research. These instruments allow the assessment of subjective concepts such as Health-Related Quality of Life, anxiety level, pain or fatigue.
Initially, the interpretation of a difference in score over time was based on the assumption that the meaning of concepts and measurement scales remains stable in individuals' minds over time. This assumption has been challenged. Indeed, the self-assessment of a concept is now understood as a contingency of the subjective meaning a subject has of this concept, which can change over time especially as a result of a salient medical event: the "response shift" phenomenon.
Since the end of the 1990s, researches on response shift phenomenon has become of prime interest in the field of health-related research. If developments have been made, it is still a young field with various scientific debates on a theoretical, methodological and statistical level. Thus, the broad objective of this thesis is to investigate some methodological and statistical issues regarding response shift concept, detection and integration into PRO data analyses.
The manuscript is composed of three main works: a state of the art and synthesis of the works conducted at an international level since response shift phenomenon is investigated, a pilot study investigating the statistical performances of the Oort's Procedure (a popular method of response shift detection using Structural Equation Modeling) by simulations and a theoretical work about the links between response shift occurrence and semantic complexity of concepts measured and items used.

Key Words: Response shift, Patient-Reported Outcomes, Health-Related Quality of Life, Structural Equation Modeling, Methodology, Psychometrics

Résumé

Les données rapportées par les patients sont maintenant fréquemment utilisées en recherche biomédicale. Ces instruments permettent la mesure de concepts subjectifs tels que la qualité de vie, les niveaux d'anxiété, de douleur, de fatigue.
L'interprétation d'une différence de score au cours du temps était basée sur l'hypothèse que le sens des concepts et échelles restai stable au cours du temps dans l'esprit des individus. Cette hypothèse semble aujourd'hui dépassée. L'auto-évaluation d'un concept est maintenant comprise comme contingente de la représentation subjective qu'à un sujet du dit concept, cette représentation pouvant changer au cours du temps, surtout après avoir vécu un évènement de santé : ce phénomène est connu comme le « response shift ».
Depuis la fin des années 1990s, l'investigation de ce phénomène est devenue un sujet d'intérêt majeur en psychométrie. Si des développements ont vu le jour, ce sujet reste récent et donc accompagné de débats variés que ce soit sur le plan théorique ou méthodologique. Aussi, l'objectif général de cette thèse est d'investiguer certaines problématiques méthodologiques et statistiques liées au response shift.
Ce manuscrit est composé de trois travaux principaux : un état de l'art et une synthèse des travaux conduits à un niveau international depuis que le response shift est étudié, une étude pilote des performances de la procédure d'Oort (une méthode populaire de détection de response shift) par simulations et un travail théorique sur les liens entre response shift et complexité sémantique des concepts mesurés et items utilisés.

Mots clés : Response shift, Patient-Reported Outcomes, Health-Related Quality of Life, Structural Equation Modeling, Methodology, Psychometrics