Comparaison des approches CTT et IRT pour l'analyse des effets temps et groupe de données longitudinales de type Patient-Reported Outcomes et impact du dropout

Myriam Blanchin

UNIVERSITE DE NANTES

UFR DE SCIENCES PHARMACEUTIQUES

ECOLE DOCTORALE BIOLOGIE SANTE

Année 2011

N ° attribué par la bibliothèque

Comparaison des approches CTT et IRT pour l'analyse des effets temps et groupe de données longitudinales de type Patient-Reported Outcomes et impact du dropout

THESE DE DOCTORAT

Discipline : Biologie, médecine et santé

Spécialité : Sciences physico-chimiques et ingénierie appliquée à la santé

Présentée et soutenue publiquement par

Myriam BLANCHIN

le 10 juin 2011, devant le jury ci-dessous,

Président	M. Mounir Mesbah, Professeur, Paris VI
Rapporteur	Mme Hélène Jacqmin-Gadda, Directrice de Recherche, Bordeaux
Rapporteur	M. Bruno Falissard, PU-PH, Paris-Sud
Examinateur	M. Alain Leplège, Professeur, Paris VII
Directrice de thèse	Mme Véronique Sébille-Rivain, Professeur, Nantes
Co-encadrant de thèse	M. Jean-Benoit Hardouin, Maître de Conférences, Nantes

Table des matières

Table des figures

Liste des tableaux

Glossaire

Notation	Description	
AIC	Akaike Information Criterion	17
AR(1)	structure auto-régressive d'ordre 1	16
ARH(1)	structure auto-régressive hétérogène d'ordre 1	16
BIC	Bayesian Information Criterion	17
CAT	Computerized Adaptative Testing ou testing adaptatif sur ordinateur	34
CML	Conditional Maximum Likelihood ou maximum de vraisemblance conditionnelle	36
CS	structure 'compound symmetry'	16
CSH	structure 'compound symmetry' hétérogène	16
CTT	Classical Test Theory ou théorie classique des tests	27
DIF	Differential Item Functioning ou fonctionnement différentiel de l'item	34
EAP	Expected A Posteriori Estimator ou estimateur attendu a posteriori de Bayes	38
ICC	Item Characteristic Curve ou courbe caractéristique de l'item	32
IRT	Item Response Theory ou théorie de réponse aux items	30
JML	Joint Maximum Likelihood ou maximum de vraisemblance jointe	35

Notation	Description	
LRM	méthode Longitudinal Rasch Mixed model	55
MAR	Missing At Random - manquant au hasard	19
MCAR	Missing Completely At Random - manquant complètement au hasard	19
ML	Maximum Likelihood - maximum de vraisemblance	14
MLE	Maximum Likelihood Estimator ou estimateur du maximum de vraisemblance	37
MML	Marginal Maximum Likelihood ou maximum de vraisemblance marginale	36
MNAR	Missing Non At Random - manquant non au hasard	20
PCM	Partial Credit Model	40
PCML	Pairwise Conditional Maximum Likelihood	36
PRO	Patient-Reported Outcomes	25
PV	méthode Plausible Values model	53
REML	REstricted Maximum Likelihood - maximum de vraisemblance restreinte	14
RM	méthode Rasch and Mixed Models	53
RSM	Rating Scale Model	39
SM	méthode Score and Mixed Models	52
UN	structure sans contraintes - unstructured	16
WLE	Weighted Likelihood Estimator ou estimateur de la vraisemblance pondéré	38

Introduction

Les avancées médicales du vingtième siècle ont vu l'apparition de traitements permettant de guérir des maladies aigües mortelles comme certaines infections. La proportion de maladies chroniques pour lesquelles il n'existe pas de traitement curatif s'en est trouvée augmentée. Après la seconde guerre mondiale, l'amélioration des traitements et l'allongement de la survie des patients mais aussi les effets secondaires parfois lourds induits par les traitements sont venus poser la question de la qualité de la vie des patients. Le monde médical a alors souhaité pouvoir évaluer l'impact de la maladie et des traitements sur les dimensions physique, psychologique, familiale et sociale de la qualité de vie. En raison du caractère subjectif et multidimensionnel de la qualité de vie, il a fallu développer des instruments de mesure adaptés. Ce développement est désormais bien avancé avec l'existence de plusieurs instruments de mesure largement utilisés dans le domaine de la recherche clinique et de l'épidémiologie. Les importants développements déjà réalisés permettent maintenant l'évaluation de l'impact de la maladie et des traitements sur la qualité de vie des patients.

En France, un plan pour l'amélioration de la qualité de vie des personnes atteintes de maladies chroniques a été mis en place en 2007. Il fait partie des cinq plans stratégiques prévus par la loi du 9 août 2004 relative à la politique de santé publique. Dans ce plan, le nombre de personnes atteintes de maladies chroniques en France est estimé à 15 millions, soit près de 20% de la population. On compte dans les maladies chroniques :

- des maladies comme l'insuffisance rénale chronique, les bronchites chroniques, l'asthme, les maladies cardio-vasculaires, le cancer ou le diabète

- des maladies rares, comme la mucoviscidose
- des maladies transmissibles persistantes, comme le Sida ou l'hépatite C
- mais aussi des troubles mentaux de longue durée (dépression, schizophrénie,...) ou la douleur chronique

Un des quatre objectifs de ce plan est de mieux connaître les conséquences de la maladie sur la qualité de vie des patients. Dans ce but, il est nécessaire de pouvoir étudier l'évolution de la qualité de vie des patients tout au long de leur maladie et la décrire avec des méthodes appropriées.

De la qualité de vie aux Patient-Reported Outcomes

La terminologie associée à la notion de qualité de vie a évolué dans le temps. De multiples définitions de la qualité de vie existent en raison de son caractère multidimensionnel. L'organisation Mondiale de la Santé donnera en 1996 [82] la définition suivante de la qualité de vie :

> "la qualité de vie, c'est la perception qu'a un individu de sa place dans l'existence, dans le contexte de la culture et du système de valeurs dans lequel il vit, en relation avec ses objectifs, ses attentes, ses normes et ses inquiétudes. Il s'agit d'un large champ conceptuel, englobant de manière complexe la santé physique de la personne, son état psychologique, son niveau d'indépendance, ses relations sociales, ses croyances personnelles et sa relation avec les spécificités de son environnement."

La notion de qualité de vie est étudiée dans de nombreuses disciplines telles que la philosophie, l'économie, les sciences politiques. Dans le domaine de la santé, on s'intéresse à la qualité de vie liée à la santé (Health Related Quality of Life - HRQoL), c'est-à-dire aux effets de la maladie et des traitements sur certaines dimensions de la qualité de vie. En santé, les études ne s'intéressent pas uniquement à cette notion de qualité de vie liée à la santé mais aussi aux notions de bien-être, d'anxiété, de dépression, de douleur, de satisfaction ou de santé perçue. Le terme de Patient-Reported

Outcomes (PRO) a émergé dans les années 2000 [99]. Ce terme générique rassemble les mesures de tout aspect de la santé d'un patient ou de son traitement, rapporté directement par le patient lui-même. Acquadro et al. [2] regroupent sous le terme PRO les mesures de la qualité de vie liée à la santé, les symptômes, la satisfaction du patient, le bien-être psychologique et l'adhésion au traitement, entre autres. Dans ce travail, l'intérêt se porte sur l'analyse de mesures de type Patient-Reported Outcomes.

Problèmes liés à l'analyse des PRO

Théorie classique des tests et théorie de réponse aux items

Les PRO sont évalués à travers les réponses à des questionnaires, remplis par les patients eux-mêmes. Les réponses aux items, qui sont observables, permettent de mesurer une variable non directement mesurable et observable, le trait latent qui peut représenter des concepts tels que la qualité de vie, l'intelligence, l'état de santé, la douleur, la pharmacodépendance... Les PRO sont, par définition, des mesures subjectives dont le développement et l'analyse sont basés sur des théories issues de la psychométrie : la théorie classique des tests (CTT) et la théorie de réponse aux items (IRT). L'approche classique [64] est basée sur un score, généralement calculé en sommant les réponses aux items. Il est supposé mesurer la vraie valeur du critère d'intérêt (la qualité de vie par exemple). Dans la théorie de réponse aux items, la probabilité de répondre positivement à un item en fonction de la variable latente est utilisée pour représenter ce critère.

Parmi les modèles issus de l'IRT, le modèle de Rasch [90], adapté aux items dichotomiques (réponses de type oui/non), est couramment employé en raison de sa simplicité et de ses propriétés psychométriques, notamment l'objectivité spécifique et l'obtention de mesure d'intervalles. La propriété d'objectivité spécifique facilite les comparaisons puisque les propriétés du questionnaire sont indépendantes de la population étudiée. De même, l'estimation du concept est indépendante de l'ensemble d'items, issus du questionnaire, utilisé pour la mesure rendant l'estimation possible en cas de données manquantes, si l'individu n'a pas répondu à l'ensemble d'items. Les mesures obtenues

avec l'IRT (et avec le modèle de Rasch en particulier) sont des mesures d'intervalle. L'écart entre deux mesures est indépendante du niveau de ces mesures sur l'échelle. Il est donc possible de déterminer si deux intervalles ont la même longueur et donc si les mesures entre des individus sont séparées par la même distance.

Les approches CTT et IRT sont très employées pour le développement et la validation de questionnaires [19, 100, 16]. Dans la pratique, l'usage de la CTT semble plus répandu pour analyser les PRO. Ceci peut s'expliquer par la facilité à raisonner au niveau individuel en CTT, ce qui permet son usage dans le domaine médical. Un clinicien peut aisément calculer le score d'un individu et ainsi avoir de manière assez directe une évaluation du niveau du concept étudié. Il peut également comparer le score obtenu à des scores de référence dans des populations données. Il est plus difficile d'appréhender les modèles IRT en raison de leur complexité et du raisonnement au niveau populationnel plutôt qu'individuel.

Toutefois, lorsque le questionnaire utilisé pour l'étude a été validé à la fois en CTT et en IRT, l'analyse avec un modèle IRT pourrait améliorer la précision des estimations des critères mesurés et ainsi fournir des méthodes plus précises et plus puissantes que la CTT. De plus, il paraît dommage, au moment de l'analyse, de se priver des propriétés spécifiques des modèles issus de l'IRT, en particulier de ceux issus de la famille de Rasch. La stratégie d'analyse la plus adéquate n'a pas encore été clairement identifiée à ce jour. Il paraît donc nécessaire de comparer les deux approches pour déterminer si l'analyse avec un modèle IRT est plus adéquate que l'analyse avec un modèle CTT pour des données validées à la fois en CTT et en IRT.

Lawson [55] fut un des premiers à comparer de manière empirique la CTT et l'IRT sur 3 jeux de données différents. Fan [31] a ensuite étendu cette première étude par la comparaison des deux approches dans une analyse transversale de données réelles de tests de mathématiques et de lecture. MacDonald et Paunonen [65] ont mené une étude de simulation dans laquelle sont étudiées la comparabilité, l'invariance et la précision des paramètres pour les deux approches. Ces études ont toutes été menées, en phase de construction et de validation, sur des données transversales et en sciences

de l'éducation. Dans le domaine de la santé, peu de travaux ont été réalisés. On trouve des comparaisons des approches en phase de validation ou de réduction de questionnaires [80, 87]. En phase d'analyse, une étude de simulation compare plusieurs méthodes basées sur l'IRT sur des mesures répétées en deux temps mais l'approche IRT n'est pas comparée à l'approche CTT [43]. Nous nous proposons de comparer les approches CTT et IRT en phase d'analyse dans le cadre de mesures répétées dans le temps en santé.

Données longitudinales quantitatives

L'intérêt pour les PRO est croissant en raison du besoin de suivi des maladies chroniques. Dans ce contexte, il est souvent intéressant d'étudier l'évolution du critère d'intérêt au cours du temps. Or, dans le cas de données longitudinales, la corrélation entre les mesures effectuées en chacun des temps pour chaque patient est à prendre en compte dans l'analyse à travers des méthodes appropriées. L'analyse des données longitudinales au moyen d'un modèle linéaire mixte [117, 41] permet de faire face à ce type de données et est devenue la méthode la plus utilisée pour l'analyse de l'évolution d'un critère quantitatif au cours du temps. Le modèle linéaire mixte a l'avantage d'être un modèle très flexible permettant notamment la modélisation des variations intra- et inter-individuelles séparément et l'ajout de covariables dépendantes du temps ou non. Il est également adapté à l'analyse de données déséquilibrées en cas de mesures inégalement espacées dans le temps ou de données incomplètes sous certaines conditions, relatives au mécanisme de données manquantes rencontré.

Gestion des données incomplètes

Comme toutes les études, les études longitudinales sont confrontées au problème des données manquantes. Ces données manquantes peuvent suivre différents schémas. Quelques réponses peuvent manquer à un temps donné pour un individu ou un individu peut ne pas être observé à un temps donné puis être observé à nouveau les temps suivants. On parle alors de données manquantes intermittentes. De plus, dans les études longitudinales, il est fréquent que des patients soient perdus de vue

("dropout") et sortent prématurément de l'étude. A partir d'un temps donné, aucune donnée n'est plus disponible pour ces patients. Les données incomplètes ont des répercussions sur l'analyse car elles entraînent une perte d'information. De plus, les données manquantes peuvent être informatives lorsqu'elles sont associées au processus de mesure du critère étudié et doivent et peuvent introduire du biais dans l'analyse si elles ne sont pas prises en compte. La validité de l'analyse choisie dépendra du mécanisme qui sous-tend la survenue de données manquantes. Dans le cadre de ce travail, l'intérêt se portera sur les sorties d'étude, informatives ou non.

Objectif et plan du mémoire

Le principal objectif de ce travail est de déterminer la méthode la plus adéquate pour analyser des Patient-Reported Outcomes recueillis de manière longitudinale et issus d'un questionnaire validé avec un modèle de Rasch. Cette méthode doit pouvoir être appliquée aux études d'évolution d'un critère de type PRO au cours du temps fréquemment rencontrées en santé, notamment pour le suivi des maladies chroniques. Il s'agit donc

- de prendre en compte le caractère latent du critère d'intérêt en utilisant des méthodes basées sur les approches existant en psychométrie : la théorie classique des tests et la théorie de réponse aux items

- de modéliser la corrélation entre les mesures d'un même patient en utilisant des méthodes adéquates en conjonction avec la CTT et l'IRT

- d'étudier les problèmes liés aux données manquantes informatives ou non, fréquemment rencontrées en pratique, et leur gestion possible en CTT et en IRT

Dans ce travail, différentes méthodes d'analyse de données longitudinales, basées sur la CTT ou l'IRT, sont comparées au moyen d'études de simulation. Ces études permettent d'évaluer les performances de chaque méthode en terme de risque de première espèce ou de puissance des tests et de biais des estimations des effets temps et groupe.

Dans une première partie, l'ensemble des méthodes statistiques utilisées et développées dans ce travail sont exposées. Le chapitre 1 présente le modèle linéaire mixte utilisé pour l'analyse de données longitudinales quantitatives gaussiennes. Il décrit également la typologie des données manquantes telle qu'elle a été définie par Little et Rubin [63] et la notion d'informativité des données manquantes. Enfin, les méthodes d'analyse les pus utilisées pour la gestion des données incomplètes sont présentées.
Le chapitre 2 présente les principales notions de la théorie classique des tests et de la théorie de réponse aux items. Le modèle de Rasch, issu de l'IRT, ainsi que ses propriétés psychométriques sont décrits de manière plus détaillée. Les deux approches sont ensuite comparées d'un point de vue théorique.

La deuxième partie de ce travail s'attachera à comparer différentes méthodes d'analyse de données longitudinales basées sur la CTT ou l'IRT au moyen de plusieurs études de simulation se plaçant dans des cadres différents. Le chapitre 3 décrit le socle commun à toutes les études de simulation présentées dans les chapitres 4, 5 et 6. Ce chapitre présente notamment les méthodes d'analyse utilisées, la méthode de simulation des données ainsi que les critères utilisés pour l'évaluation et la comparaison des méthodes d'analyse.
Dans le chapitre 4, les résultats d'une première étude de simulation sont présentés. Cette étude s'attache à comparer différentes méthodes d'analyse dans un cadre favorable de données complètes afin d'évaluer leur capacité à correctement analyser l'évolution d'un critère latent au cours du temps.
L'impact de la quantité de données manquantes et du mécanisme sous-jacent sur les performances des différentes méthodes d'analyse lorsque les données ont été recueillies dans un ou deux groupes de patients est évalué dans les chapitres 5 et 6 respectivement.

Les méthodes comparées dans les études de simulation sont appliquées sur deux jeux de données réelles dans le chapitre 7. La première application se place dans un cadre proche de celui des simulations. Une dimension issue d'un questionnaire de qualité de vie générique et formé d'items dichotomiques est étudiée. Au contraire, la deuxième application se place dans un cadre plus large. Les dimensions étudiées

sont issues d'un questionnaire de qualité de vie spécifique aux patients atteints de pathologie cancéreuse et sont composées d'items polytomiques.

Les résultats de chaque étude de simulation seront discutés dans les chapitres correspondants. Le dernier chapitre présentera une discussion plus générale des résultats en synthétisant les forces et les limites des méthodes utilisées et décrira les perspectives de ce travail.

Première partie

Etat des connaissances

Chapitre 1

Analyse de données longitudinales quantitatives

En santé, il est fréquent d'étudier l'évolution d'un critère d'intérêt au cours du temps pour suivre, par exemple, l'évolution d'une maladie chronique ou les effets à long terme des traitements. Souvent, le critère d'intérêt est une variable quantitative. Par exemple, les marqueurs biologiques habituellement utilisés pour étudier l'évolution du VIH sont le taux de CD4 et la charge virale. Le suivi après transplantation rénale s'intéresse à l'évolution de la clairance de la créatinine. Le retentissement d'une maladie ou de ses traitements dans la vie des patients peut être mesuré à travers l'évolution de scores de qualité de vie.
Pour étudier son évolution, les mesures du critère d'intérêt sont répétées dans le temps sur les mêmes patients. L'analyse de telles données nécessite l'emploi de méthodes adéquates car la corrélation entre les mesures d'un même patient doit être prise en compte tout en permettant l'estimation de l'évolution du critère au cours du temps.

L'une des premières méthodes utilisée pour l'analyse de données quantitatives longitudinales fut basée sur l'analyse de variance développée au début du vingtième siècle par Fisher [40]. L'ANOVA à mesures répétées inclut un effet sujet aléatoire qui permet la prise en compte de la corrélation entre les mesures d'un même patient. La structure de la matrice de variance-covariance est très restrictive puisque seule une matrice de type 'compound symmetry' peut être utilisée. L'hypothèse est faite que les variances

sont égales à chaque temps et que les covariances sont les mêmes quel que soit l'espace de temps entre les mesures. Ces hypothèses sont relâchées dans le cas de l'utilisation de la MANOVA à mesures répétées. Pour cette méthode, aucune hypothèse n'est faite sur la structure de la matrice de variance-covariance. Ces deux méthodes présentent l'inconvénient majeur de ne pouvoir être utilisée qu'en cas de données équilibrées en terme de nombre de mesures et de temps d'observation. L'équilibre des données est plus facilement atteint dans des domaines expérimentaux tels que l'industrie ou l'agriculture où les unités expérimentales sont plus facilement contrôlables.
En santé, l'unité d'étude est le patient et il est fréquent de ne pas avoir le même nombre de mesures pour tous les patients car ils ne sont pas venus à toutes les visites prévues ou d'avoir des mesures irrégulièrement espacées dans le temps.
L'analyse des données longitudinales au moyen d'un modèle linéaire mixte permet de faire face à ce type de données et est devenue la méthode la plus utilisée pour l'analyse de l'évolution d'un critère quantitatif au cours du temps.

1.1 Modèle linéaire mixte

Le modèle linéaire mixte consiste à modéliser l'évolution de la moyenne de la population au cours du temps par la spécification d'effets fixes. La corrélation entre les mesures d'un même individu est prise compte en spécifiant la structure de la matrice de variance-covariance. Les évolutions individuelles au cours du temps sont spécifiées au moyen d'effets aléatoires modélisant une déviation individuelle par rapport à l'évolution moyenne de la population.

1.1.1 Modèle

Laird et Ware [54] sont considérés comme les premiers à avoir souligné l'intérêt de l'utilisation des modèles linéaires mixtes pour l'analyse de données longitudinales, particulièrement en santé. Le modèle proposé par Laird et Ware est une classe flexible de modèles linéaires mixtes pour les données longitudinales basée sur les travaux d'Harville [51].

La formulation du modèle est la suivante. Soit $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{in_i})$ (i=1,...,N) le vecteur des n_i réponses aux mesures répétées sur l'individu i, le modèle linéaire mixte s'écrit

$$\begin{aligned} &\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i, \\ &\boldsymbol{b}_i \sim N_q(0, \boldsymbol{D}), \\ &\boldsymbol{e}_i \sim N_{n_i}(0, \boldsymbol{\Sigma}_i), \\ &\boldsymbol{b}_1, ..., \boldsymbol{b}_N, \boldsymbol{e}_1, ..., \boldsymbol{e}_N \textit{ indépendants} \end{aligned} \tag{1.1}$$

où p est le nombre de paramètres des effets fixes, q est le nombre de paramètres des effets aléatoires, $\boldsymbol{X}_i$ est la matrice d'expérience des effets fixes de taille $n_i \times p$, $\boldsymbol{Z}_i$ est la matrice d'expérience des effets aléatoires de taille $n_i \times q$, $\boldsymbol{\beta}$ est le vecteur des paramètres des effets fixes de taille $(p \times 1)$, $\boldsymbol{b}_i$ est le vecteur des paramètres des effets aléatoires de taille $(q \times 1)$, $\boldsymbol{e}_i$ est le vecteur des erreurs de taille $(n_i \times 1)$, $\boldsymbol{D}$ est la matrice de variance-covariance des effets aléatoires de taille $(q \times q)$, $\boldsymbol{\Sigma}_i$ est la matrice de variance-covariance des erreurs de taille $(n_i \times n_i)$.

L'équation 1.1 entraîne l'écriture du modèle sous sa forme marginale :

$$\boldsymbol{Y}_i \sim N_{n_i}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \boldsymbol{\Sigma}_i) \tag{1.2}$$

1.1.2 Estimation des paramètres

La méthode usuelle pour l'inférence est basée sur le modèle marginal à partir d'estimateurs du maximum de vraisemblance. Les paramètres du modèle à estimer sont $\boldsymbol{\beta}$, les composants de la moyenne et $\boldsymbol{\omega}$, les composants de la variance, formé par tous les paramètres de variance-covariance de $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \boldsymbol{\Sigma}_i$.

La vraisemblance marginale s'écrit

$$L_{ML}(\boldsymbol{\beta}, \boldsymbol{\omega}) = \prod_{i=1}^{N} (2\pi)^{-\frac{n_i}{2}} |V_i(\boldsymbol{\omega})|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})' V_i^{-1}(\boldsymbol{\omega})(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right\} \tag{1.3}$$

Le plus souvent, les méthodes d'estimation des paramètres utilisées sont l'estimation par maximum de vraisemblance (ML) ou l'estimation par maximum de vraisemblance restreinte (REML). Les estimateurs du maximum de vraisemblance sont obtenus par maximisation du log de la vraisemblance (équation 1.3).

Un certain nombre de travaux ont été réalisés pour étudier les performances de ces méthodes d'estimation. L'estimateur du maximum de vraisemblance est biaisé pour les composants de la variance [114]. Patterson et Thompson [85] ont proposé la méthode du maximum de vraisemblance restreinte (REML) comme alternative à la méthode du maximum de vraisemblance. Dans la méthode REML, le biais d'estimation des composants de la variance ω est corrigé par l'introduction d'un terme supplémentaire dans la vraisemblance. L'estimateur du maximum de vraisemblance restreinte est obtenu en maximisant le logarithme de la vraisemblance suivante :

$$L_{REML}(\boldsymbol{\beta}, \boldsymbol{\omega}) = \left| \sum_{i=1}^{N} \boldsymbol{X}_i' V_i^{-1}(\boldsymbol{\omega}) \boldsymbol{X}_i \right|^{-1/2} \times L_{ML}(\boldsymbol{\beta}, \boldsymbol{\omega}) \tag{1.4}$$

L'estimateur ML est invariant aux changements de paramétrisation des effets fixes au contraire de l'estimateur REML. En effet, le terme supplémentaire de la vraisemblance REML dépend la spécification du modèle. L'estimateur ML est utilisé pour comparer des modèles emboîtés pour la structure de la moyenne.
L'estimateur REML donne un estimateur moins biaisé des composants de la variance que l'estimateur ML [42]. Il est recommandé d'utiliser les estimations REML quand l'intérêt est dans l'estimation des composants de la variance. Pour comparer des modèles emboîtés pour la structure de covariance, on utilise l'AIC et le BIC basé sur l'estimation REML de modèles ayant la même structure de moyenne.

Des algorithmes itératifs sont utilisés pour maximiser la vraisemblance. L'algorithme EM proposé par Dempster et al. [25] permet d'obtenir les estimations ML ou REML. Cependant, l'algorithme de Newton-Raphson est de plus en plus utilisé car sa convergence est plus rapide que l'algorithme EM [58].

1.1.3 Inférence sur les effets fixes

L'estimation du vecteur des effets fixes $\boldsymbol{\beta}$ est donnée par l'estimateur des moindres carrés généralisés

$$\hat{\beta}(\omega) = \left(\sum_{i=1}^{N} X_i' V_i^{-1}(\hat{\omega}) X_i \right)^{-1} \sum_{i=1}^{N} X_i' V_i^{-1}(\hat{\omega}) y_i \tag{1.5}$$

en remplaçant ω par son estimation ML ou REML. Conditionnellement à ω et sous le modèle marginal (equation 1.2), $\hat{\beta}(\omega)$ suit une loi normale de moyenne β et de matrice de variance-covariance $var(\hat{\beta}) = \left(\sum_{i=1}^{N} X_i' V_i^{-1}(\hat{\omega}) X_i \right)^{-1}$. L'estimation de la matrice de variance-covariance β est obtenue en remplaçant ω par son estimation ML ou REML.

Pour chaque paramètre β_j $(j = 1, ..., p)$, un test de Wald peut être réalisé et s'écrit :

$$H_0 : \beta_j = 0 \; versus \; H_1 : \beta_j \neq 0 \tag{1.6}$$

Sous H_0, $(\hat{\beta}_j - \beta_j)/s.e.(\hat{\beta}_j) \sim N(0,1)$ avec $s.e.(\hat{\beta}_j) = \sqrt{var(\hat{\beta}_j)}$.

Pour une matrice de contrastes L donnée, le test de Wald s'écrit :

$$H_0 : L\beta = 0 \; versus \; H_1 : L\beta \neq 0 \tag{1.7}$$

$$W = (\hat{\beta} - \beta)' L' \left[L \left(\sum_{i=1}^{N} X_i' V_i^{-1}(\hat{\omega}) X_i \right)^{-1} L' \right]^{-1} L(\hat{\beta} - \beta) \tag{1.8}$$

Sous H_0, W suit approximativement une loi du χ^2 à rang(L) degrés de liberté.

Des tests de Student et de Fisher peuvent être préférés aux tests de Wald. En effet, la statistique du test de Wald utilise les estimations des erreurs standards qui sous-estiment la variabilité des effets fixes. Les tests de Student et de Fisher permettent de corriger ce biais à travers l'estimation du nombre de degrés de liberté du dénominateur des statistiques de ces tests.

Pour chaque paramètre β_j $(j = 1, ..., p)$, un test de Student peut être réalisé et s'écrit :

$$H_0 : \beta_j = 0 \; versus \; H_1 : \beta_j \neq 0 \tag{1.9}$$

Sous H_0, $(\hat{\beta}_j - \beta_j)/s.e.(\hat{\beta}_j)$ suit approximativement une loi de Student à d degrés de liberté. Le nombre de degrés de liberté d est estimé à partir des données. Plusieurs méthodes d'estimation du nombre de degrés de liberté existent. Le logiciel SAS utilise entre autres l'approximation de Satterthwaite [95] et l'approximation de Kenward et Roger [53]. Dans le cadre de l'analyse de données longitudinales, il semble que les grands nombres de degrés de liberté généralement obtenus permettent d'obtenir des p-value similaires quelle que soit la méthode d'estimation du nombre de degrés de liberté choisie [115].

Pour une matrice de contrastes L donnée, le test de Fisher s'écrit :

$$H_0 : L\beta = 0 \; versus \; H_1 : L\beta \neq 0 \tag{1.10}$$

Sous H_0, $W/rang(L)$ suit approximativement une loi de Fisher à rang(L) et d degrés de liberté, où W réfère à la statistique de l'équation 1.8 et d réfère au nombre de degrés de libertés estimé à partir des données comme dans le test de Student (équation 1.9).

1.1.4 Composants de la variance

La matrice de variance-covariance des erreurs $\mathbf{\Sigma}_i$ peut être égale à $\sigma^2 I_{n_i}$ lorsque les erreurs sont supposées indépendantes entre elles et de même variance. Cette structure est appelée 'Variance Components' (VC). Ces hypothèses sur la structure peuvent être relâchées et une autre structure peut être choisie pour $\mathbf{\Sigma}_i$. Les structures les plus souvent utilisées sont la structure auto-régressive d'ordre 1 (AR(1)) , la structure auto-régressive hétérogène d'ordre 1 (ARH(1)) , la structure 'compound symmetry' (CS) comme dans l'ANOVA à mesures répétées, la structure 'compound symmetry' hétérogène (CSH) ou la structure sans contraintes (UN) .

Un test de rapport de vraisemblance basé sur les estimations du maximum de vraisemblance restreint REML peut être utilisé pour comparer des modèles emboîtés avec la même structure de moyenne mais des structures de covariance différentes. Les modèles non emboîtés sont comparés à l'aide de critères d'information : l'Akaike Information Criterion (AIC) [3] ou le Bayesian Information Criterion ou Schwarz Criterion (BIC) [97] .

1.1.5 Robustesse

La mauvaise spécification du modèle de covariance a très peu d'impact sur les estimations des effets fixes $\hat{\beta}$. En revanche, l'estimation des variances des effets fixes $var(\hat{\beta})$ peuvent être biaisées [41]. L'inférence sur les effets fixes peut donc être faussée par une mauvaise spécification de la covariance. Liang et Zeger [57] ont proposé l'estimateur sandwich pour $var(\hat{\beta})$, robuste aux mauvaises spécifications de la covariance si la moyenne est correctement spécifiée.
Les estimateurs des effets fixes sont consistants dans le cas de non-normalité des effets aléatoires [113]. Dans ce cas, une correction de type sandwich est également utilisée pour corriger les estimations de la variance.

Le modèle linéaire mixte a l'avantage d'être un modèle très flexible permettant notamment la modélisation des variations intra- et inter-individuelles séparément et l'ajout de covariables dépendantes du temps ou non. Comme il est basé sur le principe de vraisemblance, ces estimateurs sont consistants, asymptotiquement normaux et efficaces. Enfin, il est adapté à l'analyse de données déséquilibrées en cas de mesures inégalement espacées dans le temps ou de données incomplètes sous certaines conditions, relatives au mécanisme de données manquantes rencontré, détaillées dans la section suivante.

1.2 Données incomplètes

Comme toutes les études, les études longitudinales sont confrontées au problème des données manquantes. Ces données manquantes peuvent suivre différents schémas. Quelques réponses peuvent manquer à un temps donné pour un individu ou un individu peut ne pas être observé à un temps donné puis être observé à nouveau les temps suivants. On parle alors de données manquantes intermittentes. De plus, dans les études longitudinales, il est fréquent que des patients soient perdus de vue ("dropout") et sortent prématurément de l'étude. A partir d'un temps donné, aucune donnée n'est plus disponible pour ces patients. Dans le cadre de ce travail, l'intérêt se portera sur les sorties d'étude.
Les données incomplètes ont des répercussions sur l'analyse car elles entraînent une perte d'information et peuvent introduire du biais dans l'analyse si elles ne sont pas prises en

compte. La validité de l'analyse choisie dépendra du mécanisme qui sous-tend la survenue de données manquantes.

1.2.1 Typologie

La distinction entre trois types de données manquantes a été introduite par Rubin [91, 63]. Dans une étude où Y_{ij} mesures sont prévues pour chaque individu i $(i = 1, ..., N)$ à chaque temps de mesure t_{ij} $(j = 1, ..., n_i)$, le vecteur $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{in_i})'$ désigne l'ensemble des mesures qui auraient été observées si aucune donnée n'avait été manquante ou *données complètes*. Soit R_{ij} l'indicateur de données manquantes défini comme suit,

$$R_{ij} = \begin{cases} 1 & \text{si } Y_{ij} \text{ est observé} \\ 0 & \text{sinon} \end{cases} \tag{1.11}$$

Pour chaque individu i, les indicateurs de données manquantes sont groupés dans un vecteur $\boldsymbol{R}_i$. Il est alors possible de partitionner $\boldsymbol{Y}_i$ en deux sous-vecteurs : $\boldsymbol{Y}_i^o$ le vecteur des données observées contenant les Y_{ij} pour lesquels $R_{ij} = 1$ et $\boldsymbol{Y}_i^m$ le vecteur des données manquantes contenant les données restantes. Dans une étude, seuls les $\boldsymbol{Y}_i^o$ sont observés ainsi que les $\boldsymbol{R}_i$. X_i désigne la matrice des variables explicatives complètement observées. La typologie de Rubin est basée sur la factorisation de la densité suivante,

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, \boldsymbol{\phi}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i | X_i, \boldsymbol{\phi}) f(\boldsymbol{r}_i | \boldsymbol{y}_i, X_i, \boldsymbol{\psi}) \tag{1.12}$$

où $\boldsymbol{\phi} = (\boldsymbol{\beta}', \boldsymbol{\omega}')'$ est le vecteur des paramètres du processus de mesure et $\boldsymbol{\psi}$ est le vecteur des paramètres du processus de données manquantes.
Nous pouvons également écrire :

$$f(\boldsymbol{r}_i | \boldsymbol{y}_i, X_i, \boldsymbol{\psi}) = f(\boldsymbol{r}_i | \boldsymbol{y}_i^o, \boldsymbol{y}_i^m, X_i, \boldsymbol{\psi}) \tag{1.13}$$

Les trois types de données manquantes sont les suivants :

- Un processus de données manquantes est dit *manquant complètement au hasard* ou missing completely at random (MCAR) si la probabilité d'avoir des données man-

quantes sur $\boldsymbol{Y}$ ne dépend pas des données observées $\boldsymbol{y}_i^o$ ou non observées $\boldsymbol{y}_i^m$.

$$f(\boldsymbol{r}_i|\boldsymbol{y}_i, X_i, \boldsymbol{\psi}) = f(\boldsymbol{r}_i|X_i, \boldsymbol{\psi}) \tag{1.14}$$

Dans un contexte d'évaluation longitudinale de qualité de vie, les sorties d'étude de patients sont complètement aléatoires si elles sont indépendantes du niveau de qualité de vie des évaluations précédant la sortie d'étude et au moment de la sortie d'étude.

- Un processus de données manquantes est dit *manquant au hasard* ou missing at random (MAR) si la probabilité d'avoir des données manquantes sur $\boldsymbol{Y}$ est indépendante des données non observées $\boldsymbol{y}_i^m$, conditionnellement aux données observées $\boldsymbol{y}_i^o$.

$$f(\boldsymbol{r}_i|\boldsymbol{y}_i, X_i, \boldsymbol{\psi}) = f(\boldsymbol{r}_i|\boldsymbol{y}_i^o, X_i, \boldsymbol{\psi}) \tag{1.15}$$

Dans un contexte d'évaluation longitudinale de qualité de vie, les sorties d'étude de patients sont aléatoires si elles ne dépendent pas du niveau de qualité de vie au moment de la sortie d'étude mais du niveau de qualité de vie des évaluations précédant la sortie d'étude.

- Un processus de données manquantes est dit *manquant non au hasard* ou missing non at random (MNAR) si la probabilité d'avoir des données manquantes sur $\boldsymbol{Y}$ dépend des données non observées $\boldsymbol{y}_i^m$. Il peut également dépendre des données observées $\boldsymbol{y}_i^o$. Dans un contexte d'évaluation longitudinale de qualité de vie, les sorties d'étude de patients sont non manquantes au hasard si elles sont dépendantes du niveau de qualité de vie qui aurait été observé au moment de la sortie d'étude.

1.2.2 Informativité/Ignorabilité

Dans un contexte d'estimation par vraisemblance, les données de type MCAR et MAR sont dites *ignorables* et les données de type MNAR sont dites *informatives*. En réalité, ceci est vrai si la condition de séparabilité est vérifiée pour les données MAR [116], c'est-à-dire si $\boldsymbol{\phi}$ et $\boldsymbol{\psi}$ sont disjoints, en raison de la factorisation développée ci-dessous.

Soit L la vraisemblance pour les données observées, elle s'écrit :

$$L(\boldsymbol{\phi}, \boldsymbol{\psi} | X_i, \boldsymbol{y}_i^o, \boldsymbol{r}_i) \propto f(\boldsymbol{y}_i^o, \boldsymbol{r}_i | X_i, \boldsymbol{\phi}, \boldsymbol{\psi}) \tag{1.16}$$

avec

$$f(\boldsymbol{y}_i^o, \boldsymbol{r}_i | X_i, \boldsymbol{\phi}, \boldsymbol{\psi}) = \int f(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, \boldsymbol{\phi}, \boldsymbol{\psi}) d\boldsymbol{y}_i^m \tag{1.17}$$

$$= \int f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m | X_i, \boldsymbol{\phi}) f(\boldsymbol{r}_i | \boldsymbol{y}_i^o, \boldsymbol{y}_i^m, X_i, \boldsymbol{\psi}) d\boldsymbol{y}_i^m \tag{1.18}$$

Sous l'hypothèse MAR, on a

$$f(\boldsymbol{r}_i | \boldsymbol{y}_i, X_i, \boldsymbol{\psi}) = f(\boldsymbol{r}_i | \boldsymbol{y}_i^o, X_i, \boldsymbol{\psi}) \tag{1.19}$$

ce qui donne,

$$f(\boldsymbol{y}_i^o, \boldsymbol{r}_i | X_i, \boldsymbol{\phi}, \boldsymbol{\psi}) = \int f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m | X_i, \boldsymbol{\phi}) f(\boldsymbol{r}_i | \boldsymbol{y}_i^o, X_i, \boldsymbol{\psi}) d\boldsymbol{y}_i^m \tag{1.20}$$

$$= f(\boldsymbol{y}_i^o | X_i, \boldsymbol{\phi}) f(\boldsymbol{r}_i | \boldsymbol{y}_i^o, X_i, \boldsymbol{\psi}) \tag{1.21}$$

Ainsi, la vraisemblance des données observées se factorise en deux composantes tout comme la densité de l'équation 1.12. Le premier facteur représente la densité marginale du processus de mesure des données observées et le deuxième facteur représente la densité du processus de données manquantes conditionnellement aux données observées. Si la condition de séparabilité est vérifiée, c'est-à-dire si $\boldsymbol{\phi}$ et $\boldsymbol{\psi}$ sont disjoints, alors l'inférence peut être basée uniquement sur la densité marginale du processus de mesure des données observées. Le processus de données manquantes MAR est alors ignorable.

Lorsque les données manquantes sont ignorables, les estimateurs obtenus par maximum de vraisemblance sur les données observées $\boldsymbol{y}_i^o$ sont asymptotiquement sans biais [62]. Dans le cas de données manquantes informatives, il est nécessaire de modéliser conjointement le modèle de mesure et le processus de données manquantes.

1.2.3 Méthodes d'analyse

Face aux problèmes posés par la présence de données manquantes lors des analyses, plusieurs techniques de traitement des données manquantes ont été développées.

Analyse des cas complets

L'analyse des cas complets consiste à inclure dans l'analyse uniquement les cas pour lesquels toutes les mesures ont été observées. Cette méthode a l'avantage d'être simple à mettre en œuvre. Evidemment, elle souffre de gros inconvénients. Outre la perte d'information induite par la suppression des cas incomplets, cette analyse n'est valide que si les données sont de type MCAR.

Analyse des cas disponibles

Ce type d'analyse utilise autant de données disponibles que possible. Si elle a l'avantage de prendre en compte plus d'information que l'analyse des cas complets, elle peut mener à des matrices de variance-covariance non définies positives. De plus, elle n'est valide que dans les cas de données MCAR ou de données MAR si celles-ci sont ignorables et que le modèle a été correctement spécifié [75].

Techniques d'imputation

Afin de pouvoir travailler sur des données complètes sans perdre des cas, il est fréquent d'imputer pour remplacer les valeurs manquantes. L'imputation est dite *simple* lorsqu'une seule valeur est imputée pour chaque valeur manquante. La valeur imputée peut être basée sur des informations disponibles sur l'individu (comme dans le cas de l'utilisation de la last observation carried forward où la dernière valeur observée de l'individu est utilisée pour imputer), sur les autres individus (comme pour l'imputation de la moyenne où la moyenne des valeurs observées de la variable sur les autres individus est imputée) ou sur l'ensemble des individus. Les techniques d'imputation simple ont l'avantage de pouvoir permettre l'utilisation des méthodes d'analyses habituelles sur les données complètes. En revanche, l'incertitude liée à l'imputation des valeurs manquantes n'est pas prise en compte. La variabilité est en général sous-estimée quelle que soit la technique utilisée [118]. De plus, si le modèle d'imputation choisi est incorrect, les estimations obtenues seront biaisées.

L'imputation *multiple* [92] consiste à remplacer chaque valeur manquante par plusieurs valeurs probables générant ainsi plusieurs jeux de données complètes. L'analyse est alors réalisée sur l'ensemble des jeux. Les estimations du paramètre d'intérêt sont alors combinées adéquatement pour obtenir une seule estimation. L'estimation du paramètre d'intérêt est la moyenne des estimations obtenues à chaque imputation. La variance est une somme pondé-

rée des variances intra- et inter-imputation. La multiplicité de l'imputation vise à réduire la sous-estimation de la variabilité inhérente aux techniques d'imputation simple.

Modélisation conjointe du modèle de mesure et du processus de données manquantes

L'ensemble des techniques précédentes ne sont pas adaptées à l'analyse de données longitudinales dont les données manquantes sont informatives. Deux grands types de modèle sont utilisés pour l'analyse de telles données : les modèles de sélection et les modèles de mélange de schémas d'observation (Pattern Mixture models).

Ces deux types de modèle reposent sur des factorisations différentes de la densité conjointe du processus de mesure et du processus de données manquantes. Les modèles de sélection factorisent la densité en deux facteurs comme dans l'équation 1.12. Le premier facteur représente la densité marginale du processus de mesure et le deuxième facteur représente la densité du processus de données manquantes $\boldsymbol{r}_i$ conditionnellement aux données observées $\boldsymbol{y}_i^o$ et non observées $\boldsymbol{y}_i^m$. Les modèles proposés par Diggle and Kenward [26] pour les données longitudinales continues et par Molenberghs et al. [76] pour les données ordinales reposent sur cette factorisation. Les modèles de sélection permettent de modéliser directement les deux facteurs d'intérêt de l'analyse mais nécessitent un modèle pour le vecteur des données complètes $\boldsymbol{Y}_i$. Les résultats des modèles de sélection sont donc sensibles aux hypothèses faites sur le modèle de mesure ainsi que sur le modèle du processus de données manquantes. Le principal inconvénient de ces modèles est de devoir poser des hypothèses sur le processus de données manquantes qui ne peuvent être testées.

Les modèles de mélange de schémas d'observation ont été proposé comme une alternative aux modèles de sélection [45, 60]. Ils sont basés sur la factorisation suivante :

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, \boldsymbol{\phi}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i, | \boldsymbol{r}_i, X_i, \boldsymbol{\phi}) f(\boldsymbol{r}_i | X_i, \boldsymbol{\psi}) \tag{1.22}$$

La densité du processus de mesure est spécifiée conditionnellement au schéma d'observation ou profil de données manquantes. La distribution marginale de $\boldsymbol{Y}$ est donc un mélange de distributions. Les modèles de mélange sont souvent sous-identifiées car l'ensemble des paramètres n'est pas identifiable pour tous les schémas d'observation. Des contraintes d'identification ont été proposées, notamment par Little [61] et Molenberghs et al.[77]. Tout comme

pour les modèles de sélection, les hypothèses faites pour les modèles de mélange ne sont pas vérifiables.

Les modèles de sélection et les modèles de mélange sont deux factorisations différentes de la même distribution. Mais l'obligation de poser des hypothèses dans ces deux modèles entraîne des résultats différents d'un modèle à l'autre. Lors de l'utilisation de ces modèles, une analyse de sensibilité est recommandée par l'emploi d'un modèle de mélange pour évaluer la sensibilité d'un modèle de sélection et inversement [69, 70].

Chapitre 2

Théorie classique des tests et théorie de réponse aux items

2.1 Patient-Reported Outcomes

Le terme Patient-Reported Outcomes (PRO) semble avoir émergé dans les années 2000 [99] suite aux travaux du groupe d'harmonisation des PRO de la Food and Drug Agency (FDA) [2]. Ce terme est un terme générique qui rassemble les mesures de tout aspect de la santé d'un patient ou de son traitement, rapporté directement par le patient lui-même. Acquadro et al. [2] regroupent sous le terme PRO les mesures de la qualité de vie liée à la santé, l'anxiété, la dépression, les symptômes, la satisfaction du patient, le bien-être psychologique et l'adhésion au traitement, entre autres.

L'intérêt pour les PRO est croissant en raison du besoin de suivi des maladies chroniques ou l'absence de marqueurs biologiques pour le suivi de certaines pathologies. Et même lorsque des marqueurs biologiques existent, l'évaluation de PRO peut être une source d'information complémentaire sur l'efficacité d'un traitement en rapportant la fréquence et la sévérité des symptômes et des effets secondaires et leur impact sur les activités et le bien-être du patient. L'intérêt pronostique des PRO a également été discuté, notamment en oncologie, où ce type de mesures pourrait se révéler être un bon prédicteur de la survie [46]. L'intégration de plus en plus fréquente de PRO dans les essais cliniques et l'existence d'une multitude de définitions du terme PRO ont amené les autorités de régulation à mettre à disposition des guides pour l'utilisation des PRO dans l'industrie [112, 20].

Les PRO sont évalués au moyen de questionnaires. On utilise le terme *échelle* pour désigner l'ensemble des questions ou *items* formant un questionnaire. Lorsque l'échelle mesure un seul concept, elle est dite unidimensionnelle. Si elle mesure plusieurs concepts, elle est multidimensionnelle. Les items sont regroupés en *dimensions* mesurant chacune un concept. Les items peuvent être *dichotomiques* (à deux modalités de réponse de type "oui"/"non") ou *polytomiques* (à plus de deux modalités de réponse souvent ordonnées comme "pas du tout", "un peu", "beaucoup", "tout à fait"). On trouve plus rarement des échelles visuelles analogiques (réglettes graduées ou non), discrètes ou continues.
Pour chaque item est définie une modalité ou *réponse négative* qui est associée à la modalité de réponse la plus défavorable. Les autres modalités sont dites *positives*. Ces appellations peu adéquates en santé sont dues à l'utilisation du vocabulaire des sciences de l'éducation et de la psychométrie.
Les réponses aux items, qui sont observables, permettent de mesurer une variable non directement mesurable et observable, le *trait latent* qui peut représenter des concepts tels que la qualité de vie, l'intelligence, l'état de santé, la douleur, la pharmacodépendance...
Une échelle de mesure est dite *générique* si elle mesure de façon globale un concept dans une population générale. C'est le cas d'échelles telles que le WHOQOL-BREF [107] ou le SF-36 [67]. Une échelle peut également être *spécifique* à une maladie comme l'EORTC-QLQ C30 [1], spécifique à l'évaluation de la qualité de vie de patients atteints de cancer.

Si les PRO sont par définition des mesures subjectives, ils n'en sont pas forcément moins fiables que des critères objectifs. En effet, le développement et l'analyse de ces mesures est basée sur des théories issues de la psychométrie (la théorie classique des tests et la théorie de réponse aux items). Les PRO sont des mesures valides, fiables, reproductibles et satisfaisant des propriétés psychométriques si l'instrument de mesure a été construit et validé correctement.

2.2 La théorie classique des tests

La théorie classique des tests s'est développée autour des tests d'intelligence et de personnalité en psychométrie et sciences de l'éducation. Si les premières pierres ont été posées par Thurstone[109], Spearman [102] et Brown [17], il faudra attendre l'ouvrage de Gulliksen [47] pour avoir une synthèse des recherches de la première moitié du vingtième siècle. Cet

ouvrage restera la référence jusqu'à la publication du livre de Lord and Novick [64].

Des concepts tels que la qualité de vie, l'intelligence sont représentés par une variable latente que l'on cherche à mesurer mais qui n'est pas directement observable. En revanche, le concept que l'on cherche à mesurer se manifeste par une multitude de caractères observables qui peuvent être transformés en items (modèle de l'univers d'items [30]). Si tous les items étaient observés, la somme des réponses aux items pourrait être considérée comme la vraie mesure du concept à mesurer. Or, avec un instrument de mesure, on n'évalue qu'un échantillon de l'ensemble des items. La *théorie classique des tests* (CTT) fait l'hypothèse qu'il existe un lien linéaire entre la mesure observée et la vraie mesure représentant la variable latente à mesurer.

Le *modèle général de la mesure* s'écrit :

$$Y_{ij} = T_{ij} + \epsilon_{ij} \tag{2.1}$$

où Y_{ij} est la réponse de l'individu i à l'item j, T_{ij} est la vraie mesure normalement distribuée de moyenne μ et de variance σ_T^2 et ϵ_{ij} est l'erreur de mesure aléatoire normalement distribuée de moyenne nulle et de variance σ_ϵ^2. La covariance entre T_{ij} et ϵ_{ij} est supposée nulle. L'erreur de mesure aléatoire représente l'imprécision introduite par le fait que seul un échantillon d'items est évalué par l'instrument de mesure.

Le modèle général de la mesure présente deux quantités inobservables. Pour que le modèle soit résolu, un ensemble d'hypothèses doit être défini. Les hypothèses posées déterminent le modèle de mesure dans lequel on se place, parmi lesquels le modèle parallèle et le modèle tau-équivalent.

Pour J items ($j = 1...J$) observés sur N individus ($i = 1...N$), le *modèle parallèle* [101] est un modèle linéaire à effets aléatoires et s'écrit :

$$Y_{ij} = T_{ij} + \epsilon_{ij} = \mu + a_i + \epsilon_{ij} \tag{2.2}$$

avec les cinq hypothèses suivantes :

1. μ est un effet fixe constant

2. a_i est un effet aléatoire de moyenne nulle et de variance σ_a^2. La vraie mesure est indépendante de l'item car $T_{ij} = \mu + a_i$.

3. ϵ_{ij} est un effet aléatoire de moyenne nulle et de variance σ_ϵ^2.

4. $cov(a_i, \epsilon_{ij}) = 0$. L'erreur de mesure ϵ_{ij} est indépendante de la vraie mesure T_{ij}.

5. $E(Y_{ij}) = \mu$, $var(Y_{ij}) = \sigma_a^2 + \sigma_\epsilon^2$, $cov(Y_{ij}, Y_{ij'}) = \sigma_a^2$

Ce modèle est restrictif et l'hypothèse d'indépendance de la mesure avec l'item peut être relâchée. Dans le *modèle tau-équivalent*, chaque item est toujours une répétition de la même mesure mais la moyenne de chaque item diffère. Ainsi, la vraie mesure diffère pour chaque item.

$$Y_{ij} = T_{ij} + \epsilon_{ij} = \mu_j + a_i + \epsilon_{ij} \tag{2.3}$$

Les hypothèses 1 et 5 sont alors modifiées :

1. μ_j est un effet fixe constant

2. a_i est un effet aléatoire de moyenne nulle et de variance σ_a^2. La vraie mesure diffère pour chaque item car $T_{ij} = \mu_j + a_i$.

3. ϵ_{ij} est un effet aléatoire de moyenne nulle et de variance σ_ϵ^2.

4. $cov(a_i, \epsilon_{ij}) = 0$. L'erreur de mesure ϵ_{ij} est indépendante de la vraie mesure T_{ij}.

5. $E(Y_{ij}) = \mu_j$, $var(Y_{ij}) = \sigma_a^2 + \sigma_\epsilon^2$

Ces modèles de mesure permettent de définir le cadre dans lequel est notamment développé la notion de fiabilité. La *fiabilité* réfère à la notion de reproductibilité des résultats obtenus, à la précision et la sensibilité de l'instrument de mesure. La fiabilité est définie comme la proportion de variance des réponses aux items imputable aux vraies mesures. Le coefficient de fiabilité d'un item est compris entre 0 et 1 et s'écrit :

$$\rho = \frac{V(T_{ij})}{V(Y_{ij})} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_\epsilon^2} \tag{2.4}$$

Sous le modèle tau-équivalent, la corrélation entre deux items est supposée constante et vaut la fiabilité, on a $corr(Y_{ij}, Y_{ij'}) = \rho \; \forall j \neq j'$. Il est aussi appelé coefficient intra-classe.

Soit le score S_i, somme de J réponses aux items. Sous le modèle tau-équivalent, la *fiabilité du score* S_i est donnée par la formule de Spearman-Brown :

$$\tilde{\rho} = \frac{J\rho}{1 + (J-1)\rho} \tag{2.5}$$

La fiabilité est habituellement estimée à l'aide du *coefficient* α *de Cronbach* [22], l'estimateur du maximum de vraisemblance de $\tilde{\rho}$ sous le modèle tau-equivalent.

$$\alpha = \frac{J}{J-1}\left(1 - \sum_{j=1}^{J} S_j^2/S_t^2\right) \tag{2.6}$$

avec $S_j^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_{ij} - \bar{y}_j)^2$ la variance de l'item j et
$S_t^2 = \frac{1}{NJ-1}\sum_{i=1}^{N}\sum_{j=1}^{J}(y_{ij} - \bar{y})^2$ la variance de la somme des items. α est borné entre 0 et 1 (items parfaitement corrélés).

Sous le modèle tau-équivalent, le coefficient alpha de Cronbach est un bon estimateur de la fiabilité [34, 83]. Si le modèle n'est ni parallèle ni tau-équivalent, le coefficient alpha est une borne inférieure de la fiabilité. En général, la valeur acceptable pour la fiabilité est $\alpha > 0.7$ [81]. La fiabilité augmente avec le nombre d'items.

2.3 La théorie de réponse aux items

La *théorie de réponse aux items* (IRT) est également appelée théorie moderne des tests car elle s'est développée plus récemment que la CTT. Comme son nom l'indique, cette théorie se concentre sur l'information au niveau de l'item en modélisant le lien entre la réponse à l'item et la variable latente à mesurer alors que la CTT se concentre sur l'information au niveau du test par la modélisation du lien entre le score total observé sur l'ensemble des items et la variable latente. En IRT, le lien entre les réponses aux items, qui sont des manifestations de la variable latente non observable, et la variable latente est modélisé. Ce lien n'est pas forcément linéaire, contrairement au modèle utilisé en CTT. Les fonctions de réponse aux items s'écrivent $P(Y_{ij} = y|\theta_i, \boldsymbol{v}_j)$ où Y_{ij} est la réponse de l'individu i à l'item j, θ_i la valeur du trait latent pour l'individu i et $\boldsymbol{v}_j$ le vecteur des paramètres caractérisant l'item j.
Dans les années 60, Rasch [90] propose un modèle à un seul paramètre. Toute une philosophie s'est ensuite développée autour de ce modèle et de ses propriétés psychométriques, avec

notamment les travaux d'Andersen [7], Wright [122], Andrich [9]. En parallèle, Lord et Novick [64] proposent le modèle à ogive normal et Birnbaum [12] introduit le modèle logistique, two-parameter logistic model ou 2-PLM. Ces modèles n'ayant pas les propriétés psychométriques du modèle de Rasch, la distinction sera ensuite faite entre les modèles de la famille de Rasch et les modèles de la famille de Lord. Les modèles de la famille de Rasch prônent la suprématie du modèle sur les données. On peut être amené à faire de la sélection d'items en supprimant des items pour améliorer l'adéquation des données au modèle. Dans les modèles de la famille de Lord, la suprématie des données amène à raisonner en terme d'adéquation du modèle aux données.

Tous ces modèles, adaptés à l'analyse d'items dichotomiques, seront étendus à l'analyse d'items polytomiques avec notamment le rating-scale model [9] et le partial-credit model [66] pour les modèles de la famille de Rasch et le graded-response model [93] pour les modèles de la famille de Lord.

2.3.1 Hypothèses fondamentales de L'IRT

La plupart des modèles de l'IRT reposent sur trois hypothèses fondamentales.

L'unidimensionnalité L'instrument ne mesure qu'un seul et même concept. Autrement dit, l'ensemble des items mesurent tous le même trait latent.

La monotonicité La probabilité d'une réponse positive (ou au moins celle-ci) à un item augmente avec la variable latente. Les fonctions de réponse aux items $P(Y_{ij} = y|\theta_i, \boldsymbol{\upsilon}_j)$ sont non-décroissantes en θ soit

$$\frac{dP(Y_{ij} = y|\theta_i, \boldsymbol{\upsilon}_j)}{d\theta_i} \geq 0 \tag{2.7}$$

L'indépendance locale Les réponses aux items sont indépendantes conditionnellement au trait latent. La réponse d'un individu à un item ne dépend pas de ce qu'il a répondu aux autres items. La probabilité d'un vecteur de réponse est égale au produit des J probabilités de réponse du sujet aux items.

$$P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, ..., Y_{iJ} = y_{iJ}|\theta_i) = \prod_{j=1}^{J} P(Y_{ij} = y_{ij}|\theta_i) \tag{2.8}$$

2.3.2 Le modèle de Rasch

Soit Y_{ij} la variable dichotomique représentant la réponse de l'individu i $(i = 1, ..., N)$ à l'item j $(j = 1, ..., J)$. Pour un questionnaire composé de J items dichotomiques, le modèle de Rasch s'écrit sous la forme d'un modèle logistique :

$$P(Y_{ij} = y|\theta_i; \delta_j) = \frac{\exp\left(y(\theta_i - \delta_j)\right)}{1 + \exp(\theta_i - \delta_j)} \tag{2.9}$$

où $y = 0$ pour une réponse négative et $y = 1$ pour une réponse positive. Le modèle suppose que la probabilité de réponse ne dépend que de deux paramètres. Le paramètre θ_i est la valeur individuelle du trait latent pour le patient i et représente la capacité de l'individu. Plus la capacité de l'individu est élevée, plus la probabilité que l'individu réponde positivement est grande. δ_j est le paramètre de difficulté associé à l'item j et représente le fait que les individus répondent plus ou moins facilement positivement aux items. Plus la difficulté de l'item est faible, plus la probabilité que l'individu réponde positivement est grande.

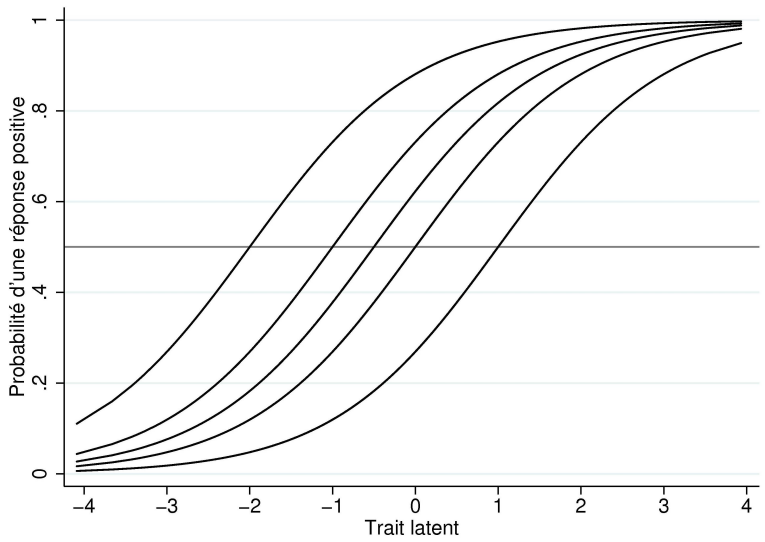

FIGURE 2.1 – Courbes caractéristiques de 5 items vérifiant un modèle de Rasch

La courbe représentative de la probabilité de réponse positive à un item en fonction du trait latent est appelé *courbe caractéristique de l'item* (ICC - Item Characteristic Curve). Le modèle logistique utilisé assure la croissance monotone des ICC comme on peut le voir sur la figure 2.1. Les pentes étant égales, les ICC sont non sécantes. Le point d'inflexion de la courbe a pour coordonnées (δ_j,0.5). Ainsi, lorsque $\theta = \delta_j$, on obtient 50% de réponses

positives à l'item j.

Dans le modèle de Rasch à effets fixes, les paramètres θ_i, $i = 1, ..., N$ et δ_j, $j = 1, ..., J$ sont des paramètres fixes à estimer. Le modèle de Rasch à effets fixes est un modèle linéaire généralisé dont la vraisemblance s'écrit :

$$L(\theta_1, ..., \theta_N, \delta_1, .., \delta_J|\mathbf{y}) = \prod_{i=1}^{N} \prod_{j=1}^{J} \frac{\exp\left(y_{ij}(\theta_i - \delta_j)\right)}{1 + \exp(\theta_i - \delta_j)} \tag{2.10}$$

Le modèle n'est pas identifiable. La contrainte d'identifiabilité la plus souvent adoptée est la nullité de la somme des paramètres de difficulté :

$$\sum_{j=1}^{J} \delta_j = 0 \tag{2.11}$$

La nullité d'un paramètre de difficulté en particulier peut aussi être adoptée comme contrainte d'identifiabilité.

Dans le modèle de Rasch à effets aléatoires, les paramètres θ_i, $i = 1, ..., N$ sont les réalisations d'une variable aléatoire Θ alors que les paramètres d'items (δ_j) sont fixes. La variable aléatoire Θ a pour distribution $g(\Theta/\boldsymbol{\nu})$ où $\boldsymbol{\nu}$ est le vecteur des paramètres de distribution. La loi normale est souvent choisie comme distribution de la variable aléatoire Θ pour des raisons pratiques. C'est notamment la loi utilisée dans les logiciels. Dans ce cas, les paramètres de distribution sont la moyenne et la variance. Le modèle de Rasch à effets aléatoires fait partie de la famille des modèles linéaires généralisés mixtes.

$$L(\delta_1, .., \delta_J, \boldsymbol{\nu}|\mathbf{y}) = \prod_{i=1}^{N} \int \prod_{j=1}^{J} \frac{\exp\left(y_{ij}(\theta - \delta_j)\right)}{1 + \exp(\theta - \delta_j)} g(\theta/\boldsymbol{\nu}) d\theta \tag{2.12}$$

Le modèle de Rasch à effets aléatoires n'est pas identifiable. La contrainte d'identifiabilité la plus courante est la nullité de la moyenne de la distribution du trait latent. Il est également possible de poser une contrainte sur les paramètres de difficulté comme dans le modèle de Rasch à effets fixes.

2.3.2.1 Propriétés du modèle de Rasch

Le modèle de Rasch appartient à la famille exponentielle. En posant

$$\begin{aligned} &\phi = 1 \\ &\theta = log(\frac{exp(\theta_i - \delta_j)}{1 + exp(\theta_i - \delta_j)}) \\ &b(\theta) = log(1 + exp(\theta_i - \delta_j)) \\ &c(z, \phi) = 0 \end{aligned} \tag{2.13}$$

on peut alors montrer que la densité du modèle de Rasch s'écrit sous la forme d'une densité de loi de la famille exponentielle : $f(z|\theta, \phi) = exp\left(\frac{z\theta - b(\theta)}{\phi} + c(z, \theta)\right)$.

Sous l'hypothèse d'indépendance locale, la densité conjointe des variables $Y_{i1}, ..., Y_{ij}, ..., Y_{iJ}$ s'écrit :

$$\begin{aligned} f(Y_{i1} = y_{i1}, ..., Y_{ij} = y_{ij}, ..., Y_{iJ} = y_{iJ}|\theta_i, \boldsymbol{\delta}) &= \prod_{j=1}^{J} \frac{exp\left[y_{ij}(\theta_i - \delta_j)\right]}{1 + exp(\theta_i - \delta_j)} \\ &= \frac{exp\left[s_i\theta_i - \sum_{j=1}^{J} y_{ij}\delta_j\right]}{\prod_{j=1}^{J} 1 + exp(\theta_i - \delta_j)} \end{aligned} \tag{2.14}$$

Par le théorème de Darmois [94], on montre que le score de l'individu i, $S_i = \sum_{j=1}^{J} Y_{ij}$, est une statistique exhaustive de θ_i. De le même façon, on peut montrer que le score de l'item j, $T_j = \sum_{i=1}^{N} Y_{ij}$, est une statistique exhaustive de δ_j. Pour le modèle de Rasch, on parle d'*exhaustivité du score S_i sur le trait latent* ce qui signifie que le score d'un individu contient toute l'information disponible sur l'individu. Le score observé est suffisant pour l'estimation des paramètres des individus. De plus, à chaque valeur du score correspond une et une seule estimation du trait latent. Tous les individus avec le même score auront la même estimation du trait latent, quelque soit le profil de réponse permettant d'obtenir le score.

Une autre propriété mise en avant par Rasch [89] est l'*objectivité spécifique*. Il y a indépendance de la mesure par rapport à l'instrument de mesure. Il est possible de comparer deux individus indépendamment de l'instrument de mesure, on parle d'invariance des comparaisons. Dans un modèle de Rasch, les estimations des capacités des individus ne dépendent pas de l'ensemble d'items utilisé. De même, les estimations des difficultés des items ne dépendent pas de l'échantillon. On parle d'invariance des paramètres.

La propriété d'objectivité spécifique est intéressante dans le cas de données manquantes. Etant donné que l'estimation du trait latent est indépendante de l'ensemble d'item utilisé pour la mesure alors le trait latent d'un individu n'ayant pas répondu à l'ensemble des items peut être estimé sans biais. Une application importante de la propriété d'objectivité spécifique a été le développement du testing adaptatif sur ordinateur (CAT - Computerized Adaptative Testing). Les individus ne répondent qu'à un sous-ensemble d'items du questionnaire adapté à leur capacité. A chaque réponse à un item de l'individu, sa capacité est réestimée et un nouvel item adapté le mieux possible à sa capacité lui est proposé. Le CAT a pour but d'estimer le plus précisément possible le trait latent des individus testés.
Il arrive que la difficulté d'un item ne soit pas la même d'un groupe d'individus à l'autre. Il y a alors violation de l'hypothèse d'invariance des paramètres d'items. On parle de fonctionnement différentiel de l'item (DIF - Differential Item Functioning). Il est important de détecter le DIF et de le prendre en compte dans l'analyse.

2.3.2.2 Estimation des paramètres

La méthode d'estimation des paramètres la plus couramment utilisée est celle du maximum de vraisemblance.

Maximum de Vraisemblance Jointe - JML

Dans le cadre du modèle de Rasch à effets fixes, la *méthode du maximum de vraisemblance jointe* (JML - Joint Maximum Likelihood) [73] consiste simplement à estimer conjointement les paramètres du modèle en maximisant la vraisemblance jointe.
Les estimateurs JML sont non consistants pour $N \to \infty$ à J fixé [6]. Les estimateurs sont consistants pour $N \to \infty$, $J \to \infty$, $N/J \to \infty$. En pratique, la méthode JML est à éviter, surtout lorsque N est petit comme c'est le cas dans le domaine de la recherche clinique par opposition aux sciences de l'éducation.

Maximum de Vraisemblance Conditionnelle - CML

Pour résoudre le problème de la non consistance des estimateurs JML dans le cadre du modèle de Rasch à effets fixes, Andersen [5] a proposé la *méthode du maximum de vraisemblance conditionnelle* (CML - Conditional Maximum Likelihood) . Cette méthode repose sur l'exhaustivité du score. Par une propriété de la famille exponentielle, la vraisemblance

conditionnée par le score S_i, qui est une statistique exhaustive de θ_i sur le trait latent, ne dépend plus des paramètres du trait latent.

Les estimateurs CML sont non biaisés, asymptotiquement efficaces, normalement distribués et consistants pour $N \to \infty$, J fixé [4]. La méthode CML nécessite la propriété d'exhaustivité du score, les paramètres des modèles IRT n'ayant pas cette propriété ne peuvent être estimés avec cette méthode. L'estimation des paramètres par la méthode CML n'est donc possible que pour les modèles de la famille de Rasch.

L'estimation par la méthode CML ne permet pas de gérer les données manquantes. Les individus dont le score total ne peut être calculé (ayant au moins une réponse manquante) sont exclus lors de l'estimation des paramètres. Pour surmonter ce problème, l'estimation Pairwise Conditional Maximum Likelihood (PCML) a été proposée [10, 125]. Elle consiste à maximiser la pseudo-vraisemblance de toutes les paires d'item sachant θ. La perte d'information est moindre qu'avec le CML car l'information portée par les individus ayant répondu à au moins deux items est utilisée avec l'estimation PCML. Les estimateurs PCML sont consistants pour $N \to \infty$, J fixé [125].

Maximum de Vraisemblance Marginale - MML

Les paramètres du modèle de Rasch à effets aléatoires à estimer sont les paramètres des items (δ_j) et les paramètres ν de la distribution du trait latent $g(\Theta)$. Avec la *méthode du maximum de vraisemblance marginale* (MML - Marginal Maximum Likelihood) [15], il s'agit de maximiser la fonction de vraisemblance marginale L_{MML} obtenue en intégrant la vraisemblance du modèle de Rasch par la fonction de distribution de Θ. La vraisemblance marginale s'écrit :

$$L_{MML}(\delta_1, .., \delta_J, \nu|\mathbf{y}) = \prod_{i=1}^{N} \int \prod_{j=1}^{J} \frac{\exp\left(y_{ij}(\theta - \delta_j)\right)}{1 + \exp(\theta - \delta_j)} g(\theta/\nu) d\theta \tag{2.15}$$

Le principal inconvénient d'utiliser les méthodes JML, CML et PCML pour l'estimation des paramètres est que les paramètres de difficulté et du trait latent ne tiennent pas compte des scores parfaits ($S_i = J$ ou $t_j = N$) ou nuls ($S_i = 0$ ou $t_j = 0$). Dans ces cas extrêmes, les estimations des paramètres tendraient vers plus ou moins l'infini. Les méthodes JML, CML

et PCML suppriment les items et les individus à score parfait ou nul ce qui n'est pas le cas de la méthode MML. De plus, les propriétés asymptotiques des estimateurs MML sont les mêmes que celles des estimateurs CML si la distribution du trait latent a été correctement spécifiée [86].

2.3.2.3 Estimation des valeurs individuelles du trait latent

Les méthodes CML, PCML et MML ne permettent pas d'obtenir directement les estimations individuelles du trait latent. Ces estimations peuvent être obtenues conditionnellement aux paramètres de difficulté obtenus par estimation.

Estimateur du maximum de vraisemblance - MLE

Lorsque les paramètres du trait latent sont considérés comme fixes, les *estimateurs du maximum de vraisemblance* (Maximum Likelihood Estimator - MLE) sont obtenus en maximisant la vraisemblance dans laquelle ont été "injectés" les estimations des paramètres de difficulté ($\hat{\delta}_j$) obtenues par CML. La moyenne des estimations MLE est une estimation non biaisée de la moyenne de la population. En revanche, la variance des MLE sous-estime la variance de la population [71]. De plus, il n'est pas possible d'obtenir d'estimation pour les scores nul ou parfaits.

Estimateur de la vraisemblance pondéré - WLE

Lorsque les paramètres du trait latent sont considérés comme fixes, l'*estimateur de la vraisemblance pondéré* (Weighted Likelihood Estimator - WLE) est obtenu en maximisant la densité postérieure de θ conditionnellement à $\boldsymbol{y}$ et $\boldsymbol{\delta}$.

$$P(\theta|\boldsymbol{y},\boldsymbol{\delta}) = \frac{B(\theta)exp(\theta s)g(\theta)}{\int B(\theta)exp(\theta s)g(\theta)d\theta} \propto B(\theta)exp(\theta s)g(\theta) \tag{2.16}$$

avec s le score observé pour l'individu i, $B(\theta) = \prod_j [1 + \exp(\theta - \delta_j)]^{-1}$.
L'estimateur WLE est habituellement obtenu en posant

$$g(\theta) = \sqrt{I(\theta)} \tag{2.17}$$

avec $I(\theta)$ l'information de Fisher. L'estimateur WLE est moins biaisé que l'estimateur MLE tout en ayant les mêmes propriétés asymptotiques [121]. De plus, cet estimateur permet

d'obtenir des estimations pour les scores parfaits et nuls.

Estimateur attendu a posteriori de Bayes - EAP

Lorsque le trait latent est considéré comme une variable aléatoire, l'*estimateur attendu a posteriori de Bayes* (Expected A Posteriori Estimator - EAP) [15] est défini comme la moyenne de la distribution postérieure :

$$EAP(\theta_i) = E(\theta|\boldsymbol{y}, \boldsymbol{\delta}) = \frac{\int \theta B(\theta) \exp(\theta s) g(\theta) d\theta}{\int B(\theta) \exp(\theta s) g(\theta) d\theta} \tag{2.18}$$

La moyenne des estimations EAP est une estimation non biaisée de la moyenne de Θ. En revanche, la variance des EAP sous-estime la variance de Θ [71].

2.3.3 Modèles pour items polytomiques

Le modèle de Rasch a rapidement été étendu aux items polytomiques car ils souvent employés dans les questionnaires. C'est notamment le cas dans l'évaluation des Patient-Reported Outcomes pour laquelle sont souvent employés des items à modalités de réponse ordinales telles que "pas du tout", "un peu", "beaucoup", "tout à fait". La réponse la moins favorable est toujours dite négative et prend pour valeur 0. Les autres réponses sont dites positives, sont ordonnées et prennent pour valeur $1, ..., m_j$.
Parmi les nombreux modèles développés pour des items polytomiques, le Rating-Scale model et le Partial Credit Model sont très utilisés car ils appartiennent aux modèles de la famille de Rasch. Autrement dit, ces modèles présentent les mêmes propriétés que le modèle de Rasch : l'exhaustivité du score et l'objectivité spécifique. La fonction de réponse aux items exprime désormais la probabilité de réponse à une catégorie (modalité) donnée en fonction du niveau du trait latent et d'un ou plusieurs paramètres d'item.

Rating Scale Model - RSM

Le Rating Scale Model proposé par Andrich [9] et basé sur les travaux d'Andersen [7] fait l'hypothèse que tous les items ont le même nombre de modalités. On a alors $\forall j, m_j = m$.

Chaque item est caractérisé par un paramètre de difficulté globale de l'item δ_j qui correspond à la valeur du trait latent pour laquelle la probabilité de répondre négativement à

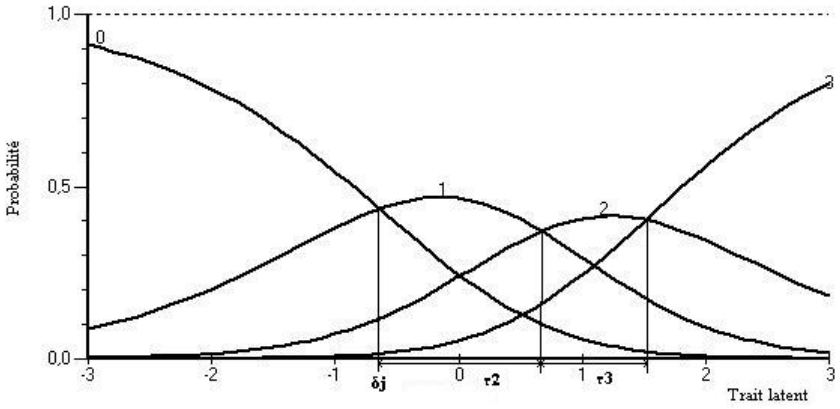

FIGURE 2.2 – Courbes caractéristiques d'un item à 4 modalités de réponse suivant un RSM

l'item j est égale à la probabilité de répondre positivement à la première modalité positive de l'item j (codée 1).

Un paramètre d'ajustement τ_h est également défini pour chaque modalité positive h avec $h > 1$. Ce paramètre correspond à la difficulté de la réponse h pour les individus ayant une réponse au moins égalé à $h-1$. La somme $\delta_j + \sum_{p=2}^{h} \tau_p$ correspond à la valeur du trait latent pour laquelle la probabilité de répondre positivement à la modalité $h-1$ d'un item est égale à la probabilité de répondre positivement à la modalité h d'un item. Les paramètres $(\tau_h)_{h=2,...,m}$ sont les mêmes pour tous les items.
La figure 2.2 présente les courbes caractéristiques d'un item à 4 modalités de réponse suivant un RSM. Trois paramètres d'item sont estimés : le paramètre de difficulté globale de l'item δ_j et deux paramètres d'ajustement (τ_2) et (τ_3).

Pour le Rating Scale Model, la probabilité de réponse de l'individu i à la modalité positive h de l'item j s'écrit :

$$P(Y_{ij} = h|\theta_i, \delta_j, \tau_2, ..., \tau_m) = \frac{exp(h(\theta_i - \delta_j) - \sum_{p=2}^{h} \tau_p)}{\sum_{l=0}^{m} exp(l(\theta_i - \delta_j) - \sum_{p=2}^{l} \tau_p)} \tag{2.19}$$

Les paramètres peuvent être estimés par les méthodes du maximum de vraisemblance conditionnelle (CML) ou du maximum de vraisemblance marginale (MML).

Partial Credit Model - PCM

Dans le Partial Credit Model (PCM), développé par Masters [66], le nombre de modalités d'un item à l'autre peut être différent. Chaque modalité positive h de chaque item j est caractérisée par un paramètre δ_{jh}.

Pour le Partial Credit Model, la probabilité de réponse de l'individu i à la modalité positive h de l'item j s'écrit :

$$P(Y_{ij} = h|\theta_i, \delta_{j1}, \delta_{j2}, ..., \delta_{jm_j}) = \frac{exp(h\theta_i - \sum_{p=1}^{h} \delta_{jp})}{\sum_{l=0}^{m_j} exp(l\theta_i - \sum_{p=1}^{l} \delta_{jp})} \tag{2.20}$$

Le Rating Scale Model peut être considéré comme un modèle particulier du Partial Credit Model pour lequel $\delta_{j1} = \delta_j$ et $\delta_{jk} = \delta_j + \tau_k, \forall k > 1$. Les paramètres peuvent être estimés par les méthodes du maximum de vraisemblance jointe (JML), du maximum de vraisemblance conditionnelle (CML) ou du maximum de vraisemblance marginale (MML).

2.3.4 Modèles IRT longitudinaux

Le modèle de Rasch peut être facilement étendu à l'analyse de données longitudinales. De manière générale, la probabilité de réponse d'un individu i à un item j au temps t, $t = 1, ..., T$ s'écrit [68] :

$$P(Y_{ij}^{(t)} = y^{(t)}|\theta_i^{(t)}, \delta_j^{(t)}) = \frac{exp(y^{(t)}(\theta_i^{(t)} - \delta_j^{(t)}))}{1 + exp(\theta_i^{(t)} - \delta_j^{(t)})} \tag{2.21}$$

De multiples extensions du modèle de Rasch et des modèles IRT ont été proposées comme le linear logistic test model (LLTM) et le linear logistic test model with relaxed assumptions (LLRA) de Fischer, le multidimensional Rasch model for learning and change (MRMLC) de Embretson et le modèle de Rasch multidimensionnel d'Andersen.

Dans le linear logistic test model (LLTM) [38], le changement est supposé le même pour tous les individus.

$$P(Y_{ij}^{(t)} = 1|\theta_i, \delta_j, \lambda_t) = \frac{exp(\theta_i + \lambda_t - \delta_j)}{1 + exp(\theta_i + \lambda_t - \delta_j)} \tag{2.22}$$

Les paramètres θ_i et δ_j sont constants dans le temps. λ_t peut être interprété comme le changement moyen du trait latent dans le temps. Le modèle LLTM est un modèle unidimensionnel. Il a été étendu au cas multidimensionnel avec le linear logistic test model with

relaxed assumptions (LLRA) [39].

Andersen [8] a proposé un modèle de Rasch multidimensionnel pour lequel le changement n'est pas obligatoirement le même pour tous les individus. Les paramètres de difficulté sont constants dans le temps.

$$P(Y_{ij}^{t} = 1|\theta_i^{(t)}, \delta_j) = \frac{exp(\theta_i^{(t)} - \delta_j)}{1 + exp(\theta_i^{(t)} - \delta_j)} \tag{2.23}$$

Le multidimensional Rasch model for learning and change (MRMLC) proposé par Embretson est également un modèle multidimensionnel. Il a été développé pour des situations où ce ne sont pas obligatoirement les mêmes items qui sont proposés à chaque temps de mesure.

$$P(Y_{ij}^{(t)} = y^{(t)}|\boldsymbol{\theta}_i, \delta_j) = \frac{exp\left(\sum_{t=1}^{T} \theta_i^{(t)} - \delta_j\right)}{1 + exp\left(\sum_{t=1}^{T} \theta_i^{(t)} - \delta_j\right)} \tag{2.24}$$

avec $\boldsymbol{\theta}_i$ le vecteur contenant $\theta_i^{(1)}$ la valeur du trait latent au premier temps et $\theta_i^{(2)}, ..., \theta_i^{(T)}$ les changements du trait latent d'un temps à l'autre.

Une extension longitudinale du RSM peut être trouvée dans Fischer et Parzer [36]. De même, une extension longitudinale du PCM a été proposée par Fischer et Ponocny [37].

2.4 Comparaison des deux théories

2.4.1 Les types de mesure

Comme le rapporte Stevens [106], Campbell définit la mesure comme l'attribution de nombres à des objets ou à des événements selon certaines règles. Stevens distingue 4 types de mesures [105], en définissant différentes règles pour l'attribution des nombres, les propriétés mathématiques des échelles ainsi obtenues et les opérations statistiques sur les mesures obtenues avec chaque échelle.

La *mesure nominale* est celle qui est le moins contrainte. Un nombre est attribué pour définir une caractéristique de l'individu. Ce nombre sert uniquement de libellé, une lettre ou un autre symbole peut être utilisé. Seules des statistiques telles que le calcul des fréquences

et le mode sont adéquates pour les mesures nominales. Les mesures nominales répondent aux trois postulats fondamentaux suivants :

- Postulat 1 : identité ou différence. Soit a=b, soit $a \neq b$.
- Postulat 2 : symétrie de la relation d'égalité. Si a=b alors b=a.
- Postulat 3 : transitivité de l'égalité. Si a=b et b=c alors a=c.

La mesure est *ordinale* lorsque les modalités de réponses sont ordonnées (ex : 1. très difficile 2. difficile 3. facile 4. très facile). Pour ces mesures, la médiane et les percentiles sont utilisés. L'interprétation de la moyenne et l'écart type est inappropriée pour les mesures ordinales car l'écart entre les modalités adjacentes n'ont pas forcément la même longueur. Outre les trois postulats précédents, les mesures ordinales répondent aux deux postulats suivants :

- Postulat 4 : relation d'ordre antisymétrique. Si a<b alors b>a.
- Postulat 5 : transitivité de la relation d'ordre. Si a<b et b<c alors a<c.

Pour les *mesures d'intervalles*, les écarts entre les modalités adjacentes sont de la même longueur. Un zéro existe mais il est fixé arbitrairement. Les deux échelles de température en Celsius et en Farenheit sont deux exemples d'échelles de mesure d'intervalle. L'interprétation de la moyenne et de l'écart type est possible. Outre les cinq postulats précédents, les mesures d'intervalles répondent aux quatres postulats suivants :

- Postulat 6 : possibilité de sommation. Si a=p et b>0 alors a+b>p.
- Postulat 7 : commutativité de l'addition. a+b=b+a.
- Postulat 8 : possibilité de substitution de termes égaux. Si a=p et b=q alors a+b=p+q.
- Postulat 9 : associativité de l'addition. (a+b)+c=a+(b+c).

La *mesure de rapports* (ou proportionnelle) connaît l'ensemble des relations d'identité, d'ordre, d'égalité d'intervalles et de proportionnalité. Une échelle de mesure de rapports a un "vrai" zéro. C'est le cas, par exemple, pour les mesures physiques telles la masse ou la longueur. Les mesures de rapports répondent à l'ensemble des 9 postulats précédents et au dixième postulat suivant :

- Postulat 10 : présence d'un zéro vrai. Si a/b=p/q alors qa=bp.

2.4.2 Types de mesure en CTT et en IRT

Fischer [35] a montré que les échelles de mesure des items et des individus sont des échelles de mesure d'intervalles pour les modèles de la famille de Rasch. La justification de la propriété de mesures d'intervalle du modèle de Rasch s'appuie notamment sur la propriété d'invariance des comparaisons comme le souligne Embretson [28].

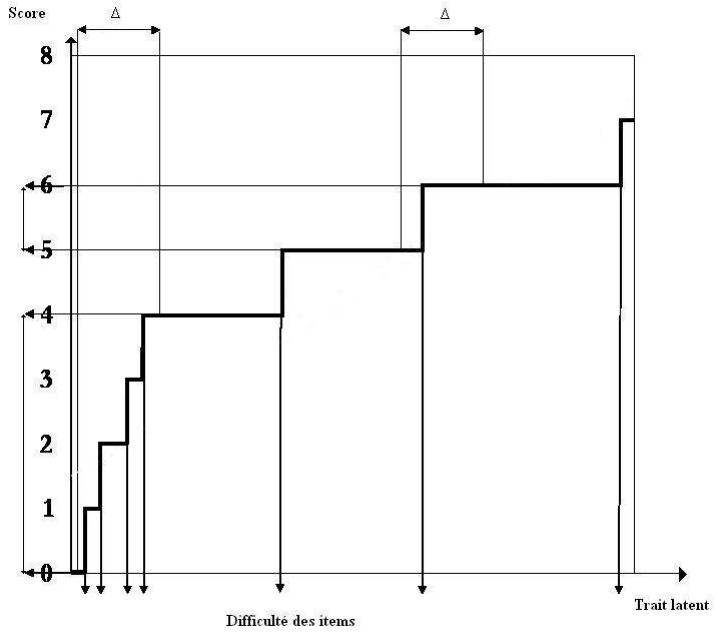

FIGURE 2.3 – Relation score et trait latent

Le graphique 2.3 présente la relation entre score et trait latent. Cette relation est ici présentée de manière déterministe pour montrer qu'une même différence de trait latent(Δ) peut correspondre à des différences de score de taille variable.

La propriété de mesure d'intervalles n'est pas systématiquement obtenue en CTT. Sur le graphique 2.4, on observe que, lorsque les difficultés d'items sont régulièrement espacées, des patients ayant une même différence de trait latent auront approximativement une même différence de score. La propriété de mesure d'intervalles de l'échelle peut être obtenue si les scores sont normalement distribués dans l'échantillon [29]. De plus, seules des transformations linéaires de ces scores garantissent le maintien de la propriété de mesure d'intervalles. Néanmoins, en CTT, la propriété de mesure d'intervalles est dépendante de l'échantillon. Si la propriété de mesure d'intervalles n'est pas obtenue, le score total est une mesure ordinale. Deux individus peuvent alors être comparés en terme de rang (l'un a un score plus élevé que

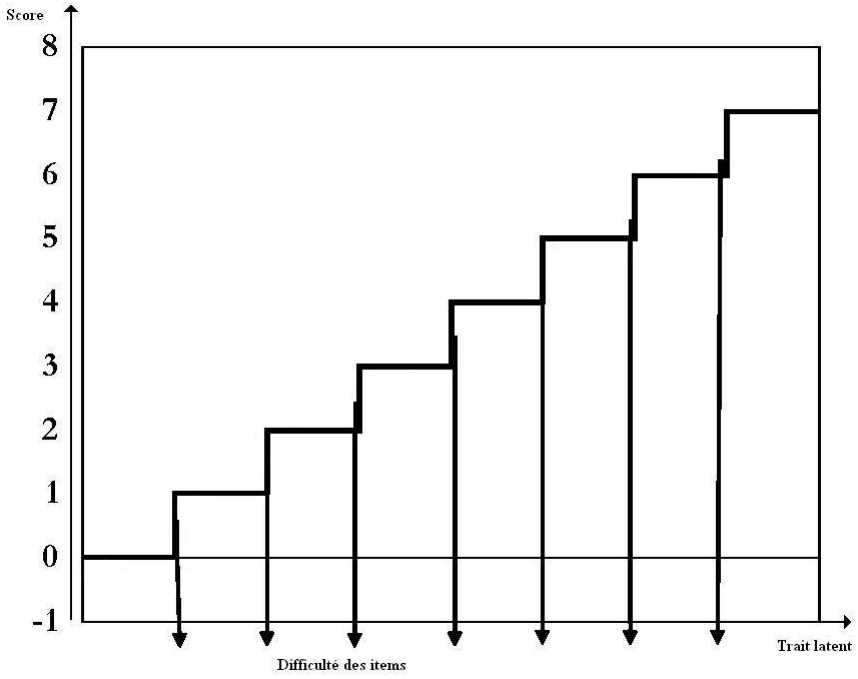

FIGURE 2.4 – Relation score et trait latent : difficultés d'items régulièrement espacées

l'autre) mais la distance entre les deux scores n'est interprétable que lorsque le niveau est connu.

2.4.3 CTT versus IRT

Le principal avantage de la théorie classique des tests est sa simplicité. La plupart des auteurs rapporte que la CTT repose sur de "faibles hypothèses théoriques" permettant de l'appliquer dans beaucoup de situations [48]. La CTT connaît plusieurs limites dont la principale est la dépendance des paramètres à l'échantillon. Les paramètres des individus (mesure observée) sont dépendants de l'ensemble d'items utilisé. De même, les paramètres d'items sont dépendants de l'échantillon d'individus. La dépendance des paramètres rend difficile les comparaisons d'individus qui n'ont pas répondu au même ensemble d'items ou les comparaisons d'items dont les paramètres ont été obtenus sur des échantillons d'individus différents. De plus, la notion de fiabilité en CTT repose sur les modèles parallèle et tau-équivalent. Ces deux modèles font l'hypothèse d'homogénéité de la variance de l'erreur de mesure pour tous les individus. Cette hypothèse est rarement rencontrée en pratique. Par exemple, lorsque les items sont dichotomiques, l'homogénéité des variances n'est obtenue que si la répartition entre réponses positives et réponses négatives est la même pour tous les items. Sans homogé-

néité de la variance, les coefficients calculés sont alors une estimation de la borne inférieure de la fiabilité ou une estimation de la fiabilité avec des biais inconnus [49].

Le développement de la théorie de réponse aux items visait à corriger les faiblesses de la CTT. Embretson [28] compare la CTT et l'IRT et décrit les nouvelles règles de la mesure, changements induits par le développement de l'IRT, dont les principales sont : l'hypothèse d'homogénéité des variances n'est plus nécessaire, des questionnaires plus courts peuvent être plus fiables que des questionnaires plus longs et l'invariance des paramètres. La propriété d'invariance des paramètres est très importante en IRT. Elle rend les comparaisons plus facilement interprétables et a permis le développement du computerized adaptative testing (CAT). L'intérêt du CAT est de réduire la taille du questionnaire en assurant une précision au moins aussi bonne que si tous les items avaient été répondus. La réduction de la taille du questionnaire est intéressante en santé car elle permet de réduire le temps de passation pour des populations potentiellement fatiguées et qui risquent de ne pas répondre à l'integralité du questionnaire si celui-ci est trop long. La propriété d'invariance des paramètres assure aussi une bonne gestion des données manquantes puisqu'il n'est pas nécessaire que l'individu ait répondu à l'ensemble des items pour pouvoir évaluer sa capacité. Au contraire, en CTT, les données manquantes peuvent empêcher le calcul du score de l'individu entraînant une perte d'information. En contrepartie, les modèles de la théorie de réponse aux items sont plus complexes. Ils reposent sur des hypothèses souvent considérées comme plus fortes. Néanmoins, nous avons vu que les approches CTT et IRT reposent sur des hypothèses très différentes. Il semble difficile de les comparer et de considérer que certaines sont plus fortes que d'autres.

Deuxième partie

Comparaison de méthodes d'analyse de données subjectives longitudinales

L'évaluation de Patient-Reported Outcomes prend une place de plus en plus importante en recherche clinique et épidémiologique. Comme nous l'avons vu au chapitre précédent, deux théories existent pour l'analyse des Patient-Reported Outcomes : la théorie classique des tests (CTT) et la théorie de réponse aux items (IRT). Ces deux approches sont très employées pour le développement et la validation de questionnaires. En revanche, l'approche classique reste la théorie la plus employée pour l'analyse des PRO. Toutefois, lorsque le questionnaire utilisé pour l'étude a été validé à la fois en CTT et en IRT, on peut se demander si la méthode la plus adéquate pour l'analyse est basée sur la CTT ou l'IRT. Dans la pratique, l'usage de la CTT semble plus répandu pour analyser les PRO. Or, il est dommage de se priver des propriétés spécifiques des modèles issus de l'IRT, en particulier de ceux issus de la famille de Rasch.

De plus, les Patient-Reported Outcomes sont souvent évalués dans les maladies chroniques ou le suivi après un traitement ou une chirurgie. Les données sont alors recueillies de manière longitudinale au sein d'un ou plusieurs groupes de patients. Leur intérêt est alors de pouvoir évaluer l'évolution du critère d'intérêt au cours du temps et de comparer cette évolution entre les groupes. Les méthodes pour l'analyse doivent donc tenir compte du caractère longitudinal des données et donc de la corrélation entre les réponses d'un même individu à chaque temps d'étude.

En raison de l'état de santé des patients, les données recueillies sont souvent sujettes au dropout. Ces sorties d'étude prématurées sont souvent considérées MNAR car elles ont un fort risque d'être liées au niveau du critère étudié (la qualité de vie par exemple). En effet, on peut émettre l'hypothèse qu'un patient dont l'état se dégrade et dont le niveau du critère étudié diminue a plus de risque de sortir de l'étude prématurément que les autres patients. Il est alors probable que l'échantillon souffre d'un biais de sélection et que la qualité de l'inférence en soit affectée. La présence de données manquantes peut avoir un impact sur les résultats des analyses. Comme nous l'avons vu, il existe plusieurs mécanismes de données manquantes et l'informativité ou non des données manquantes peut invalider une méthode d'analyse. L'impact des données manquantes sur l'analyse de PRO peut dépendre de la quantité et du type de données manquantes. Il pourrait être différent d'une approche à l'autre. En effet, la propriété d'objectivité spécifique du modèle de Rasch pourrait permettre à ce modèle d'être plus robuste face aux données manquantes qu'un modèle basé sur la CTT.

Ce travail a pour but d'identifier la stratégie la plus adéquate pour l'analyse de Patient-Reported Outcomes recueillis de manière longitudinale au moyen d'un questionnaire validé avec un modèle de Rasch. Plusieurs études de simulation sont présentées ci-après. Elles ont pour but de comparer différentes méthodes d'analyse adaptées aux données longitudinales et basées sur la CTT ou l'IRT. L'impact de la quantité et du type de données manquantes sur les différentes méthodes est également étudié. Le chapitre suivant présente les quatre méthodes d'analyse comparées, basées sur la CTT ou sur le modèle de Rasch, ainsi que la méthode de simulation des données. Trois études de simulation sont ensuite présentées. La première étude compare quatre méthodes d'analyse (trois méthodes basées sur l'IRT et une méthode basée sur la CTT) dans le cadre de données complètes et un groupe de patients pour permettre de valider ces méthodes dans un cadre simple et favorable mais peu réaliste. Plusieurs méthodes d'analyse basées sur l'IRT sont comparées car plusieurs possibilités sont envisageables : l'estimation individuelle du trait latent puis l'estimation de l'effet temps ou l'estimation simultanée de la distribution de la variable latente et de l'effet temps. Dans la deuxième étude, l'impact du dropout de type MCAR ou MNAR est étudié sur une méthode basée sur la CTT et une méthode basée sur l'IRT pour des données issues d'un seul groupe de patients. La troisième étude évalue l'impact du dropout de type MCAR ou MNAR dans l'estimation des effets temps et groupe sur une méthode basée sur la CTT et une méthode basée sur l'IRT, pour des données portant sur deux groupes de patients.

Chapitre 3

Base commune des études de simulation

L'intérêt de mettre en place une étude de simulation réside dans la possibilité de créer des données dont on connaît et maîtrise la "vérité", comme les vraies valeurs des paramètres et les hypothèses sous-jacentes. Une étude de simulation repose non sur des données recueillies, pour lesquelles les vraies valeurs des paramètres ne peuvent pas être connues, mais sur des paramètres dont on a fixé les valeurs. A partir de ces valeurs, des jeux de données sont simulés selon un modèle choisi. Ainsi, les vraies valeurs des paramètres et les hypothèses du modèle sont connues et contrôlées ce qui n'est jamais le cas lors de l'analyse de données recueillies. Les données simulées sont ensuite analysées et permettent d'estimer les valeurs des paramètres et de tester différentes hypothèses. En faisant varier les valeurs des paramètres et les hypothèses testées, il est possible d'étudier différents scénarii. Ces scenarii peuvent reproduire la réalité ou tout au moins s'en approcher pour étudier des propriétés dans des cas courants et ainsi valider un modèle ou au contraire s'éloigner de la réalité afin d'étudier des propriétés asymptotiques ou la robustesse d'un modèle. Dans ces situations variées, la validité, la performance ou la robustesse d'une méthode d'analyse peut être évaluée.

A travers les études de simulation mises en place dans ce travail, il nous a été possible d'étudier la capacité d'une méthode d'analyse à rejeter à tort ou à raison une hypothèse et donc d'évaluer le risque de première espèce et la puissance. Il a également été possible de mesurer l'écart entre les valeurs réelles (fixées a priori) et les valeurs estimées lors de l'analyse afin de connaître la qualité des estimations obtenues et de quantifier les biais associés.

3.1 Méthodes comparées

Pour l'ensemble des études de simulation, les N individus sont évalués en $T = 3$ temps au moyen d'un questionnaire composé de J items dichotomiques. La réponse du patient i $(i = 1, ..., N)$ à l'item j $(j = 1, ..., J)$ au temps t $(t = 1, ..., T)$ est notée $Y_{ij}^{(t)}$. Lorsque les patients sont divisés en deux groupes, g $(g = 0, 1)$ indique le groupe d'appartenance du patient.

3.1.1 Score and Mixed Models - SM

La méthode "Score and Mixed Models" (SM) est la seule méthode comparée basée sur la théorie classique des tests (cf. section 2.2). Elle consiste à calculer le score de l'individu à chaque temps puis à estimer l'évolution des scores de l'individu par un modèle linéaire mixte (cf. section 1.1).

$$
\begin{aligned}
& S_i^{(t)} = \sum_{j=1}^{J} Y_{ij}^{(t)} \\
& \boldsymbol{S}_i = (S_i^{(1)}, ..., S_i^{(3)})' = \boldsymbol{X}_i \boldsymbol{\beta}_S + \boldsymbol{e}_{S,i} \\
& \boldsymbol{X}_i = \begin{pmatrix} 1 & 0 & 0 & g_i \\ 0 & 1 & 0 & g_i \\ 0 & 0 & 1 & g_i \end{pmatrix}, \boldsymbol{\beta}_S = \begin{pmatrix} \beta_{S,1} \\ \beta_{S,2} \\ \beta_{S,3} \\ \beta_{S,gp} \end{pmatrix} \\
& \boldsymbol{e}_{S,i} \sim N(0, \boldsymbol{\Sigma}_{S,i}) \\
& \boldsymbol{S}_i \sim N_3(\boldsymbol{\mu}_{S,i}, \boldsymbol{\Sigma}_{S,i})
\end{aligned}
\tag{3.1}
$$

avec $\boldsymbol{\mu}_{S,i} = (\mu_{S,i}^{(1)} \cdots \mu_{S,i}^{(3)})' = \boldsymbol{X}_i \boldsymbol{\beta}_S$. Les paramètres de moyenne $\boldsymbol{\beta}_S$ et les composants de la variance $\boldsymbol{\omega}$ du modèle sont estimés par la méthode REML. Les patients sont tous évalués aux mêmes temps de mesure. La corrélation entre les mesures est modélisée à travers la matrice de variance covariance.

3.1.2 Rasch and Mixed Models - RM

La méthode "Rasch and Mixed Models" (RM) est composée de deux étapes. Dans un premier temps, les paramètres du modèle de Rasch à effets aléatoires sont estimés sur tout l'échantillon par la méthode MML. La variable aléatoire Θ est supposée suivre une loi normale. Les

valeurs individuelles du trait latent sont estimées en utilisant les EAP du modèle de Rasch (cf. section 2.3.2.3). La contrainte d'identifiabilité pour le modèle de Rasch est $\mu = 0$.

$$P(Y_{ij} = y|\theta_i, \delta_j) = \frac{exp(y(\theta_i - \delta_j))}{1 + exp(\theta_i - \delta_j)} \tag{3.2}$$

Dans un second temps, l'évolution du trait latent au cours du temps est estimée dans un modèle linéaire mixte.

$$\begin{aligned}
&\hat{\boldsymbol{\theta}}_i = (\hat{\theta}_i^{(1)}, ..., \hat{\theta}_i^{(3)})' = \boldsymbol{X}_i\boldsymbol{\beta}_\theta + \boldsymbol{e}_{\theta,i} \\
&\boldsymbol{X}_i = \begin{pmatrix} 1 & 0 & 0 & g_i \\ 0 & 1 & 0 & g_i \\ 0 & 0 & 1 & g_i \end{pmatrix}, \boldsymbol{\beta}_\theta = \begin{pmatrix} \beta_{\theta,1} \\ \beta_{\theta,2} \\ \beta_{\theta,3} \\ \beta_{\theta,gp} \end{pmatrix} \\
&\boldsymbol{e}_{\theta,i} \sim N(0, \boldsymbol{\Sigma}_{\theta,i}) \\
&\hat{\boldsymbol{\theta}}_i \sim N_3(\boldsymbol{\mu}_{\theta,i}, \boldsymbol{\Sigma}_{\theta,i})
\end{aligned} \tag{3.3}$$

avec $\boldsymbol{\mu}_{\theta,i} = (\mu_{\theta,i}^{(1)} \cdots \mu_{\theta,i}^{(3)})' = \boldsymbol{X}_i\boldsymbol{\beta}_\theta$ et $\hat{\theta}_i^{(t)}$ l'estimation individuelle du trait latent obtenue par EAP pour l'individu i au temps t.

3.1.3 Plausible Values model - PV

La méthode "Plausible Values model" (PV) est basée sur le même principe que la méthode RM. Les paramètres du modèle de Rasch sont estimés par la méthode MML sur tout l'échantillon dans un premier temps puis des estimations individuelles du trait latent à chaque temps d'observation sont obtenues dans un second temps. Dans la méthode PV, les estimations individuelles du trait latent sont obtenues par l'imputation de valeurs plausibles.

L'imputation de *valeurs plausibles* est basée sur la théorie de l'imputation multiple de Rubin [92]. Dans la méthode d'imputation de valeurs plausibles, la variable latente est considérée comme manquante pour tous les individus. Des valeurs plausibles de θ sont tirées aléatoirement dans la distribution postérieure de la variable latente. Contrairement à l'estimation par EAP, les individus avec le même profil de réponse ou simplement le même score total et donc la même distribution postérieure peuvent avoir des estimations individuelles du trait latent différentes avec la méthode des valeurs plausibles. Pour chaque individu, un ou plusieurs ti-

rages sont réalisés pour chaque valeur ce qui permet de corriger le biais de la variance observé pour les estimateurs MLE, WLE et EAP. De la même façon que pour l'imputation multiple, les analyses postérieures sur les estimations individuelles du trait latent sont réalisées sur chaque tirage afin d'estimer les paramètres d'intérêt. Les résultats des analyses postérieures sont ensuite combinés pour obtenir une seule estimation des paramètres d'intérêt [78]. Cette méthode est largement employée dans les études internationales sur les performances scolaires comme PISA (Programme International sur le Suivi des Acquis des élèves) [111] dont l'intérêt porte sur le niveau moyen des élèves et donc sur les paramètres de population de la variable latente.

Les valeurs de la variable latente ne sont pas observables et doivent être estimées. L'incertitude liée à l'estimation n'est pas prise en compte lorsque des estimateurs tels que l'EAP sont utilisés dans la méthode RM car les estimations individuelles de la variable latente sont traitées comme si elles avaient été observées. Avec les valeurs plausibles, plusieurs tirages sont effectués en général afin d'obtenir une estimation de l'incertitude liée à l'estimation du trait latent. Glas et al. [43] considèrent qu'un seul tirage est suffisant si l'incertitude liée à l'estimation doit être prise en compte dans l'analyse mais qu'il n'est pas nécessaire d'obtenir une estimation explicite de cette incertitude. De plus, Wu [123] a montré qu'un seul tirage peut être suffisant pour estimer correctement les paramètres de population. Etant donné que l'intérêt de l'analyse de PRO dans un cadre longitudinal se porte sur l'évolution du critère d'intérêt, les différentes études de simulation mises en place se concentrent sur les paramètres de population et non les trajectoires individuelles. Pour cette raison, la méthode PV n'utilisera, dans ce travail, qu'un seul tirage de valeurs plausibles.

Les valeurs plausibles sont tirées aléatoirement dans la distribution postérieure de la variable latente. La variable aléatoire Θ est supposée suivre une loi normale. Les valeurs plausibles $\hat{\theta}_i^{(t)}$ de chaque individu i à chaque temps t sont tirées dans une loi normale dont la moyenne vaut l'estimation EAP de l'individu i et l'écart-type vaut l'erreur standard estimée correspondante. Le modèle utilisé est le même que pour la méthode RM dans lequel les estimations individuelles du trait latent $\hat{\theta}_i^{(t)}$ sont les estimations obtenues par imputation de valeurs plausibles.

3.1.4 Longitudinal Rasch Mixed model - LRM

La méthode "Longitudinal Rasch Mixed model" (LRM) est basée sur le modèle de Rasch multidimensionnel d'Andersen [8]. Elle permet d'estimer en une seule étape les paramètres de moyenne et les paramètres de variance covariance de la variable latente.

$$\begin{aligned} P(Y_{ij}^{(t)} = y^{(t)}|\theta_i^{(t)}, \delta_j) &= \frac{exp(y^{(t)}(\theta_i^{(t)} - \delta_j + \beta_{\theta,gp} * g_i))}{1 + exp(\theta_i^{(t)} - \delta_j + \beta_{\theta,gp} * g_i)} \\ \boldsymbol{\theta_i} = (\theta_i^{(1)}, ..., \theta_i^{(3)})' &\sim N_3\left(\boldsymbol{\mu}_{\theta,i}, \boldsymbol{\Sigma}_{\theta,i}\right) \end{aligned} \tag{3.4}$$

Les paramètres de difficulté des items δ_j et l'effet groupe $\beta_{\theta,gp}$ sont supposés constants dans le temps. Comme $\boldsymbol{\theta}$ est supposée suivre une loi multinormale, le modèle fait partie de la famille des modèles logistiques à effets mixtes. Les paramètres de moyenne $\boldsymbol{\mu}_{\theta,i}$ et les paramètres de variance covariance $\boldsymbol{\Sigma}_{\theta,i}$ du modèle sont estimés par la méthode MML. La contrainte d'identifiabilité est $\mu_\theta^{(1)} = 0$. La matrice de variance covariance présente une structure sans contraintes.

3.2 Simulation des données

3.2.1 Simulation des réponses aux items

Toutes les études de simulation présentées font l'hypothèse que les réponses des patients ont été recueillies à partir d'une échelle validée avec un modèle de Rasch. Ce modèle a été préféré à d'autres modèles plus complexes de l'IRT comme le 2-PLM ou le 3-PLM en raison de ses propriétés psychométriques. Les méthodes d'analyse reposent également sur le modèle de Rasch pour ne pas pénaliser les résultats de ces méthodes. Les réponses de N individus à J items dichotomiques dans une étude comprenant trois temps d'évaluation ($T = 3$) ont été simulées avec un modèle de Rasch multidimensionnel. Le vecteur du trait latent $\theta = \left(\theta^{(1)}, \theta^{(2)}, \theta^{(3)}\right)'$ suit une loi normale $N_3\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ avec $\boldsymbol{\Sigma} = \sigma_\theta^2 \begin{pmatrix} 1 & \rho_\theta & \rho_\theta^2 \\ \rho_\theta & 1 & \rho_\theta \\ \rho_\theta^2 & \rho_\theta & 1 \end{pmatrix}$ avec σ_θ^2 la variance et ρ_θ le coefficient de corrélation entre deux mesures consécutives du trait latent. La matrice de variance covariance a une structure AR(1). Les corrélations sont supposées diminuer lorsque les mesures sont plus espacées dans le temps. La variance est

supposée constante dans le temps (σ^2_θ=1). Trois valeurs du coefficient de corrélation entre deux mesures consécutives ρ_θ ont été simulées : $\rho_\theta = 0,4$ (faible corrélation), $\rho_\theta = 0,7$ et $\rho_\theta = 0,9$ (forte corrélation). Ces valeurs sont fréquemment rencontrées en recherche clinique. Comme il est peu courant que des données longitudinales aient de faibles valeurs de corrélation, aucune valeur n'a été simulée inférieure à 0,4.

Le vecteur des moyennes vaut $\boldsymbol{\mu}_{g=0} = \left(-d_\theta - \Delta_\theta/2, \quad -\Delta_\theta/2, \quad d_\theta - \Delta_\theta/2\right)'$ pour le premier groupe et $\boldsymbol{\mu}_{g=1} = \left(-d_\theta + \Delta_\theta/2, \quad \Delta_\theta/2, \quad d_\theta + \Delta_\theta/2\right)'$ pour le second groupe. Le paramètre d_θ est égal à 0 ou 0,2 selon que les données ont été simulées sans ou avec un effet temps. Le paramètre Δ_θ est égal à 0 ou 0,5 selon que les données ont été simulées sans ou avec un effet groupe. Les moyennes simulées en fonction des valeurs de d_θ et Δ_θ sont représentées sur le graphique 3.1.

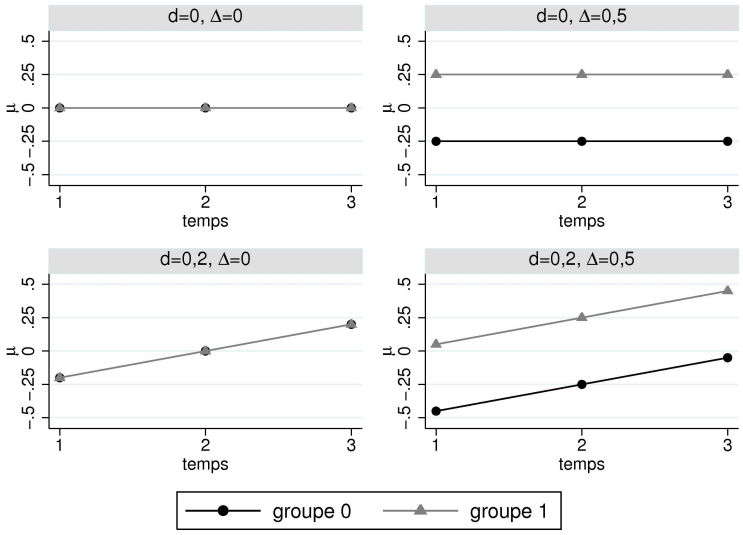

FIGURE 3.1 – Moyennes simulées en fonction de l'effet temps d_θ et de l'effet groupe Δ_θ

Les jeux de données simulés contiennent $N = 100$ ou $N = 200$ patients, tailles d'échantillon fréquemment rencontrées en pratique. Les patients ont été divisés en deux groupes de taille $N/2$ simulés indépendamment l'un de l'autre.

Les données sont supposées provenir d'une échelle composée de $J = 4$ ou $J = 7$ items dichotomiques. Les paramètres de difficulté des items ont été fixées à $\delta_1 = -1$, $\delta_2 = -0,5$, $\delta_3 = 0,5$, $\delta_4 = 1$ pour une échelle à 4 items et à $\delta_1 = -1,5$, $\delta_2 = -1$, $\delta_3 = -0,5$, $\delta_4 = 0$,

$\delta_5 = 0,5$, $\delta_6 = 1$, $\delta_7 = 1,5$ pour une échelle à 7 items. Ce choix résulte du fait que la plupart des dimensions des échelles en santé comprennent entre 1 et 10 items. Par exemple, les dimensions "limitations dues à l'état physique" et "vitalité" du SF-36 et la dimension "fonctionnement émotionnel" du QLQ-C30 sont formés de 4 items. La dimension Anxiété-Dépression du Duke Health Profile (DHP) et les dimensions Anxiété et Dépression de l'Hospital Anxiety and Depression Scale (HADS) comportent 7 items.

Les paramètres de difficulté des items sont centrés sur 0 pour se placer dans le cadre d'un questionnaire bien adapté aux patients. Ainsi, la moyenne des difficultés d'items est égale à la moyenne globale de la variable latente. Ceci reflète un questionnaire ni trop facile ni trop difficile pour les patients et un certain équilibre des réponses positives et négatives. Les valeurs des paramètres de difficulté des items sont cohérentes avec l'hypothèse de normalité de la variable latente. Le choix de ces valeurs évite la présence d'effets seuils qui auraient pénalisé la méthode basée sur le score. Le graphique 3.2 présente l'histogramme de la distribution des scores totaux dans chaque groupe de patients pour un jeu de données simulé avec les paramètres suivants : $d_\theta = 0$, $\Delta_\theta = 0$, $N = 200$, $J = 7$ et $\rho_\theta = 0,9$. Sur cet exemple, on observe que la distribution des scores semble proche de la normalité. On note également l'absence d'effets plancher ou plafond.

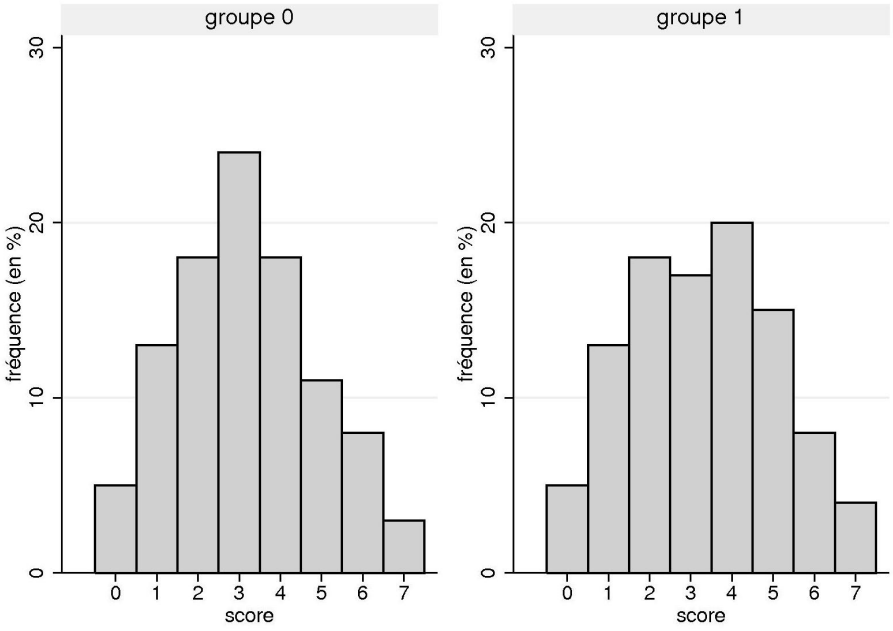

FIGURE 3.2 – Histogramme de la distribution des scores totaux dans chaque groupe ($d_\theta = 0$, $\Delta_\theta = 0$, $N = 200$, $J = 7$ et $\rho_\theta = 0,9$)

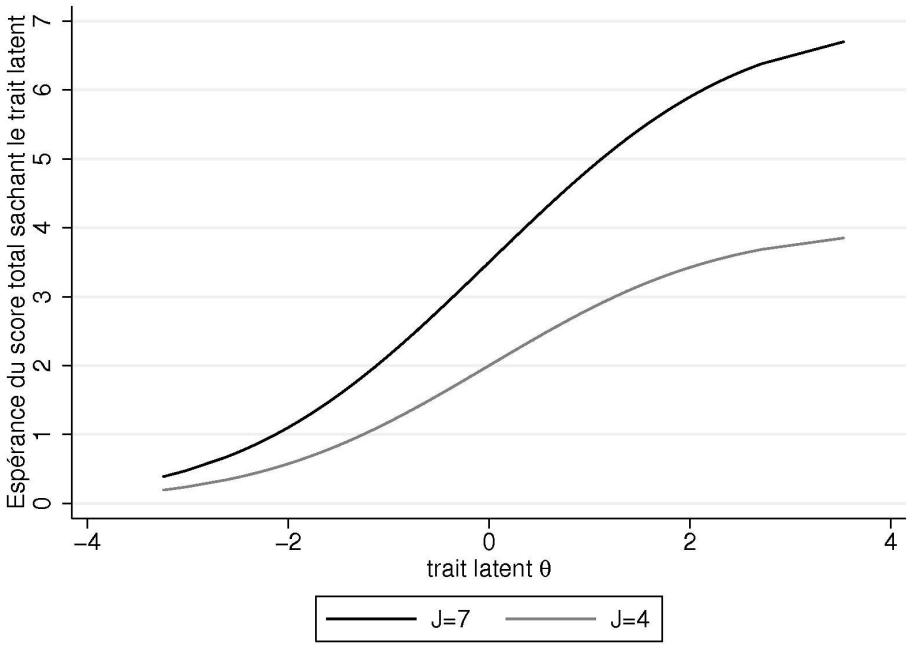

FIGURE 3.3 – Espérance du score total sachant le trait latent en fonction de la valeur du trait latent et du nombre d'items ($\theta \sim N(0,1)$)

Le graphique 3.3 présente l'espérance du score total en fonction de la valeur de la variable latente θ et du nombre d'items J pour une variable latente issue d'une loi normale centrée réduite. On observe une quasi-linéarité entre l'espérance du score et la variable latente sur l'intervalle [-2,2].

3.2.2 Simulation du dropout

Le processus de sortie d'étude a été simulé au moyen d'une deuxième variable latente : la propension au dropout notée χ. La probabilité qu'un patient sorte de l'étude au temps t dépend de sa propension au dropout. Le processus de sortie d'étude a été simulé au moyen du modèle suivant adapté du modèle 4-PLM [98] :

$$P(DO_i^{(t)} = 1|\chi_i^{(t)}, \pi_{min}^{(t)}, \pi_{max}^{(t)}) = \pi_{min}^{(t)} + (\pi_{max}^{(t)} - \pi_{min}^{(t)})\frac{exp(\chi_i^{(t)})}{1 + exp(\chi_i^{(t)})} \tag{3.5}$$

avec $DO_i^{(t)} = 1$ représente le fait que le patient i sorte de l'étude au temps t, $\pi_{min}^{(t)}$ la probabilité individuelle minimale de dropout au temps t et $\pi_{max}^{(t)}$ la probabilité individuelle maximale de dropout au temps t. $\pi_{min}^{(t)}$ et $\pi_{max}^{(t)}$ sont définis à partir de la proportion attendue de dropout au temps t : $\pi^{(t)} = \frac{\pi_{min}^{(t)} + \pi_{max}^{(t)}}{2}$.

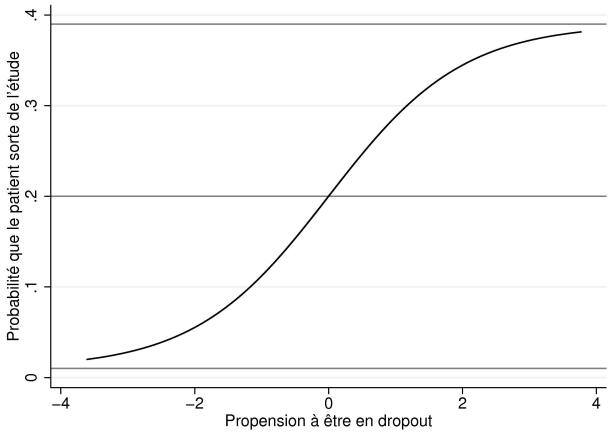

FIGURE 3.4 – Modèle 4-PLM : Probabilité qu'un patient sorte de l'étude au temps t en fonction de la propension à être en dropout au temps t pour $\pi_{min}^{(t)} = 1\%$, $\pi^{(t)} = 20\%$ et $\pi_{max}^{(t)} = 39\%$

Le graphique 3.4 présente un modèle 4-PLM pour lequel $\pi_{min}^{(t)} = 1\%$ asymptote horizontale basse de la courbe, $\pi^{(t)} = 20\%$ point d'inflexion de la courbe et $\pi_{max}^{(t)} = 39\%$ asymptote horizontale haute de la courbe. On observe que plus la propension à être en dropout augmente, plus la probabilité que le patient sorte de l'étude est grande.

Le type de processus de dropout est déterminé par la corrélation entre la valeur du trait latent au temps t, $\theta^{(t)}$ et la propension au dropout au temps t, $\chi^{(t)}$, notée $corr(\theta^{(t)}, \chi^{(t)}) = \rho_{\theta\chi}$ et supposée constante dans le temps. Selon la terminologie de Little et Rubin [63], le processus de dropout est MCAR lorsque la valeur du trait latent ne dépend pas de la propension au dropout ($\rho_{\theta\chi} = 0$). Le processus de dropout est MNAR lorsque ($\rho_{\theta\chi} \neq 0$). Les patients avec les plus bas niveaux pour le PRO étudié, en raison d'une progression de la maladie ou de l'augmentation des effets secondaires par exemple, ont plus de risque de sortir de l'étude prématurément que les autres patients [110]. Le dropout de type MNAR a donc été simulé sous l'hypothèse $\rho_{\theta\chi} < 0$. La propension au dropout χ_i suit une loi multinormale de moyenne (0 0 0)' et de matrice de variance covariance $\begin{pmatrix} 1 & \rho_{\theta\chi}^2\rho_\theta & \rho_{\theta\chi}^2\rho_\theta^2 \\ \rho_{\theta\chi}^2\rho_\theta & 1 & \rho_{\theta\chi}^2\rho_\theta \\ \rho_{\theta\chi}^2\rho_\theta^2 & \rho_{\theta\chi}^2\rho_\theta & 1 \end{pmatrix}$.

Les données sont supposées complètes au premier temps d'évaluation ($\pi^{(1)} = 0$). La quantité de dropout est ensuite la même à chaque temps et $\pi^{(t)}$ des patients restants sortent de l'étude à chaque temps ($t = 2, 3$). L'impact de la quantité de données manquantes sur les méthodes d'analyse a été investigué en faisant varier la proportion de dropout dans les données simulées : $\pi^{(t)} = 0\%$ (données complètes), 5%, 10% ou 20%. De même, l'impact de

l'informativité du dropout a été étudié en faisant varier la corrélation entre la variable latente et la propension au dropout : $\rho_{\theta\chi} = 0$ (dropout de type MCAR), $\rho_{\theta\chi} = -0,4; -0,7; -0,9$ (dropout de type MNAR de plus en plus informatif).

3.3 Analyse

3.3.1 Paramètres de variance covariance

Pour les méthodes où un modèle linéaire mixte est utilisé (SM, RM et PV), quatre structures de matrices de variance covariance ont été étudiées : la structure sans contraintes (UN), la structure auto-régressive hétérogène d'ordre 1 (ARH(1)), la structure auto-régressive d'ordre 1 (AR(1)) et la structure 'compound symmetry' hétérogène (CSH).
La structure UN est la plus générale. Elle est utilisée lorsqu'aucune hypothèse ne peut être faite sur la structure de variance covariance et nécessite donc d'estimer un grand nombre de paramètres. La structure ARH(1) prend en compte le fait que les mesures de chaque individu sont corrélées. Plus les temps de mesure sont éloignés, plus les corrélations sont supposées décroître. La structure AR(1) fait une hypothèse de plus que la structure ARH(1) : les variances sont alors supposées constantes dans le temps. Pour la structure CSH, les variances sont supposées différentes et les corrélations sont supposées constantes dans le temps.
Dans un modèle à trois temps d'évaluation, les matrices s'écrivent :

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{pmatrix} \text{ pour UN} \qquad \mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho^2 \\ \sigma_1\sigma_2\rho & \sigma_2^2 & \sigma_2\sigma_3\rho \\ \sigma_1\sigma_3\rho^2 & \sigma_2\sigma_3\rho & \sigma_3^2 \end{pmatrix} \text{ pour ARH(1)}$$

$$\mathbf{\Sigma} = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix} \text{ pour AR(1)} \quad \mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 & \sigma_2\sigma_3\rho \\ \sigma_1\sigma_3\rho & \sigma_2\sigma_3\rho & \sigma_3^2 \end{pmatrix} \text{ pour CSH}$$

où σ_t^2 désigne la variance au temps t, $\sigma_{tt'}$ la covariance entre t et t' et ρ le coefficient de corrélation de la mesure. Pour chaque méthode, la structure de variance covariance correspondant aux modèles avec les plus faibles AIC a été retenue.

3.3.2 Critères de comparaison

Les différentes méthodes sont comparées sur leur capacité à détecter un effet temps/groupe ou non et à estimer correctement cet effet. Pour chaque effet, les critères principaux étudiés sont le risque de première espèce, la puissance et le biais de l'estimation de l'effet. Le calcul de ces critères nécessite de définir la manière d'estimer les effets et de les tester.

Pour chaque méthode, l'effet temps entre deux mesures t et t' peut être estimé par : $\hat{d_{tt'}} = \hat{\mu}_{t'} - \hat{\mu}_t$. L'estimation de l'effet groupe est obtenue directement par $\hat{\beta}_{gp}$.

Le test de la présence d'un effet groupe utilise un test de Wald :

$H_0 : \mu_{g=0} = \mu_{g=1} \Leftrightarrow \beta_{gp} = 0$

$H_1 : \mu_{g=0} \neq \mu_{g=1} \Leftrightarrow \beta_{gp} \neq 0$

Le test de la présence d'un effet temps global utilise un test de Wald :

$H_0 : \mu^{(1)} = \mu^{(2)} = \mu^{(3)} = \mu \Leftrightarrow L\boldsymbol{\mu} = 0$

$H_1 : \exists i | \mu^{(i)} \neq \mu \Leftrightarrow L\boldsymbol{\mu} \neq 0$

$$\boldsymbol{L} = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}$$

Le risque de première espèce a été défini comme la proportion de rejet de H_0 sous l'hypothèse nulle. Il permet d'évaluer la capacité de la méthode à éviter de détecter à tort la présence d'un effet. De la même manière, la puissance a été définie comme la proportion de rejet de H_0 sous l'hypothèse alternative. Elle permet d'évaluer la capacité de la méthode à détecter à raison la présence d'un effet.

Le calcul du biais de l'estimation de l'effet nécessite de connaître la vraie valeur de l'effet pour pouvoir calculer l'écart entre la valeur estimée et la vraie valeur. Comme les données sont simulées à partir d'un modèle de Rasch, les vraies valeurs des effets sont connues pour le trait latent. Le paramètre d_θ vaut 0 ou 0,2 pour des données simulées sans ou avec un effet temps. Le paramètre Δ_θ vaut 0 ou 0,5 pour des données simulées sans ou avec un effet groupe. En revanche, les vraies valeurs des effets ne sont pas connues pour le score mais peuvent être approximées.

L'effet temps pour le score est approximé par d_S, la différence entre les scores attendus

à chaque temps sur l'ensemble de l'échantillon (tous groupes confondus). Pour $t = (2, 3)$,

$$d_S = E\left(S_i^{(t)}\right) - E\left(S_i^{(t-1)}\right) \tag{3.6}$$

Le score attendu à chaque temps est estimé par :

$$\begin{aligned} E\left(S_i^{(t)}\right) &= E\left(\sum_j Y_{ij}^{(t)}\right) \\ &= \sum_j E\left(Y_{ij}^{(t)}\right) \\ &= \sum_j P\left(Y_{ij}^{(t)} = 1\right) \\ &= \sum_j \int_{\mathbb{R}} P\left(Y_{ij}^{(t)} = 1 | \theta_i^{(t)}, \delta_j\right) g(\theta_i^{(t)}/\mu^{(t)}, \sigma^2) d\theta_i \\ &= \sum_j \int_{\mathbb{R}} \frac{\exp\left(\theta_i^{(t)} - \delta_j\right)}{1 + \exp(\theta_i^{(t)} - \delta_j)} g(\theta_i^{(t)}/\mu^{(t)}, \sigma^2) d\theta_i \end{aligned} \tag{3.7}$$

avec $g(\theta_i^{(t)}/\mu^{(t)}, \sigma^2)$ la distribution normale de moyenne $\mu^{(t)}$ et de variance σ^2. Les intégrales sont estimées par les quadratures de Gauss-Hermite. Lorsque l'effet temps simulé entre deux temps consécutifs pour la variable latente vaut $d_\theta = 0$, l'effet temps pour le score est approximé à $d_S = 0$. Lorsque l'effet temps simulé entre deux temps consécutifs pour la variable latente vaut $d_\theta = 0,2$, l'effet temps pour le score est approximé à $d_S = 0,15$ pour $J = 4$ et $d_S = 0,25$ pour $J = 7$.

De la même façon, l'effet groupe pour le score est approximé par Δ_S, la différence des scores attendus dans chaque groupe :

$$\Delta_S = E\left(S_i^{(t)} | g = 1\right) - E\left(S_{i'}^{(t)} | g = 0\right) \tag{3.8}$$

Le score attendu dans chaque groupe est estimé par :

$$\begin{aligned} E\left(S_i^{(t)} | g\right) &= E\left(\sum_j Y_{ij}^{(t)} | g\right) \\ &= \sum_j E\left(Y_{ij}^{(t)} | g\right) \end{aligned} \tag{3.9}$$

$$
\begin{aligned}
&= \sum_j P\left(Y_{ij}^{(t)} = 1|g\right) \\
&= \sum_j \int_{\mathbb{R}} P\left(Y_{ij}^{(t)} = 1|\theta_i^{(t)}, \delta_j\right) g(\theta_i^{(t)}/\mu_g^{(t)}, \sigma^2) d\theta_i \\
&= \sum_j \int_{\mathbb{R}} \frac{\exp\left(\theta_i^{(t)} - \delta_j\right)}{1 + \exp(\theta_i^{(t)} - \delta_j)} g(\theta_i^{(t)}/\mu_g^{(t)}, \sigma^2) d\theta_i
\end{aligned}
$$

avec $g(\theta_i^{(t)}/\mu_g^{(t)}, \sigma^2)$ la distribution normale de moyenne $\mu_g^{(t)}$ et de variance σ^2. Lorsque l'effet groupe simulé pour la variable latente vaut $\Delta_\theta = 0$, l'effet groupe pour le score est approximé à $\Delta_S = 0$. Lorsque l'effet groupe simulé pour la variable latente vaut $\Delta_\theta = 0,5$, l'effet groupe pour le score est approximé à $d_S = 0,38$ pour $J = 4$ et $d_S = 0,63$ pour $J = 7$.

L'ensemble des valeurs des effets temps et groupe simulées pour la variable latente θ et approximées pour le score sont résumées dans le tableau 3.1 en fonction du nombre d'items J.

TABLE 3.1 – Effets simulés pour le trait latent et le score

d_θ	Δ_θ	J	d_S	Δ_S
0	0	4	0	0
		7	0	0
	0,5	4	0	0,38
		7	0	0,63
0,2	0	4	0,15	0
		7	0,25	0
	0,5	4	0,15	0,38
		7	0,25	0,63

3.4 Outils logiciels

La simulation des données et leur analyse a fait appel aux logiciels généralistes Stata et SAS.

3.4.1 Stata

Le module *simirt* permet de simuler des données selon un modèle de l'IRT. Il dispose d'un large éventail de possibilités pour la simulation d'items dichotomiques ou polytomiques

selon différents modèles. Dans la version initiale du module, il n'était possible de simuler que deux ensembles de réponses aux items liés chacun à un trait latent spécifique, les deux traits latents pouvant être cependant corrélés. Le module a été adapté au sein de l'équipe afin de permettre de simuler plus de deux ensembles de réponses aux items.

Le module *raschtest* [50] estime les paramètres d'un modèle de Rasch avec différentes méthodes d'estimation possibles (maximum de vraisemblance conditionnelle (CML), maximum de vraisemblance marginale (MML) ou équations d'estimation généralisées (GEE)). Dans ce travail, l'estimation par maximum de vraisemblance marginale a permis d'obtenir les estimations EAP utilisées dans les méthodes RM et PV.

Le module *gllamm* (Generalized Linear Latent And Mixed Models) [124, 88] permet l'estimation des paramètres de modèles linéaires généralisés mixtes à variable latente. Ce module a été utilisé pour l'estimation par maximum de vraisemblance marginale des paramètres de la distribution de la variable latente ainsi que les paramètres de difficulté des items du modèle de Rasch longitudinal de la méthode LRM. Afin de diminuer les temps de calcul du module dans ce travail, les valeurs initiales sont estimées préalablement au lieu d'être fixées aux valeurs par défaut du logiciel.

3.4.2 SAS

La procédure MIXED [59] permet l'estimation des paramètres de modèles linéaires mixtes par différentes méthodes d'estimation (maximum de vraisemblance, maximum de vraisemblance restreinte) avec une grande variété de structures de matrices de variance covariance possibles. Dans ce travail, l'estimation par maximum de vraisemblance restreinte avec la procédure MIXED a été utilisée dans les méthodes SM, RM et PV.

Chapitre 4

Comparaison de méthodes d'analyse de données subjectives longitudinales complètes

Cette première étude de simulation vise à évaluer les performances des quatre méthodes Score and Mixed models (SM), Rasch and Mixed models (RM), Plausible Values (PV) et Longitudinal Rasch Mixed model (LRM) sur des données longitudinales, complètes, de patients issus d'un seul et même groupe. Les données sont supposées issues d'une échelle validée avec un modèle de Rasch. Les résultats de la méthode SM nous renseigneront sur la pertinence d'utiliser une analyse basée sur la CTT alors que les données sont issues d'un modèle IRT. La comparaison des différentes méthodes basées sur l'IRT nous permettra de définir s'il est plus adéquat d'estimer l'effet temps à partir des estimations individuelles du trait latent (méthodes RM et PV) ou à partir des estimations des paramètres de population (méthode LRM). De plus, différentes méthodes d'obtention des valeurs individuelles du trait latent seront comparées : l'estimation par EAP pour la méthode RM et l'imputation de valeurs plausibles pour la méthode PV.

Le design de l'étude comportant des données complètes et un seul groupe de patient a été choisi afin de comparer les méthodes dans un cadre simple et favorable. Ce type d'étude avec un seul groupe de patients est fréquemment rencontré dans le suivi de patients après une chirurgie ou un traitement dont le but est d'évaluer l'impact du traitement ou de la chirurgie sur le PRO évalué. C'est le cas dans l'exemple sur l'évolution de la qualité de vie avant

et après chirurgie dans l'hyperparathyroïdie primaire présenté dans le chapitre 7 p.101. Le choix des données complètes permet de s'affranchir de possibles biais dans les résultats qui pourraient être dus à l'absence de données.

Le but de cette étude de simulation était de comparer les méthodes les unes aux autre grâce aux trois critères définis précédemment : le risque de première espèce, la puissance et le biais de l'estimation de l'effet temps. L'impact de paramètres tels que la taille d'échantillon, la longueur du questionnaire et la corrélation entre les mesures de la variable latente sur les performances de chaque méthode a également été évalué. L'ensemble des paramètres de l'étude sont rappelés dans le tableau 4.1. La combinaison des différents paramètres nous a amené à considérer 24 cas différents. Cette étude a fait l'objet d'une publication dans Statistics in Medicine [13] (cf. annexe A p.129).

TABLE 4.1 – Paramètres utilisés pour la simulation des données et l'analyse

	Nombre de cas	24
	Nombre de jeux simulés/cas	500
	Taille d'échantillon (N)	100 ; 200
	Nombre d'items (J)	4 ; 7
Variable latente	Effet temps (d_θ)	0 ; 0,2
	Variance (σ^2_θ)	1
	Corrélation (ρ_θ)	0,4 ; 0,7 ; 0,9
	Effet groupe (Δ_θ)	0
Score	Effet temps (d_S)	0 ; 0,15 (J=4) ; 0,25 (J=7)
	Effet groupe (Δ_S)	0
Dropout	Proportion ($\pi^{(t)}$)	0%
Analyse	Méthodes comparées	Score and Mixed Models (SM)
		Rasch Mixed model (RM)
		Plausible Values (PV)
		Longitudinal Rasch Mixed model (LRM)
	Structures de matrice	sans contraintes (UN)
		auto-régressive d'ordre 1 (AR(1))
		AR(1) hétérogène (ARH(1))
		"compound symmetry" hétérogène (CSH)

4.1 Résultats

Les structures de matrice de variance covariance UN, AR(1), ARH(1) et CSH ont été étudiées pour les méthodes SM, RM et PV. La structure UN a été étudiée pour la méthode LRM. Pour la méthode SM, l'AIC était minimisé plus souvent dans les modèles avec une structure de variance covariance AR(1) que dans les modèles avec une autre structure pour des valeurs simulées du coefficient de corrélation $\rho_\theta = 0,4$ et $\rho_\theta = 0,7$. Lorsque $\rho_\theta = 0,9$, l'AIC était plus souvent minimisé par les modèles ayant une structure de variance covariance CSH que par les modèles ayant une autre structure. Les résultats présentés ci-dessous ont été obtenus avec une structure AR(1). Pour la méthode RM, l'AIC était minimisé plus souvent dans les modèles avec une structure de variance covariance AR(1) que dans les modèles avec une autre structure. Les résultats présentés ci-dessous ont été obtenus avec une structure AR(1).

4.1.1 Risque de première espèce et puissance

TABLE 4.2 – Risque de première espèce ($\hat{\alpha}$) et puissance ($1 - \hat{\beta}$) pour les méthodes Score and Mixed models (SM), Rasch Mixed model (RM), Plausible Values (PV) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J) et de la corrélation de la variable latente (ρ_θ). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour les méthodes SM, RM et PV et une structure UN pour la méthode LRM.

			SM		RM		PV		LRM	
N	J	ρ_θ	$\hat{\alpha}$	$1-\hat{\beta}$	$\hat{\alpha}$	$1-\hat{\beta}$	$\hat{\alpha}$	$1-\hat{\beta}$	$\hat{\alpha}$	$1-\hat{\beta}$
100	4	0,4	0,054	0,384	0,052	0,146	0,062	0,166	0,054	0,388
		0,7	0,040	0,372	0,040	0,120	0,042	0,142	0,044	0,423
		0,9	0,038	0,424	0,042	0,156	0,056	0,156	0,048	0,508
	7	0,4	0,036	0,428	0,034	0,108	0,052	0,162	0,044	0,470
		0,7	0,046	0,522	0,048	0,096	0,050	0,150	0,058	0,568
		0,9	0,044	0,564	0,044	0,070	0,054	0,160	0,046	0,688
200	4	0,4	0,060	0,634	0,058	0,280	0,058	0,288	0,052	0,654
		0,7	0,048	0,678	0,048	0,276	0,060	0,304	0,056	0,721
		0,9	0,060	0,756	0,060	0,250	0,052	0,330	0,067	0,826
	7	0,4	0,050	0,810	0,048	0,184	0,042	0,304	0,050	0,822
		0,7	0,036	0,880	0,038	0,154	0,040	0,314	0,048	0,916
		0,9	0,048	0,918	0,048	0,148	0,038	0,306	0,058	0,955

Le tableau 4.2 présente les estimations du risque de première pour chaque méthode étudiée et pour différentes valeurs de la taille d'échantillon, du nombre d'items et de la corrélation de la variable latente. Toutes les méthodes ont des résultats comparables en terme de risque de première espèce. La variation de la taille d'échantillon, du nombre d'items et du coefficient de corrélation ne semble pas avoir d'impact sur le risque de première espèce.
Pour toutes les méthodes, le risque de première espèce est correctement maintenu proche de la valeur attendue de 5%. De plus, tous les intervalles de confiance à 95% contiennent la valeur 5%. Le risque de première espèce est compris entre 3,6% et 6,0% pour la méthode SM, 3,4% et 5,8% pour la méthode RM, 3,8% et 6,2% pour la méthode PV et 4,4% et 6,7% pour la méthode LRM.

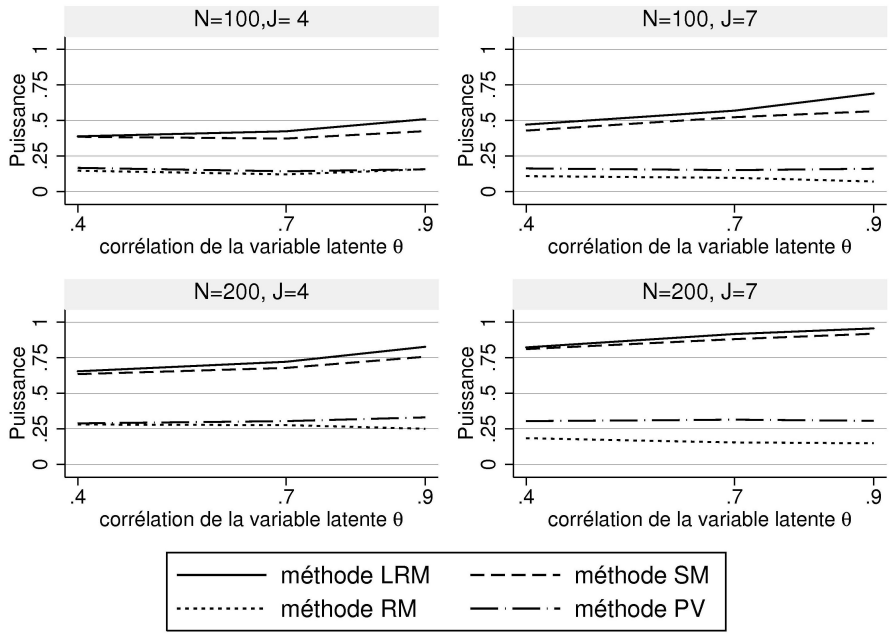

FIGURE 4.1 – Puissance pour les méthodes Score and Mixed models (SM), Rasch Mixed model (RM), Plausible Values (PV) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J) et de la corrélation de la variable latente (ρ_θ). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour les méthodes SM, RM et PV et une structure UN pour la méthode LRM.

Le tableau 4.2 et le graphique 4.1 présentent les estimations de la puissance pour chaque méthode étudiée et pour différentes valeurs de la taille d'échantillon, du nombre d'items et de la corrélation de la variable latente. La puissance de la méthode LRM est comparable à celle de SM quelle que soit la valeur des paramètres étudiés. Il semble que la puissance

de la méthode LRM soit systématiquement un peu plus élevée que celle de la méthode SM lorsque le coefficient de corrélation ρ_θ vaut 0,7 ou 0,9. Les méthodes RM et PV présentent des puissances faibles, beaucoup plus petites que celles des méthodes SM et LRM. Les différences de puissance entre les méthodes RM/PV et les méthodes SM/LRM augmentent avec la taille de l'échantillon, le nombre d'items et le coefficient de corrélation. La différence est la plus importante lorsque $N = 200$ et $J = 7$: LRM et SM présentent des puissances supérieures à 80% alors que la puissance de RM varie de 15 à 18% et la puissance de PV est proche de 30%.

Ces différences de puissance s'expliquent par le fait que la puissance des méthodes SM et LRM augmentent avec la taille de l'échantillon, le nombre d'items et la corrélation. La puissance la plus petite pour SM et LRM est proche de 38% pour $N = 100$, $J = 4$ et $\rho_\theta = 0,4$ alors que la puissance la plus grande dépasse 90% pour $N = 200$, $J = 7$ et $\rho_\theta = 0,9$. Au contraire, la puissance des méthodes RM et PV est plutôt stable quelle que soit la valeur des paramètres étudiés. On observe néanmoins une faible augmentation de la puissance lorsque la taille d'échantillon augmente pour ces deux méthodes.

4.1.2 Estimation de l'effet temps

Le tableau 4.3 présente les estimations de l'effet temps entre le premier et le deuxième temps de mesure pour chaque méthode étudiée et pour différentes valeurs de la taille d'échantillon, du nombre d'items et de la corrélation de la variable latente. Lorsqu'aucun effet temps n'a été simulé $d_\theta = 0$, toutes les méthodes présentent des estimations de l'effet temps proches de 0 donc non-biaisées.
Lorsque l'effet temps vaut $d_\theta = 0,2$, les méthodes RM et PV sous-estiment l'effet temps. Pour toutes les valeurs des paramètres étudiés, les estimations de l'effet temps sont biaisées. En revanche, les méthodes SM et LRM présentent des estimations non-biaisées de l'effet temps lorsque l'effet temps simulé vaut $d_\theta = 0,2$.

TABLE 4.3 – Estimations de l'effet temps entre le temps 1 et le temps 2 ($\hat{d}_{12}$) pour les méthodes Score and Mixed models (SM), Rasch Mixed model (RM), Plausible Values (PV) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J) et de la corrélation de la variable latente (ρ_θ). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour les méthodes SM, RM et PV et une structure UN pour la méthode LRM. Valeurs moyennes de $\hat{d}_{12}$ et erreurs standards (s.e.).

					SM		RM		PV		LRM	
N	J	ρ_θ	d_θ	d_S	$\hat{d}_{12}$	(s.e.)	$\hat{d}_{12}$	(s.e.)	$\hat{d}_{12}$	(s.e.)	$\hat{d}_{12}$	(s.e.)
100	4	0,4	0	0	0,010	(0,007)	0,006	(0,004)	0,008	(0,006)	0,013	(0,009)
		0,7	0	0	-0,003	(0,006)	-0,002	(0,004)	0,001	(0,006)	-0,004	(0,008)
		0,9	0	0	-0,007	(0,006)	-0,005	(0,003)	-0,003	(0,006)	-0,010	(0,008)
	7	0,4	0	0	0,005	(0,009)	0,002	(0,004)	0,000	(0,006)	0,005	(0,007)
		0,7	0	0	0,008	(0,009)	0,004	(0,004)	0,008	(0,006)	0,007	(0,007)
		0,9	0	0	0,011	(0,007)	0,005	(0,003)	0,005	(0,006)	0,009	(0,006)
200	4	0,4	0	0	-0,003	(0,005)	-0,001	(0,003)	0,001	(0,004)	-0,003	(0,006)
		0,7	0	0	0,007	(0,005)	0,004	(0,003)	0,007	(0,004)	0,009	(0,006)
		0,9	0	0	-0,007	(0,004)	-0,004	(0,002)	-0,004	(0,004)	-0,009	(0,006)
	7	0,4	0	0	-0,003	(0,007)	-0,001	(0,003)	-0,002	(0,004)	-0,002	(0,005)
		0,7	0	0	0,002	(0,006)	0,001	(0,003)	0,003	(0,004)	0,003	(0,005)
		0,9	0	0	-0,005	(0,005)	-0,002	(0,002)	-0,003	(0,004)	-0,004	(0,004)
100	4	0,4	0,2	0,15	0,139	(0,007)	0,023*	(0,002)	0,082*	(0,006)	0,184	(0,009)
		0,7	0,2	0,15	0,138	(0,006)	0,022*	(0,002)	0,073*	(0,006)	0,186	(0,008)
		0,9	0,2	0,15	0,156	(0,006)	0,022*	(0,002)	0,084*	(0,006)	0,211	(0,008)
	7	0,4	0,2	0,25	0,237	(0,009)	0,009*	(0,001)	0,080*	(0,006)	0,189	(0,007)
		0,7	0,2	0,25	0,246	(0,008)	0,010*	(0,001)	0,091*	(0,006)	0,197	(0,007)
		0,9	0,2	0,25	0,250	(0,007)	0,007*	(0,001)	0,084*	(0,006)	0,202	(0,006)
200	4	0,4	0,2	0,15	0,148	(0,005)	0,023*	(0,001)	0,093*	(0,004)	0,197	(0,006)
		0,7	0,2	0,15	0,155	(0,004)	0,023*	(0,001)	0,086*	(0,004)	0,207	(0,005)
		0,9	0,2	0,15	0,150	(0,004)	0,022*	(0,001)	0,087*	(0,004)	0,202	(0,006)
	7	0,4	0,2	0,25	0,255	(0,007)	0,011*	(0,001)	0,089*	(0,004)	0,203	(0,005)
		0,7	0,2	0,25	0,243	(0,006)	0,008*	(0,001)	0,082*	(0,004)	0,194	(0,005)
		0,9	0,2	0,25	0,254	(0,005)	0,008*	(0,001)	0,088*	(0,004)	0,202	(0,004)

* indique que le test de Student est significatif à 5%
($H_0 : \mu_{\hat{d}_{12}} = d_S$ ou $H_0 : \mu_{\hat{d}_{12}} = d_\theta$)

4.1.3 Estimation des paramètres de variance

4.1.3.1 Estimations de la variance

Le tableau 4.4 présente les estimations des variances pour chaque méthode étudiée et pour différentes valeurs de la taille d'échantillon, du nombre d'items et de la corrélation de la variable latente. La variance a été simulée à 1 pour la variable latente. La vraie valeur de la variance pour le score n'a pas été approximée et la présence ou l'absence d'un biais de la variance pour le score n'a pas pu être testée. Seules les estimations de la variance au temps 1 sont présentées dans le tableau pour la méthode LRM. Les estimations aux temps 2 et 3 sont proches des estimations au temps 1.

La méthode RM semble sous-estimer la variance. La variance est estimée entre 0,435 et 0,575 au lieu de la valeur simulée $\sigma^2_\theta = 1$. La méthode PV donne des estimations de la variance proches de 1 mais la plupart sont légèrement sur-estimées lorsque l'effet temps simulé vaut $d_\theta = 0,2$. La méthode LRM donne également des estimations de la variance proches de 1. L'estimation de la variance semble être légèrement biaisée lorsque la corrélation entre les mesures consécutives de la variable latente est élevée $\rho_\theta = 0,7$ ou $\rho_\theta = 0,9$. Les estimations présentées dans le tableau 4.4 pour la méthode LRM, dont la matrice de variance covariance a une structure sans contraintes, correspondent au premier temps d'évaluation. Les mêmes tendances sont observées aux deux derniers temps. La variance est moins souvent sur-estimée par la méthode LRM que par la méthode PV.

4.1.3.2 Estimations du coefficient de corrélation

Le tableau 4.5 présente les estimations du coefficient de corrélation pour chaque méthode étudiée et pour différentes valeurs de la taille d'échantillon, du nombre d'items et de la corrélation de la variable latente. La vraie valeur du coefficient de corrélation pour le score n'a pas été approximée et la présence ou l'absence d'un biais du coefficient de corrélation pour le score n'a pas pu être testée. Seules les estimations de la corrélation entre les temps 1 et 2 sont présentées dans le tableau pour la méthode LRM. Les estimations de la corrélation entre les temps 2 et 3 sont proches des estimations entre les temps 1 et 2.

Les méthodes RM et PV donnent des estimations systématiquement biaisées du coefficient de corrélation. Les estimations obtenues sont très sous-estimées. La méthode LRM montre des estimations du coefficient de corrélation proches des valeurs simulées. Néanmoins, certains cas présentent des estimations biaisées, notamment presque tous les cas où la corrélation est

TABLE 4.4 – Estimations de la variance ($\hat{\sigma}^2$) pour les méthodes Score and Mixed models (SM), Rasch Mixed model (RM), Plausible Values (PV) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J) et de la corrélation de la variable latente (ρ_θ). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour les méthodes SM, RM et PV et une structure UN pour la méthode LRM. Valeurs moyennes de $\hat{\sigma}^2$ et erreurs standards (s.e.).

					SM		RM		PV		LRM	
d_θ	N	J	ρ_θ	σ^2_θ	$\hat{\sigma}^2_S$	(s.d.)	$\hat{\sigma}^2_\theta$	(s.e.)	$\hat{\sigma}^2_\theta$	(s.e.)	$\hat{\sigma}^2_\theta$	(s.e.)
0	100	4	0,4	1	1,331	(0,088)	0,460*	(0,008)	1,040*	(0,013)	1,027	(0,021)
			0,7	1	1,328	(0,089)	0,451*	(0,008)	1,023	(0,012)	1,047*	(0,019)
			0,9	1	1,329	(0,095)	0,449*	(0,008)	1,026*	(0,013)	1,078*	(0,019)
		7	0,4	1	2,857	(0,202)	0,575*	(0,007)	1,026*	(0,013)	1,013	(0,014)
			0,7	1	2,850	(0,211)	0,568*	(0,007)	1,016	(0,013)	1,030*	(0,013)
			0,9	1	2,850	(0,228)	0,568*	(0,007)	1,012	(0,013)	1,032*	(0,013)
0	200	4	0,4	1	1,326	(0,059)	0,435*	(0,005)	1,006	(0,008)	0,999	(0,013)
			0,7	1	1,327	(0,063)	0,435*	(0,005)	1,008	(0,009)	1,016	(0,013)
			0,9	1	1,331	(0,059)	0,441*	(0,005)	1,013	(0,008)	1,049*	(0,013)
		7	0,4	1	2,849	(0,140)	0,562*	(0,004)	0,999	(0,008)	1,013	(0,010)
			0,7	1	2,849	(0,140)	0,560*	(0,004)	1,007	(0,008)	1,001	(0,009)
			0,9	1	2,840	(0,154)	0,555*	(0,005)	0,994	(0,009)	0,981*	(0,009)
0,2	100	4	0,4	1	1,319	(0,085)	0,456*	(0,008)	1,040*	(0,013)	1,004	(0,019)
			0,7	1	1,320	(0,086)	0,457*	(0,008)	1,034*	(0,013)	1,023	(0,018)
			0,9	1	1,319	(0,088)	0,453*	(0,007)	1,030	(0,012)	1,095*	(0,019)
		7	0,4	1	2,838	(0,189)	0,453*	(0,008)	1,049*	(0,012)	1,006	(0,014)
			0,7	1	2,851	(0,209)	0,453*	(0,008)	1,047*	(0,013)	1,018	(0,013)
			0,9	1	2,831	(0,227)	0,454*	(0,008)	1,054*	(0,014)	1,035*	(0,014)
0,2	200	4	0,4	1	1,322	(0,059)	0,451*	(0,005)	1,033*	(0,008)	0,997	(0,014)
			0,7	1	1,318	(0,061)	0,445*	(0,005)	1,026*	(0,009)	1,025	(0,013)
			0,9	1	1,322	(0,064)	0,454*	(0,005)	1,038*	(0,009)	1,051*	(0,013)
		7	0,4	1	2,843	(0,131)	0,436*	(0,005)	1,031*	(0,008)	0,996	(0,010)
			0,7	1	2,836	(0,145)	0,445*	(0,006)	1,041*	(0,009)	1,013	(0,009)
			0,9	1	2,840	(0,158)	0,437*	(0,006)	1,026*	(0,009)	0,997	(0,009)

* indique que le test significatif à 5% ($H_0 : \mu_{\hat{\sigma}_\theta} = \sigma_\theta$).
La vraie valeur de la variance pour le score σ^2_S n'est pas connue.

TABLE 4.5 – Estimations du coefficient de corrélation ($\hat{\rho}$) pour les méthodes Score and Mixed models (SM), Rasch Mixed model (RM), Plausible Values (PV) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J) et de la corrélation de la variable latente (ρ_θ). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour les méthodes SM, RM et PV et une structure UN pour la méthode LRM. Valeurs moyennes de $\hat{\rho}$ et erreurs standards (s.e.).

				SM		RM		PV		LRM	
d_θ	N	J	ρ_θ	$\hat{\rho}_S$	s.d.	$\hat{\rho}_\theta$	s.e.	$\hat{\rho}_\theta$	s.e.	$\hat{\rho}_\theta$	s.e.
0	100	4	0,4	0,173	(0,072)	0,173*	(0,003)	0,070*	(0,003)	0,426*	(0,011)
			0,7	0,302	(0,073)	0,302*	(0,003)	0,134*	(0,003)	0,695	(0,009)
			0,9	0,383	(0,069)	0,383*	(0,003)	0,168*	(0,003)	0,855*	(0,006)
		7	0,4	0,219	(0,072)	0,219*	(0,003)	0,064*	(0,003)	0,405	(0,008)
			0,7	0,391	(0,064)	0,391*	(0,003)	0,122*	(0,003)	0,717*	(0,006)
			0,9	0,497	(0,060)	0,496*	(0,003)	0,155*	(0,003)	0,871*	(0,004)
0	200	4	0,4	0,171	(0,051)	0,171*	(0,002)	0,069*	(0,002)	0,410	(0,008)
			0,7	0,299	(0,051)	0,299*	(0,002)	0,127*	(0,003)	0,715*	(0,007)
			0,9	0,388	(0,045)	0,387*	(0,002)	0,167*	(0,003)	0,875*	(0,005)
		7	0,4	0,222	(0,048)	0,222*	(0,002)	0,073*	(0,002)	0,405	(0,005)
			0,7	0,392	(0,044)	0,391*	(0,002)	0,125*	(0,002)	0,714*	(0,005)
			0,9	0,500	(0,040)	0,499*	(0,002)	0,157*	(0,003)	0,889*	(0,003)
0,2	100	4	0,4	0,172	(0,068)	0,172*	(0,003)	0,069*	(0,003)	0,420	(0,011)
			0,7	0,296	(0,066)	0,296*	(0,003)	0,128*	(0,003)	0,703	(0,008)
			0,9	0,386	(0,067)	0,386*	(0,003)	0,163*	(0,003)	0,856*	(0,006)
		7	0,4	0,218	(0,071)	0,169*	(0,003)	0,067*	(0,003)	0,397	(0,008)
			0,7	0,389	(0,062)	0,296*	(0,003)	0,120*	(0,003)	0,704	(0,006)
			0,9	0,498	(0,058)	0,381*	(0,003)	0,164*	(0,004)	0,878*	(0,004)
0,2	200	4	0,4	0,173	(0,049)	0,173*	(0,002)	0,074*	(0,002)	0,419*	(0,008)
			0,7	0,298	(0,046)	0,298*	(0,002)	0,126*	(0,002)	0,712	(0,006)
			0,9	0,382	(0,049)	0,382*	(0,002)	0,163*	(0,003)	0,864*	(0,005)
		7	0,4	0,224	(0,049)	0,168*	(0,002)	0,071*	(0,002)	0,407	(0,005)
			0,7	0,387	(0,046)	0,295*	(0,002)	0,125*	(0,002)	0,705	(0,005)
			0,9	0,504	(0,046)	0,387*	(0,002)	0,164*	(0,003)	0,894	(0,003)

* indique que le test significatif à 5% ($H_0 : \mu_{\hat{\rho}_\theta} = \rho_\theta$).
La vraie valeur du coefficient de corrélation pour le score ρ_S n'est pas connue.

élevée $\rho_\theta = 0,9$. Lorsque l'estimation est biaisée, le coefficient de corrélation est légèrement sur-estimé pour $\rho_\theta = 0,4$ et $\rho_\theta = 0,7$. Il est légèrement sous-estimé pour $\rho_\theta = 0,9$.

4.2 Discussion

Les méthodes PV et RM présentent des puissances peu élevées par rapport aux méthodes SM et LRM. Ce manque de puissance pour RM et PV s'explique par la sous-estimation de l'effet temps lorsqu'un effet temps avait été simulé alors que les estimations sont non-biaisées pour SM et LRM. La mauvaise estimation de l'effet temps semble due à la première étape des méthodes RM et PV qui consiste à obtenir des estimations individuelles du trait latent.

La méthode RM utilise l'estimateur EAP pour estimer les valeurs individuelles du trait latent. En effet, les valeurs de la variable latente ne sont pas observables et doivent être estimées. L'incertitude liée à l'estimation n'est pas prise en compte lorsque des estimateurs tels que l'EAP sont utilisés dans la méthode RM car les estimations individuelles de la variable latente sont ensuite traitées comme si elles avaient été observées. Le traitement des estimations individuelles comme des observations peut entraîner une mauvaise estimation des erreurs standard des paramètres de population voire même entraîner un biais dans l'estimation des paramètres de population [72].
La moyenne des estimations EAP est un estimateur non-biaisé de la moyenne de la population. En revanche, la variance des estimations EAP sous-estime la variance de la population [71]. La distribution des EAP est dite resserrée autour de sa moyenne ("shrinkage"). Dans cette étude, les estimations EAP sont obtenues sur l'ensemble de l'échantillon, tous temps de mesure confondus. La distribution des EAP est donc resserrée autour de 0, la moyenne de la variable latente simulée tous temps confondus, qu'un effet temps ait été simulé ou non. L'effet temps entre deux temps est estimé par la différence des moyennes estimées à chaque temps. Or, l'écart estimé entre les moyennes est trop petit en raison de la distribution resserrée des EAP, ce qui explique la sous-estimation de l'effet temps par la méthode RM.
Il a été montré que le biais de la variance des EAP diminue avec l'augmentation du nombre d'items et que ce biais est minime lorsque le nombre d'items est supérieur à 20 [120]. Dans cette étude de simulation, le nombre d'items était de 4 ou 7, ce qui correspond aux valeurs fréquemment rencontrées pour des dimensions d'échelles utilisées en santé où les dimensions

comptent rarement plus de 10 items. Il semble donc difficile d'éviter le biais dû au resserrement de la distribution des EAP dans l'analyse de Patient-Reported Outcomes.

La méthode PV utilise l'imputation de valeurs plausibles pour estimer les valeurs individuelles du trait latent. Les valeurs plausibles sont tirées dans la distribution postérieure du trait latent. Contrairement à l'EAP, les patients ayant le même score et la même distribution postérieure peuvent avoir des estimations différentes. La moyenne des estimations obtenues par valeurs plausibles est un estimateur non-biaisé de la moyenne de la population. L'imputation de valeurs plausibles permet de corriger le biais de la variance.
Dans cette étude, la variance simulée était de $\sigma_\theta^2 = 1$. La méthode RM présente des variances très sous-estimées proches de 0,45. La méthode PV présente des variances légèrement surestimées.
Si l'estimation de la variance a bien été corrigée par l'utilisation de valeurs plausibles au lieu de l'estimateur EAP, la méthode PV sous-estime toujours l'effet temps et présente des puissances beaucoup moins élevées que les méthodes SM et LRM qui ont des estimations non-biaisées de l'effet temps. De plus, la corrélation entre les mesures est sous-estimée. La méthode PV, tout comme la méthode RM, semble ne pas être adéquate pour l'analyse de PRO dans un cadre longitudinal. Il semble que l'utilisation d'un modèle IRT longitudinal est plus adéquate.

De plus, l'utilisation de l'imputation de valeurs plausibles avec un modèle de Rasch pose également un problème quant au non-respect d'une des propriétés du modèle. En effet, le modèle de Rasch présente la propriété d'exhaustivité du score sur le trait latent. Ainsi, tous les patients ayant le même score ont la même estimation du trait latent. Lorsque l'imputation de valeurs plausibles est utilisée, les patients avec le même score peuvent avoir des estimations du trait latent différente mais surtout des patients ayant le même profil de réponse peuvent avoir des estimations du trait latent différentes. L'utilisation des valeurs plausibles rend impossible toute analyse au niveau individuel. Seules les analyses au niveau du groupe sont possibles. Ceci explique la large utilisation des valeurs plausibles dans les grandes enquêtes de l'éducation évaluant l'apprentissage et les performances des élèves avec des nombre d'individus et d'items très grands. Dans les enquêtes, telles que PISA [111] ou NAEP [108], l'intérêt porte seulement sur le groupe et donc sur les paramètres de population.

Les valeurs plausibles sont très peu employées en santé. On peut trouver un exemple d'utilisation des valeurs plausibles en santé dans une étude de simulation menée par Glas et al [43]. Dans cette étude, un modèle IRT longitudinal estimé par MML présente des résultats comparables en terme de risque de première espèce et de puissance qu'un modèle utilisant des valeurs plausibles dans le contexte de deux groupes de patients évalués à deux temps différents. Néanmoins, pour des valeurs petites du nombre d'items (5 ou 10), le modèle IRT longitudinal présente une meilleure puissance que le modèle basé sur les valeurs plausibles. Les résultats de l'étude de simulation de Glas et al sont donc comparables à ceux de cette étude où le nombre d'items valait 4 ou 7.

Les méthodes LRM et SM ont montré des résultats comparables. Le risque de première espèce est maintenu proche de 5%. Comme attendu, la puissance augmente avec la taille d'échantillon, le nombre d'items et la corrélation de la variable latente. Les bons résultats en terme de risque de première espèce et de puissance s'expliquent par la bonne estimation de l'effet temps pour les deux méthodes. En revanche, une légère sur-estimation de la variance et de faibles biais du coefficient de corrélation de la variable latente ont été observés pour la méthode LRM. Les biais des paramètres de variance n'ont pu être évalués pour la méthode SM en raison de la non-connaissance de la vraie valeur des paramètres pour le score. Les méthodes LRM et SM semblent être adéquates pour l'analyses de PRO longitudinales complètes issues d'une échelle validée avec un modèle de Rasch.

Toutefois, la plupart des études sont souvent sujettes à des données manquantes. L'absence de données entraîne une perte d'information et de précision. La méthode choisie pour l'analyse de données incomplètes peut également aboutir à une mauvaise estimation des paramètres si elle n'est pas adaptée au type de données manquantes. En présence de données incomplètes, on peut s'attendre à une perte d'information plus importante pour la méthode SM que la méthode LRM. En effet, pour que le score puisse être calculé l'ensemble des items doit être rempli sinon l'information portée par l'individu même si elle est incomplète n'est pas prise en compte dans l'analyse. On réalise alors une analyse des cas complets. Les manuels des questionnaires les plus utilisés (SF-36, QLQ-C30) comporte des recommandations pour le traitement des données manquantes. Dans ce cas, il est préconisé d'imputer pour remplacer les valeurs manquantes ce qui permet de perdre moins d'individus lors de l'analyse.

Néanmoins, en règle générale, l'imputation n'est possible que si au moins la moitié des items ont été remplis par l'individu. De plus, la technique d'imputation utilisée a également un impact sur les résultats et les estimations obtenues après imputation peuvent être biaisées. Au contraire de la CTT, l'analyse en IRT est une analyse des cas disponibles. Toute l'information portée par les individus même ceux pour lesquels les données sont incomplètes est utilisée dans l'analyse. De plus, le modèle de Rasch possède la propriété d'objectivité spécifique. Si l'on considère que l'estimation du trait latent est indépendante de l'ensemble d'item utilisé pour la mesure alors le trait latent d'un individu n'ayant pas répondu à l'ensemble des items devrait pouvoir être estimé correctement.

Chapitre 5

Comparaison de méthodes d'analyse de données subjectives longitudinales sujettes au dropout

Les données incomplètes sont fréquemment rencontrées dans les études longitudinales, notamment les sorties prématurées de patient de l'étude encore appelées dropout. Il paraît important d'étudier l'impact du dropout sur les méthodes d'analyse de PRO longitudinales en raison de la perte de puissance et des biais qui peuvent être engendrés par l'absence de données. Afin de se placer dans un cadre plus proche de la réalité des données habituellement rencontrées en santé que le cadre de l'étude précédente, une étude de simulation comparant la méthode SM à la méthode LRM dans le cadre de données longitudinales sujettes au dropout issues d'une échelle validée avec un modèle de Rasch a été mise en place. En raison de leurs mauvais résultats sur des données complètes, les performances des méthodes RM et PV n'ont pas été étudiées dans le cadre de données incomplètes.

Cette étude de simulation vise à évaluer les performances des méthodes SM et LRM en terme de risque de première espèce, de puissance et de biais de l'estimation de l'effet temps ainsi qu'à étudier l'impact du type et de la quantité de dropout sur ces critères. En santé, le dropout est souvent considéré comme étant de type MNAR car le processus de sortie d'étude a un fort risque d'être lié au niveau du PRO étudié. On peut considérer qu'un patient dont l'état se dégrade, en raison d'une progression de la maladie ou de l'augmentation des effets secondaires par exemple, et dont le niveau du PRO étudié diminue a plus de risque de sortir

TABLE 5.1 – Paramètres utilisés pour la simulation des données et l'analyse

	Nombre de cas	312
	Nombre de jeux simulés/cas	500
	Taille d'échantillon (N)	100 ; 200
	Nombre d'items (J)	4 ; 7
Variable latente	Effet temps (d_θ)	0 ; 0,2
	Variance (σ^2_θ)	1
	Corrélation (ρ_θ)	0,4 ; 0,7 ; 0,9
	Effet groupe (Δ_θ)	0
Score	Effet temps (d_S)	0 ; 0,15 (J=4) ; 0,25 (J=7)
	Effet groupe (Δ_S)	0
Dropout	Proportion ($\pi^{(t)}$)	0% ; 5% ; 10% ; 20%
	Corrélation variable latente	MCAR : 0
	et propension au dropout ($\rho_{\theta\chi}$)	MNAR : -0,4 ; -0,7 ; -0,9
Analyse	Méthodes comparées	Score and Mixed Models (SM)
		Longitudinal Rasch Mixed model (LRM)
	Structures de matrice	sans contraintes (UN)
		auto-régressive d'ordre 1 (AR(1))
		"compound symmetry" hétérogène (CSH)

de l'étude prématurément que les autres patients [110]. Le dropout de type MNAR a donc été simulé sous l'hypothèse $\rho_{\theta\chi} < 0$ avec $\rho_{\theta\chi}$ la corrélation entre la valeur du trait latent au temps t, $\theta^{(t)}$ et la propension au dropout au temps t, $\chi^{(t)}$. Le dropout MNAR est non-ignorable ou informatif dans cette étude. L'impact de l'informativité du dropout sera étudié en faisant varier la corrélation entre la valeur du trait latent et la propension au dropout ($\rho_{\theta\chi}$) de $-0,4$ (dropout peu informatif) à $-0,9$ (dropout très informatif). Les données simulées avec une corrélation nulle entre la valeur du trait latent et la propension au dropout ($\rho_{\theta\chi} = 0$) permettront d'évaluer les performances des deux méthodes en cas de dropout MCAR qui est ignorable dans le cadre de cette étude. Les données sont supposées complètes au premier temps d'évaluation ($\pi^{(1)} = 0$). Le dropout des patients augmente linéairement et $\pi^{(t)} = 0\%$ (données complètes), 5%, 10% ou 20% des patients restants sortent de l'étude à chaque temps($t = 2, 3$). L'ensemble des paramètres de l'étude sont rappelés dans le tableau 5.1. La combinaison des différents paramètres nous a amené à considérer 312 cas différents. Cette étude a fait l'objet d'une publication dans l'International Journal of Applied Mathematics & Statistics [14] (cf. annexe B p.145).

5.1 Résultats

Les structures de matrice de variance covariance UN, AR(1) et CSH ont été étudiées pour la méthode SM. Pour la méthode SM, l'AIC était minimisé plus souvent dans les modèles avec une structure de variance covariance AR(1) que dans les modèles avec une autre structure pour des valeurs simulées du coefficient de corrélation $\rho_\theta = 0,4$ et $\rho_\theta = 0,7$. Lorsque $\rho_\theta = 0,9$, l'AIC était plus souvent minimisé par les modèles ayant une structure de variance covariance CSH que par les modèles ayant une autre structure. Les résultats présentés ci-dessous ont été obtenus avec une structure AR(1). La méthode LRM estime une matrice de type UN.

5.1.1 Risque de première espèce et puissance

Le tableau 5.2 présente lee estimations du risque de première espèce pour chaque méthode étudiée et pour différentes valeurs de la taille d'échantillon, du nombre d'items, de la corrélation de la variable latente, de la proportion et du type de dropout. Les deux méthodes ont des résultats comparables en terme de risque de première espèce. La variation de la taille d'échantillon, du nombre d'items et du coefficient de corrélation ne semble pas avoir d'impact sur le risque de première espèce.

Comme observé dans l'étude précédente, le risque de première espèce lorsque les données sont complètes est maintenu proche de 5%. Il varie de 4,4% à 6,6% et de 3,6% à 6,0% pour LRM et SM respectivement. Quelle que soit la quantité de dropout, les données sujettes à un dropout de type MCAR présentent un risque de première espèce proche de celui des données complètes. Pour les données sujettes à un dropout de type MNAR, le risque de première espèce augmente avec la proportion de dropout. Si, de plus la proportion de dropout est grande ($\pi^{(t)} = 20\%$), le risque de première espèce augmente avec l'informativité (corrélation entre la variable latente et la propension au dropout). Ainsi, dans les pires cas où $\pi^{(t)} = 20\%$ et $\rho_{\theta\chi} = -0.9$, le risque de première espèce peut atteindre 10% pour les deux méthodes. Dans le cas MNAR, certains intervalles de confiance à 95% du risque de première espèce ne contiennent pas la valeur attendue 5%. Ce nombre d'intervalles augmente avec l'informativité du dropout.

TABLE 5.2 – Risque de première espèce pour les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J), de la corrélation de la variable latente (ρ_θ), de la proportion de dropout ($\pi^{(t)}$) et du type de dropout (ρ_{θ_χ}). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour la méthode SM et une structure UN pour la méthode LRM.

				no dropout		MCAR		MNAR					
								$\rho_{\theta\chi} = -0.4$		$\rho_{\theta\chi} = -0.7$		$\rho_{\theta\chi} = -0.9$	
N	J	ρ_θ	$\pi^{(t)}$	LRM	SM	LRM	SM	LRM	SM	LRM	SM	LRM	SM
100	4	0,4	0	0,054	0,054								
			0,05			0,054	0,050	0,066	0,066	0,060	0,054	0,038	0,038
			0,1			0,046	0,042	0,050	0,046	0,048	0,052	0,054	0,054
			0,2			0,050	0,048	0,072*	0,064	0,080*	0,078*	0,070	0,066
		0,7	0	0,044	0,040								
			0,05			0,048	0,042	0,040	0,034	0,042	0,044	0,036	0,034
			0,1			0,048	0,038	0,044	0,044	0,034	0,032	0,058	0,052
			0,2			0,068	0,060	0,052	0,060	0,052	0,046	0,076*	0,080*
		0,9	0	0,048	0,038								
			0,05			0,048	0,048	0,032	0,034	0,048	0,050	0,050	0,054
			0,1			0,040	0,046	0,024*	0,030*	0,032	0,038	0,034	0,036
			0,2			0,048	0,046	0,034	0,048	0,050	0,046	0,044	0,038
	7	0,4	0	0,044	0,036								
			0,05			0,054	0,048	0,066	0,058	0,048	0,050	0,040	0,038
			0,1			0,060	0,058	0,054	0,058	0,068	0,064	0,072*	0,066
			0,2			0,046	0,046	0,054	0,046	0,062	0,062	0,088*	0,080*
		0,7	0	0,058	0,046								
			0,05			0,056	0,046	0,038	0,038	0,054	0,052	0,056	0,042
			0,1			0,054	0,040	0,052	0,046	0,050	0,042	0,056	0,036
			0,2			0,052	0,054	0,064	0,044	0,090*	0,084*	0,090*	0,072*
		0,9	0	0,046	0,044								
			0,05			0,049	0,050	0,052	0,038	0,056	0,046	0,036	0,036
			0,1			0,038	0,034	0,046	0,046	0,048	0,040	0,064	0,060
			0,2			0,034	0,034	0,064	0,050	0,054	0,054	0,054	0,068
200	4	0,4	0	0,052	0,060								
			0,05			0,054	0,052	0,040	0,046	0,062	0,060	0,048	0,052
			0,1			0,052	0,044	0,066	0,058	0,080*	0,072*	0,046	0,048
			0,2			0,046	0,050	0,052	0,052	0,074*	0,076*	0,106*	0,104*
		0,7	0	0,056	0,048								
			0,05			0,042	0,038	0,040	0,042	0,080*	0,072*	0,050	0,044
			0,1			0,054	0,050	0,044	0,048	0,062	0,058	0,086*	0,078*
			0,2			0,038	0,040	0,044	0,044	0,050	0,056	0,088*	0,086*
		0,9	0	0,066	0,060								
			0,05			0,044	0,046	0,052	0,044	0,050	0,048	0,032	0,030*
			0,1			0,057	0,050	0,044	0,042	0,057	0,058	0,060	0,052
			0,2			0,046	0,046	0,063	0,052	0,069	0,070	0,074*	0,082*
	7	0,4	0	0,050	0,050								
			0,05			0,056	0,050	0,052	0,052	0,064	0,058	0,064	0,062
			0,1			0,042	0,038	0,064	0,054	0,066	0,052	0,058	0,052
			0,2			0,066	0,070	0,080*	0,072*	0,058	0,054	0,106*	0,100*
		0,7	0	0,048	0,036								
			0,05			0,054	0,054	0,066	0,052	0,046	0,042	0,068	0,066
			0,1			0,056	0,038	0,060	0,046	0,058	0,054	0,064	0,056
			0,2			0,052	0,032	0,084*	0,074*	0,086*	0,084*	0,080*	0,086*
		0,9	0	0,058	0,048								
			0,05			0,056	0,050	0,054	0,046	0,059	0,058	0,060	0,060
			0,1			0,044	0,028*	0,050	0,054	0,038	0,036	0,048	0,044
			0,2			0,053	0,046	0,059	0,040	0,088*	0,080*	0,103*	0,104*

* l'intervalle de confiance à 95% ne contient pas la valeur attendue 5%.

TABLE 5.3 – Puissance pour les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J), de la corrélation de la variable latente (ρ_θ), de la proportion de dropout ($\pi^{(t)}$) et du type de dropout (ρ_{θ_χ}). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour la méthode SM et une structure UN pour la méthode LRM.

				no dropout		MCAR		MNAR					
								$\rho_{\theta\chi} = -0.4$		$\rho_{\theta\chi} = -0.7$		$\rho_{\theta\chi} = -0.9$	
N	J	ρ_θ	$\pi^{(t)}$	LRM	SM	LRM	SM	LRM	SM	LRM	SM	LRM	SM
100	4	0,4	0	0,388	0,384								
			0,05			0,399	0,400	0,356	0,354	0,383$^\sharp$	0,360	0,404	0,404
			0,1			0,335	0,320	0,411	0,408	0,439	0,426	0,440	0,426
			0,2			0,325	0,316	0,384	0,376	0,474	0,464	0,513	0,508
		0,7	0	0,423$^\sharp$	0,372								
			0,05			0,439$^\sharp$	0,382	0,426$^\sharp$	0,390	0,457$^\sharp$	0,400	0,498$^\sharp$	0,448
			0,1			0,441$^\sharp$	0,406	0,453$^\sharp$	0,400	0,480	0,464	0,499$^\sharp$	0,466
			0,2			0,399$^\sharp$	0,368	0,399$^\sharp$	0,378	0,486$^\sharp$	0,466	0,543	0,524
		0,9	0	0,508$^\sharp$	0,424								
			0,05			0,471$^\sharp$	0,390	0,505$^\sharp$	0,432	0,511$^\sharp$	0,432	0,560$^\sharp$	0,474
			0,1			0,462$^\sharp$	0,398	0,427	0,366	0,511$^\sharp$	0,436	0,508$^\sharp$	0,462
			0,2			0,379$^\sharp$	0,342	0,477$^\sharp$	0,434	0,509	0,496	0,598$^\sharp$	0,578
	7	0,4	0	0,470$^\sharp$	0,428								
			0,05			0,454$^\sharp$	0,424	0,488$^\sharp$	0,468	0,504	0,492	0,518$^\sharp$	0,496
			0,1			0,482$^\sharp$	0,454	0,522$^\sharp$	0,494	0,510$^\sharp$	0,496	0,556	0,546
			0,2			0,398	0,396	0,500	0,482	0,534$^\sharp$	0,514	0,600	0,588
		0,7	0	0,568$^\sharp$	0,522								
			0,05			0,558$^\sharp$	0,502	0,586$^\sharp$	0,548	0,598$^\sharp$	0,534	0,626$^\sharp$	0,576
			0,1			0,541$^\sharp$	0,470	0,549$^\sharp$	0,510	0,623$^\sharp$	0,574	0,607$^\sharp$	0,562
			0,2			0,446$^\sharp$	0,414	0,568$^\sharp$	0,518	0,655	0,632	0,685	0,658
		0,9	0	0,688$^\sharp$	0,564								
			0,05			0,670$^\sharp$	0,548	0,679$^\sharp$	0,604	0,723$^\sharp$	0,622	0,725$^\sharp$	0,620
			0,1			0,635$^\sharp$	0,526	0,686$^\sharp$	0,584	0,743$^\sharp$	0,654	0,719$^\sharp$	0,632
			0,2			0,531$^\sharp$	0,452	0,687$^\sharp$	0,590	0,745$^\sharp$	0,674	0,737$^\sharp$	0,684
200	4	0,4	0	0,654$^\sharp$	0,634								
			0,05			0,654	0,644	0,682	0,668	0,718$^\sharp$	0,690	0,678$^\sharp$	0,658
			0,1			0,631	0,628	0,658	0,646	0,738	0,726	0,718	0,714
			0,2			0,552	0,540	0,694	0,688	0,745	0,736	0,820	0,818
		0,7	0	0,721$^\sharp$	0,678								
			0,05			0,729$^\sharp$	0,694	0,747$^\sharp$	0,700	0,787	0,742	0,753$^\sharp$	0,702
			0,1			0,701$^\sharp$	0,670	0,752$^\sharp$	0,724	0,806$^\sharp$	0,772	0,825	0,786
			0,2			0,669$^\sharp$	0,626	0,745$^\sharp$	0,710	0,843$^\sharp$	0,830	0,820	0,808
		0,9	0	0,826$^\sharp$	0,756								
			0,05			0,765$^\sharp$	0,686	0,807$^\sharp$	0,736	0,809$^\sharp$	0,764	0,823$^\sharp$	0,744
			0,1			0,768$^\sharp$	0,704	0,797$^\sharp$	0,740	0,857$^\sharp$	0,784	0,847$^\sharp$	0,806
			0,2			0,698$^\sharp$	0,616	0,781$^\sharp$	0,744	0,861$^\sharp$	0,846	0,875	0,872
	7	0,4	0	0,822$^\sharp$	0,810								
			0,05			0,758	0,750	0,778	0,778	0,826$^\sharp$	0,812	0,798$^\sharp$	0,782
			0,1			0,778$^\sharp$	0,760	0,794	0,788	0,826	0,826	0,846	0,832
			0,2			0,696	0,698	0,812	0,786	0,874	0,870	0,896	0,894
		0,7	0	0,916$^\sharp$	0,880								
			0,05			0,858$^\sharp$	0,826	0,882$^\sharp$	0,846	0,892$^\sharp$	0,860	0,880$^\sharp$	0,850
			0,1			0,814$^\sharp$	0,786	0,894$^\sharp$	0,878	0,904$^\sharp$	0,870	0,936$^\sharp$	0,910
			0,2			0,776$^\sharp$	0,732	0,876	0,856	0,900	0,890	0,958	0,944
		0,9	0	0,955$^\sharp$	0,918								
			0,05			0,935$^\sharp$	0,882	0,966$^\sharp$	0,932	0,946$^\sharp$	0,910	0,938$^\sharp$	0,912
			0,1			0,931	0,876	0,956$^\sharp$	0,932	0,946$^\sharp$	0,932	0,962$^\sharp$	0,942
			0,2			0,893$^\sharp$	0,810	0,922$^\sharp$	0,900	0,956	0,942	0,960	0,950

$^\sharp$ test de McNemar significatif à 5%

H_0 : puissance LRM=puissance SM vs H_1 : puissance LRM>puissance SM

en gris : l'intervalle de confiance à 95% du risque de première espèce ne contient pas la valeur attendue 5%

Le tableau 5.3 présente les estimations de la puissance pour chaque méthode étudiée et pour différentes valeurs de la taille d'échantillon, du nombre d'items, de la corrélation de la variable latente, de la proportion et du type de dropout. Lorsque la corrélation entre les mesures de la variable latente est faible ($\rho_\theta = 0,4$), les puissances obtenues pour chaque méthode sont proches avec une puissance pour LRM légèrement supérieure à celle de SM. Pour des valeurs de corrélation plus élevées ($\rho_\theta = 0,7$ ou $\rho_\theta = 0,9$), la puissance de LRM semble être systématiquement légèrement supérieure à celle de SM. Les tests de McNemar correspondant sont tous significatifs à 5% concluant à la supériorité de la puissance de LRM. La puissance augmente avec la taille d'échantillon, le nombre d'items et le coefficient de corrélation.

Les données sujettes à un dropout de type MCAR présentent une puissance moins élevée que les données complètes. Cette perte de puissance est la plus élevée quand la proportion de dropout est élevée ($\pi^{(t)} = 20\%$). Les données sujettes à un dropout de type MNAR présentent une puissance plus élevée que les données complètes. La puissance des données avec dropout MNAR devient de plus en plus éloignée de la puissance des données complètes à mesure que la proportion de dropout et l'informativité augmentent.

5.1.2 Estimation de l'effet temps

La figure 5.1 présente l'estimation de l'effet temps entre le premier et le deuxième temps de mesure pour chaque méthode étudiée et pour différentes valeurs de la taille d'échantillon, du nombre d'items et de la corrélation de la variable latente lorsque l'effet temps simulé vaut $d_\theta = 0$. Les résultats présentés proviennent de données complètes ($\pi^{(t)} = 0\%$), sujettes à 20% de dropout MCAR ou sujettes à 20% de dropout MNAR ($\rho_{\theta_\chi} = -0,9$). Les résultats pour l'ensemble des valeurs des paramètres de simulation sont présentés en annexe dans le tableau C.1 p.166. Les résultats en terme de biais sont comparables d'une méthode à l'autre. Comme pour les données complètes, la plupart des estimations de l'effet temps sont proches de 0 et donc non-biaisées lorsque les données sont sujettes à du dropout de type MCAR. Seuls deux cas présentent une sous-estimation de l'effet temps pour les deux méthodes, lorsque $N = 200$, $J = 7$, $\pi_{(t)} = 20\%$, $\rho_{\theta\chi} = 0$ (MCAR) et que la corrélation de la variable latente vaut $\rho_\theta = 0,4$ dans le premier cas ou vaut $\rho_\theta = 0,9$ dans le deuxième cas.

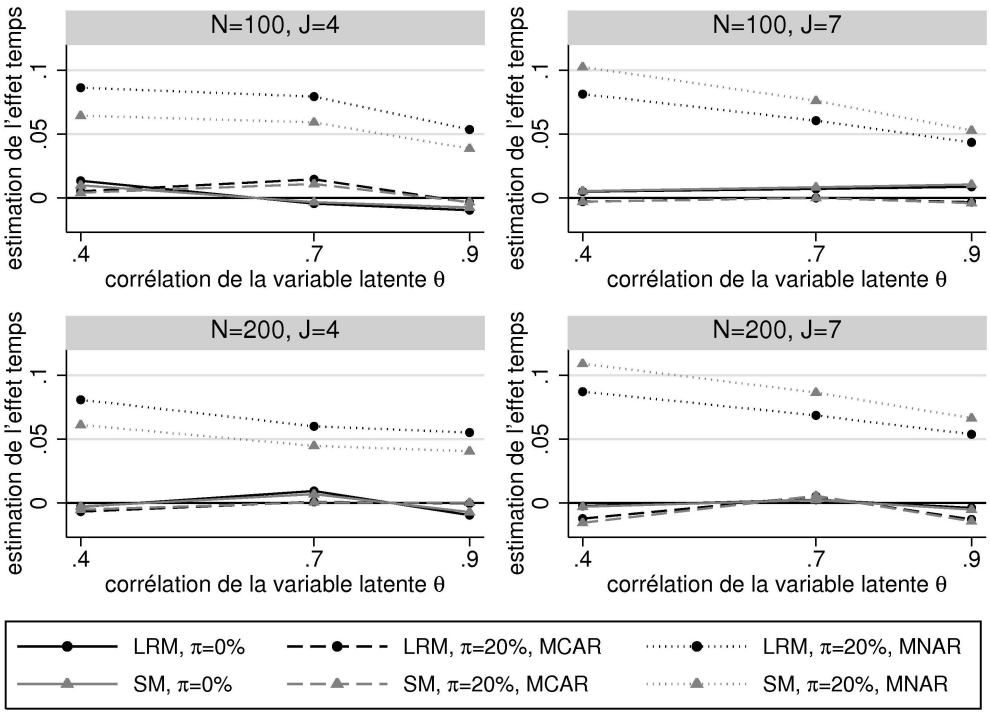

FIGURE 5.1 – Estimations de l'effet temps entre le temps 1 et le temps 2 pour les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J) et de la corrélation de la variable latente (ρ_θ). Données complètes ($\pi^{(t)} = 0\%$), sujettes à 20% de dropout MCAR ou sujettes à 20% de dropout MNAR ($\rho_{\theta_\chi} = -0,9$). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour la méthode SM et une structure UN pour la méthode LRM. Effet temps simulé : $d_\theta = 0$.

Pour les données sujettes à un dropout de type MNAR, les méthodes SM et LRM sur-estiment l'effet temps. Le biais diminue lorsque la corrélation de la variable latente ρ_θ augmente. Le biais augmente avec le taux de dropout. Le nombre de cas biaisés augmente avec l'informativité du dropout (quand la corrélation entre la variable latente et la propension au dropout diminue). Lorsque l'informativité est faible ($\rho_{\theta\chi} = -0.4$), les cas où le taux de dropout est élevé ($\pi^{(t)} = 20\%$) sont tous biaisés. Lorsque l'informativité est plus élevée ($\rho_{\theta\chi} = -0.7$ ou $\rho_{\theta\chi} = -0.9$), l'effet temps est systématiquement sur-estimé lorsque le taux de dropout est supérieur ou égal à 10%.

La figure 5.2 présente l'estimation de l'effet temps entre le premier et le deuxième temps de mesure pour chaque méthode étudiée et pour différentes valeurs de la taille d'échantillon, du nombre d'items et de la corrélation de la variable latente lorsque l'effet temps simulé vaut $d_\theta = 0,2$. L'effet temps simulé est approximé à $d_S = 0,15$ pour $J = 4$ et $d_S = 0,25$ pour $J = 7$. Les résultats présentés proviennent de données complètes ($\pi^{(t)} = 0\%$), sujettes à 20% de dropout MCAR ou sujettes à 20% de dropout MNAR ($\rho_{\theta_\chi} = -0,9$). Les résultats pour

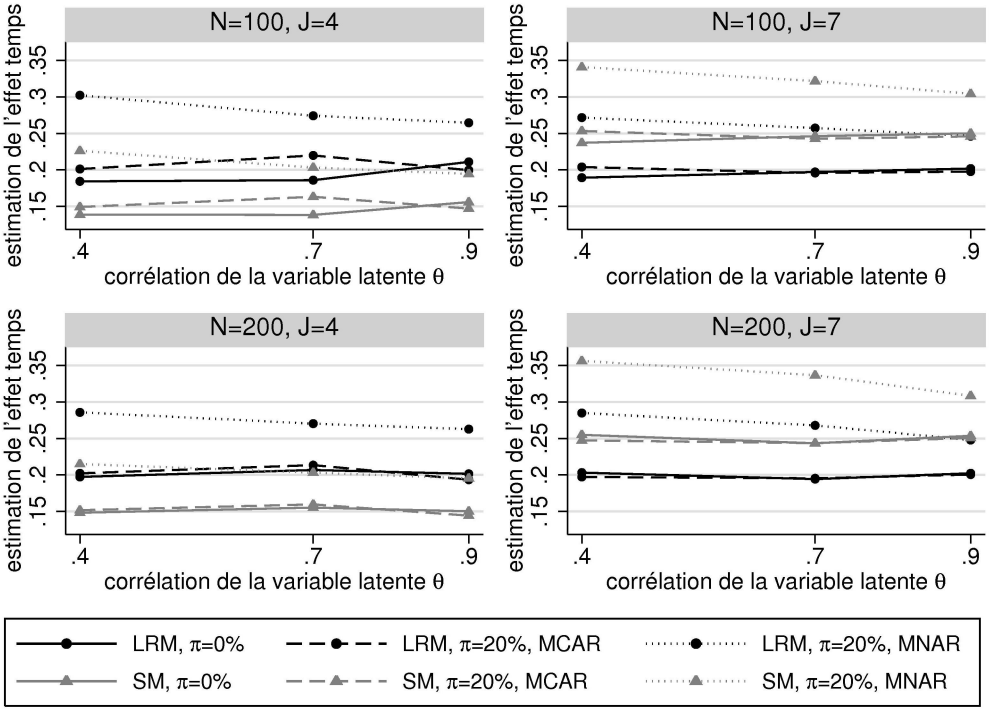

FIGURE 5.2 – Estimations de l'effet temps entre le temps 1 et le temps 2 pour les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J) et de la corrélation de la variable latente (ρ_θ). Données complètes ($\pi^{(t)} = 0\%$), sujettes à 20% de dropout MCAR ou sujettes à 20% de dropout MNAR ($\rho_{\theta_\chi} = -0,9$). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour la méthode SM et une structure UN pour la méthode LRM. Effet temps simulé : $d_\theta = 0,2$, $d_S = 0,15$ pour $J = 4$ et $d_S = 0,25$ pour $J = 7$.

l'ensemble des valeurs des paramètres de simulation sont présentés en annexe dans le tableau C.2 p.168. Les mêmes tendances que pour les données avec un effet temps simulé à $d_\theta = 0$ peuvent être observées. Les résultats de SM et LRM sont comparables. Les estimations de l'effet temps sont non-biaisées pour les données complètes et les données sujettes à un dropout de type MCAR. Dans deux cas, les effets temps sont sur-estimés pour les deux méthodes, lorsque $J = 4$, $\rho_\theta = 0,7$, $\pi^{(t)} = 20\%$, $\rho_{\theta\chi} = 0$ (MCAR) et $N = 100$ ou $N = 200$.

Pour les données sujettes à du dropout de type MNAR, l'effet temps est sur-estimé dans de nombreux cas. Le biais diminue lorsque la corrélation de la variable latente augmente et le biais augmente avec l'informativité et le taux de dropout. Le nombre de cas biaisés augmente avec l'informativité du dropout.

5.2 Discussion

Les deux méthodes SM et LRM ont montré des résultats comparables en terme de risque de première espèce. Les puissances de la méthode LRM semblent être légèrement mais systématiquement significativement plus hautes que les puissances de la méthode SM lorsque le coefficient de corrélation de la variable latente vaut $\rho_\theta = 0,7$ ou $\rho_\theta = 0,9$.
Les deux méthodes étudiées sont basées sur la vraisemblance. Dans ce cadre, des données manquantes de type MCAR sont ignorables et les analyses de cas disponibles telles les méthodes SM et LRM sont supposées valides. Comme attendue, les données sujettes à un dropout de type MCAR ont montré des résultats proches des données complètes avec des risques de premières espèce proches de 5% ainsi qu'une perte de puissance. Néanmoins, un léger biais de l'estimation de l'effet temps a été observé dans de rares cas.
La survenue de dropout de type MNAR (données manquantes non-ignorables) a eu un impact négatif sur les performances des deux méthodes. Le risque de première espèce n'est plus maintenu à 5% et peut atteindre 10% dans les cas les plus défavorables où la quantité et l'informativité du dropout sont élevées. La puissance observée pour les données avec dropout MNAR est plus élevée que la puissance des données complètes. La sur-estimation de l'effet temps augmente avec la quantité et l'informativité du dropout. Ceci explique que la puissance augmente également avec la quantité et l'informativité du dropout. Les méthodes SM et LRM semblent être peu adaptées à l'analyse de données sujettes au dropout MNAR. Cette étude confirme que les données manquantes informatives ne peuvent pas être ignorées dans le cadre d'une analyse basée sur la vraisemblance. Pour l'analyse de PRO longitudinales sujettes à du dropout de type MNAR, il est nécessaire d'envisager l'utilisation de modèles de mélange de schémas d'observation ou de modèles de sélection qui permettent la modélisation conjointe du modèle de mesure et du processus de données manquantes.

Les méthodes SM et LRM semblent se comporter de façon comparable face aux données manquantes. Or, dans cette étude, seul l'impact du dropout a été étudié. Dans le cas du dropout, les réponses de l'individu à l'ensemble des items sont manquantes à partir d'un temps donné. Ainsi, toute l'information du patient est perdue à partir de la sortie d'étude autant pour la méthode SM que pour la méthode LRM. Avec la propriété d'objectivité spécifique du modèle de Rasch, on considère que l'estimation du trait latent est indépendante de l'ensemble d'item utilisé pour la mesure. Ainsi le trait latent d'un individu n'ayant répondu qu'à une partie des items à un temps donné devrait pouvoir être estimé correctement. Alors

que le score de l'individu à chaque temps ne peut être calculé si l'individu n'a pas répondu à assez d'items. Dans le cas du dropout, les données manquantes ont le même impact sur les deux méthodes car il n'y a pas d'information partielle pour le patient : soit l'information est complète aux temps où le patient a été mesuré soit il n'y a aucune information pour le patient lorsque celui-ci est sorti de l'étude. Dans ce cadre, la propriété d'objectivité spécifique du modèle de Rasch ne permet pas de prendre en compte plus d'information dans l'analyse avec la méthode LRM qu'avec la méthode SM.

Chapitre 6

Comparaison de méthodes d'analyse des effets temps et groupe pour données subjectives longitudinales sujettes au dropout

Dans le traitement et le suivi des patients, les études évaluant des Patient-Reported Outcomes ont souvent pour but d'étudier l'évolution d'un PRO au cours du temps dans une population donnée. Il est également fréquent que les études visent à comparer l'évolution d'un PRO au cours du temps dans deux ou plusieurs groupes distincts par une caractéristique clinique comme le traitement reçu ou par une caractéristique socio-démographique. C'est le cas, par exemple, dans l'étude de Durna et al. [27] qui compare la qualité de vie de survivantes de cancer du sein ayant reçu ou non une hormonothérapie substitutive. Dans ce type d'études, outre l'intérêt dans l'évolution du critère au cours du temps, la détection de la présence ou non d'un effet groupe et sa quantification sont également importantes. Les performances des méthodes d'analyse de PRO longitudinales ont également été évaluées dans ce cadre à travers une étude de simulation.

L'étude de simulation suivante vise à évaluer les performances des méthodes SM et LRM dans le cadre de données subjectives longitudinales sujettes au dropout recueillies auprès de deux groupes de patients. Elle a été menée de la même manière que les études de simulation précédentes. Lorsqu'un effet groupe a été simulé, il vaut $\Delta_\theta = 0,5$ pour la variable latente

et a été approximé à $\Delta_S = 0,38$ lorsque $J = 4$ et $\Delta_S = 0,63$ lorsque $J = 7$ pour le score (cf. equations 3.8 et 3.9 p.60). L'ensemble des paramètres de l'étude sont rappelés dans le tableau 6.1. La combinaison des différents paramètres nous a amené à considérer 624 cas différents.

TABLE 6.1 – Paramètres utilisés pour la simulation des données et l'analyse

	Nombre de cas	624
	Nombre de jeux simulés/cas	500
	Taille d'échantillon (N)	100 ; 200
	Nombre d'items (J)	4 ; 7
Variable latente	Effet temps (d_θ)	0 ; 0,2
	Variance (σ^2_θ)	1
	Corrélation (ρ_θ)	0,4 ; 0,7 ; 0,9
	Effet groupe (Δ_θ)	0 ; 0,5
Score	Effet temps (d_S)	0 ; 0,15 (J=4) ; 0,25 (J=7)
	Effet groupe (Δ_S)	0 ; 0,38 (J=4) ; 0,63 (J=7)
Dropout	Proportion ($\pi^{(t)}$)	0% ; 5% ; 10% ; 20%
	Corrélation variable latente	MCAR : 0
	et propension au dropout ($\rho_{\theta\chi}$)	MNAR : -0,4 ; -0,7 ; -0,9
Analyse	Méthodes comparées	Score and Mixed Models (SM)
		Longitudinal Rasch Mixed model (LRM)
	Structures de matrice	sans contraintes (UN)
		auto-régressive d'ordre 1 (AR(1))
		"compound symmetry" hétérogène (CSH)

6.1 Résultats

Les structures de matrice de variance covariance UN, AR(1) et CSH ont été étudiées pour la méthode SM. Pour la méthode SM, l'AIC était minimisé plus souvent dans les modèles avec une structure de variance covariance AR(1) que dans les modèles avec une autre structure pour des valeurs simulées du coefficient de corrélation $\rho_\theta = 0,4$ et $\rho_\theta = 0,7$. Lorsque $\rho_\theta = 0,9$, l'AIC était plus souvent minimisé par les modèles ayant une structure de variance covariance CSH que par les modèles ayant une autre structure. Les résultats présentés ci-dessous ont été obtenus avec une structure AR(1). La méthode LRM estime une matrice de structure UN.

6.1.1 Risque de première espèce et puissance de l'effet groupe

TABLE 6.2 – Risque de première espèce de l'effet groupe pour les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J), de la corrélation de la variable latente (ρ_θ) et du type de dropout (ρ_{θ_χ}). Structure de matrice de variance covariance AR(1) pour SM et structure UN pour LRM. Données simulées avec ($d_\theta = 0.2$) ou sans effet temps ($d_\theta = 0$). Proportion de dropout : $\pi^{(t)} = 20\%$.

				no dropout		MCAR		MNAR			
								$\rho_{\theta\chi} = -0.4$		$\rho_{\theta\chi} = -0.9$	
d_θ	N	J	ρ_θ	LRM	SM	LRM	SM	LRM	SM	LRM	SM
0	100	4	0,4	0,050	0,040	0,066	0,068	0,070	0,060	0,050	0,052
			0,9	0,058	0,064	0,070	0,072[!]	0,072[!]	0,072[!]	0,034	0,044
		7	0,4	0,068	0,064	0,066	0,052	0,064	0,060	0,072[!]	0,068
			0,9	0,060	0,054	0,076[!]	0,054	0,070	0,052	0,074[!]	0,064
	200	4	0,4	0,048	0,046	0,048	0,046	0,044	0,050	0,040	0,042
			0,9	0,063	0,062	0,054	0,064	0,055	0,066	0,062	0,076[!]
		7	0,4	0,054	0,048	0,068	0,068	0,044	0,042	0,032	0,042
			0,9	0,044	0,054	0,053	0,044	0,077[!]	0,076[!]	0,060	0,052
0,2	100	4	0,4	0,060	0,058	0,074[!]	0,056	0,060	0,062	0,060	0,052
			0,9	0,053	0,054	0,062	0,060	0,048	0,056	0,066	0,070
		7	0,4	0,048	0,046	0,066	0,056	0,048	0,042	0,062	0,068
			0,9	0,070	0,062	0,066	0,056	0,062	0,052	0,062	0,070
	200	4	0,4	0,052	0,048	0,058	0,056	0,064	0,052	0,054	0,062
			0,9	0,057	0,064	0,043	0,070	0,060	0,074[!]	0,054	0,064
		7	0,4	0,056	0,056	0,036	0,028[!]	0,044	0,040	0,046	0,052
			0,9	0,048	0,058	0,067	0,056	0,054	0,062	0,046	0,058

[!] l'intervalle de confiance à 95% ne contient pas la valeur attendue 5%.

Le tableau 6.2 présente lee estimations du risque de première espèce de l'effet groupe pour chaque méthode étudiée et pour certaines valeurs de la taille d'échantillon, du nombre d'items, de la corrélation de la variable latente et du type de dropout. Les résultats ont été obtenus à partir de données simulées sans ($d_\theta = 0$) ou avec un effet temps ($d_\theta = 0.2$). Les résultats complets pour l'ensemble des valeurs de paramètres simulées sont présentés en annexe dans les tableaux D.1 p.172 et D.2 p.173. Les deux méthodes ont des résultats comparables en terme de risque de première espèce, maintenus proches de la valeur attendue 5% dans la plupart des cas. Le risque de première espèce peut atteindre 10,2% et 9,8% pour LRM et SM respectivement lorsque $N = 100$, $J = 4$, $\rho_\theta = 0,9$, $\pi^{(t)} = 10\%$, $\rho_{\theta\chi} = -0.9$ et que l'effet temps simulé vaut $d_\theta = 0.2$.

La variation de la taille d'échantillon (N), du nombre d'items (J) et du coefficient de corrélation (ρ_θ) ne semble pas avoir d'impact sur le risque de première espèce. Les données

sujettes au dropout montrent des résultats similaires aux données complètes. Le taux et le type de dropout ne semblent pas avoir d'impact sur le risque de première espèce. Celui-ci est très légèrement supérieur lorsque les données sont sujettes au dropout que lorsque les données sont complètes. Les résultats sont comparables qu'un effet temps ait été simulé ou non.

TABLE 6.3 – Puissance de l'effet groupe pour les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J), de la corrélation de la variable latente (ρ_θ) et du type de dropout (ρ_{θ_χ}). Structure de matrice de variance covariance AR(1) pour SM et structure UN pour LRM. Données simulées avec ($d_\theta = 0.2$) ou sans effet temps ($d_\theta = 0$). Proportion de dropout : $\pi^{(t)} = 20\%$.

				no dropout		MCAR		MNAR			
								$\rho_{\theta\chi} = -0.4$		$\rho_{\theta\chi} = -0.9$	
d_θ	N	J	ρ_θ	LRM	SM	LRM	SM	LRM	SM	LRM	SM
0	100	4	0,4	0,729	0,734	0,636	0,626	0,609	0,622	0,606	0,604
			0,9	0,579	0,598	0,540	0,562	0,504	0,524	0,499	0,532
		7	0,4	0,780	0,782	0,750	0,746	0,731	0,720	0,708	0,702
			0,9	0,623	0,658	0,614	0,622	0,551	0,576	0,599	0,634
	200	4	0,4	0,960	0,972	0,904	0,912	0,872	0,882	0,908	0,910
			0,9	0,887	0,898	0,842	0,858	0,777	0,802	0,806	0,824
		7	0,4	0,968	0,976	0,960	0,960	0,929	0,938	0,956	0,948
			0,9	0,897	0,912	0,869	0,880	0,859	0,866	0,867	0,894
0,2	100	4	0,4	0,718	0,722	0,624	0,634	0,633	0,636	0,624	0,642
			0,9	0,582	0,610	0,525	0,534	0,554	0,576	0,570	0,584
		7	0,4	0,784	0,792	0,722	0,718	0,716	0,718	0,754	0,742
			0,9	0,610	0,628	0,600	0,608	0,582	0,592	0,640	0,638
	200	4	0,4	0,908	0,916	0,916	0,918	0,911	0,920	0,888	0,896
			0,9	0,851	0,876	0,815	0,816	0,809	0,840	0,810	0,818
		7	0,4	0,974	0,980	0,952	0,952	0,935	0,944	0,944	0,946
			0,9	0,911	0,926	0,844	0,880	0,860	0,892	0,861	0,880

en gris : l'intervalle de confiance à 95% du risque de première espèce ne contient pas la valeur attendue 5%

Le tableau 6.3 présente les estimations de la puissance de l'effet groupe pour chaque méthode étudiée et pour certaines valeurs de la taille d'échantillon, du nombre d'items, de la corrélation de la variable latente et du type de dropout. Les résultats ont été obtenus à partir de données simulées sans ($d_\theta = 0$) ou avec un effet temps ($d_\theta = 0.2$). Les résultats complets pour l'ensemble des valeurs de paramètres simulées sont présentés en annexe dans les tableaux D.3 p.174 et D.4 p.175. Les puissances pour la méthode SM semblent légèrement supérieures aux puissances pour la méthode LRM dans la plupart des cas. Comme attendu, la puissance des deux méthodes augmente avec la taille d'échantillon et le nombre d'items.

Au contraire, la puissance décroît lorsque le coefficient de corrélation de la variable latente ρ_θ augmente.

La puissance pour les données complètes semble être généralement supérieure à la puissance des données sujettes au dropout quelle que soit la méthode d'analyse utilisée. Comme attendu, le dropout semble entraîner une perte de puissance. Les résultats sont comparables qu'un effet temps ait été simulé ou non.

6.1.2 Estimation de l'effet groupe

Le tableau 6.4 présente l'estimation de l'effet groupe et de son erreur standard pour chaque méthode étudiée et pour certaines valeurs de la taille d'échantillon, du nombre d'items, de la corrélation de la variable latente et du type de dropout. Les résultats ont été obtenus à partir de données simulées sans ($d_\theta = 0$) ou avec un effet temps ($d_\theta = 0.2$) et sans ($\Delta_\theta = 0$) ou avec ($\Delta_\theta = 0,5$) effet groupe. Les résultats complets pour l'ensemble des valeurs de paramètres simulées sont présentés en annexe dans les tableaux D.5 p.176, D.6 p.178, D.7 p.180 et D.8 p.182. La plupart des estimations de l'effet groupe sont non-biaisées. Lorsque l'effet temps simulé est nul, la plupart des cas biaisés l'étaient pour les deux méthodes.

En revanche, l'estimation de l'effet groupe a tendance à être plus souvent biaisée pour la méthode SM que pour la méthode LRM lorsque l'effet temps simulé vaut $d_\theta = 0.2$. Sous cette hypothèse, la méthode SM semble sous-estimer l'effet groupe alors que la méthode LRM semble le sur-estimer. L'augmentation de la taille d'échantillon, du nombre d'items et du coefficient de corrélation de la variable latente ne semble pas avoir d'impact sur les estimations de l'effet groupe.

TABLE 6.4 – Estimations de l'effet groupe ($\hat{\beta}_{gp}$) et de son erreur standard (s.e.) pour les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J), de la corrélation de la variable latente (ρ_θ) et du type de dropout (ρ_{θ_χ}). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour la méthode SM et une structure UN pour la méthode LRM. Données simulées avec ($d_\theta = 0, 2$) ou sans effet temps ($d_\theta = 0$) et avec ($\Delta_\theta = 0, 5$) ou sans ($\Delta_\theta = 0$) effet groupe. Proportion de dropout : $\pi^{(t)} = 20\%$. $\Delta_\theta = 0 : \Delta_S = 0$. $\Delta_\theta = 0, 5 : \Delta_S = 0.38$ pour J=4, $\Delta_S = 0.63$ pour J=7.

					no dropout				MCAR				MNAR							
													$\rho_{\theta\chi} = -0.4$				$\rho_{\theta\chi} = -0.9$			
					LRM		SM		LRM		SM		LRM		SM		LRM		SM	
Δ_θ	d_θ	N	J	ρ_θ	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)
0	0	100	4	0,4	-0,021*	(0,200)	-0,015*	(0,149)	-0,015	(0,220)	-0,010	(0,164)	-0,014	(0,216)	-0,011	(0,162)	0,019	(0,218)	0,013	(0,163)
				0,9	0,008	(0,237)	0,006	(0,169)	-0,020	(0,251)	-0,010	(0,181)	-0,004	(0,249)	-0,004	(0,180)	-0,007	(0,252)	-0,008	(0,181)
			7	0,4	-0,010	(0,181)	-0,010	(0,225)	0,013	(0,195)	0,015	(0,244)	0,022*	(0,195)	0,031*	(0,245)	0,001	(0,194)	-0,002	(0,244)
				0,9	-0,018	(0,218)	-0,021	(0,263)	-0,017	(0,228)	-0,009	(0,278)	-0,001	(0,228)	-0,002	(0,279)	-0,024*	(0,228)	-0,028*	(0,278)
		200	4	0,4	0,002	(0,141)	0,001	(0,105)	0,007	(0,153)	0,005	(0,115)	0,006	(0,154)	0,004	(0,115)	0,013	(0,153)	0,009	(0,115)
				0,9	0,007	(0,166)	0,005	(0,119)	0,002	(0,176)	0,002	(0,128)	-0,013	(0,176)	-0,009	(0,128)	0,022*	(0,175)	0,016*	(0,127)
			7	0,4	-0,005	(0,127)	-0,007	(0,158)	-0,014*	(0,138)	-0,018*	(0,173)	-0,006	(0,139)	-0,006	(0,173)	0,000	(0,138)	-0,001	(0,172)
				0,9	0,010	(0,153)	0,013	(0,185)	-0,004	(0,162)	-0,001	(0,197)	0,010	(0,161)	0,011	(0,197)	0,002	(0,161)	0,007	(0,196)
0	0,2	100	4	0,4	0,003	(0,200)	0,002	(0,149)	-0,007	(0,218)	-0,005	(0,163)	0,009	(0,219)	0,006	(0,162)	0,008	(0,217)	0,007	(0,162)
				0,9	0,016	(0,235)	0,012	(0,168)	0,007	(0,253)	0,004	(0,181)	0,006	(0,251)	0,004	(0,180)	0,008	(0,250)	0,005	(0,179)
			7	0,4	0,002	(0,179)	0,001	(0,224)	-0,01	(0,196)	-0,017	(0,244)	0,001	(0,195)	-0,001	(0,243)	-0,006	(0,194)	-0,007	(0,242)
				0,9	-0,009	(0,218)	-0,014	(0,263)	0,002	(0,230)	0,003	(0,280)	0,004	(0,228)	0,004	(0,278)	-0,017	(0,225)	-0,021	(0,274)
		200	4	0,4	-0,007	(0,140)	-0,005	(0,105)	0,002	(0,154)	0,002	(0,115)	0,018*	(0,154)	0,014*	(0,114)	0,001	(0,153)	0,000	(0,114)
				0,9	-0,001	(0,165)	-0,001	(0,118)	-0,004	(0,176)	-0,002	(0,127)	-0,002	(0,176)	-0,002	(0,127)	0,010	(0,175)	0,008	(0,126)
			7	0,4	0,002	(0,127)	0,003	(0,158)	0,003	(0,138)	0,004	(0,172)	0,008	(0,138)	0,010	(0,172)	-0,003	(0,138)	-0,004	(0,172)
				0,9	0,011	(0,153)	0,012	(0,185)	0,010	(0,162)	0,011	(0,197)	0,002	(0,162)	0,002	(0,197)	-0,010	(0,160)	-0,015	(0,195)

Suite page suivante

TABLE6.4-Suite

				no dropout				MCAR				MNAR							
												$\rho_{\theta\chi} = -0.4$				$\rho_{\theta\chi} = -0.9$			
				LRM		SM		LRM		SM		LRM		SM		LRM		SM	
Δ_θ	d_θ	N	J	ρ_θ	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)
0,5	0	100	4	0,4	0,508	(0,201)	0,380	(0,149)	0,511	(0,219)	0,381	(0,162)	0,500	(0,221)	0,373	(0,163)	0,484	(0,219)	0,361*	(0,161)
				0,9	0,506	(0,235)	0,375	(0,167)	0,506	(0,254)	0,372	(0,180)	0,507	(0,252)	0,375	(0,180)	0,500	(0,251)	0,370	(0,180)
			7	0,4	0,492	(0,181)	0,618	(0,224)	0,510	(0,196)	0,635	(0,243)	0,510	(0,197)	0,633	(0,244)	0,491	(0,196)	0,610	(0,243)
				0,9	0,503	(0,219)	0,627	(0,261)	0,513	(0,227)	0,633	(0,276)	0,493	(0,228)	0,608	(0,276)	0,501	(0,227)	0,625	(0,276)
		200	4	0,4	0,511	(0,141)	0,384	(0,104)	0,496	(0,154)	0,373	(0,114)	0,484*	(0,154)	0,364*	(0,114)	0,509	(0,154)	0,382	(0,114)
				0,9	0,521*	(0,167)	0,388	(0,119)	0,506	(0,176)	0,376	(0,127)	0,484	(0,177)	0,363	(0,127)	0,509	(0,176)	0,379	(0,127)
			7	0,4	0,494	(0,128)	0,618	(0,158)	0,499	(0,139)	0,624	(0,172)	0,496	(0,138)	0,619	(0,172)	0,507	(0,139)	0,632	(0,172)
				0,9	0,507	(0,154)	0,632	(0,185)	0,504	(0,161)	0,626	(0,195)	0,507	(0,161)	0,631	(0,195)	0,493	(0,161)	0,619	(0,196)
0,5	0,2	100	4	0,4	0,516	(0,200)	0,385	(0,148)	0,507	(0,219)	0,379	(0,162)	0,512	(0,220)	0,379	(0,162)	0,498	(0,219)	0,371	(0,161)
				0,9	0,511	(0,238)	0,377	(0,168)	0,514	(0,251)	0,381	(0,179)	0,516	(0,250)	0,378	(0,179)	0,529*	(0,252)	0,384	(0,178)
			7	0,4	0,497	(0,181)	0,619	(0,224)	0,505	(0,196)	0,627	(0,243)	0,506	(0,196)	0,629	(0,243)	0,507	(0,195)	0,629	(0,241)
				0,9	0,503	(0,219)	0,622	(0,262)	0,507	(0,229)	0,628	(0,277)	0,508	(0,230)	0,620	(0,277)	0,514	(0,227)	0,631	(0,274)
		200	4	0,4	0,499	(0,142)	0,373	(0,104)	0,523*	(0,155)	0,390	(0,114)	0,513	(0,154)	0,383	(0,114)	0,498	(0,154)	0,371	(0,113)
				0,9	0,509	(0,167)	0,378	(0,118)	0,500	(0,177)	0,369	(0,127)	0,503	(0,177)	0,373	(0,126)	0,494	(0,177)	0,364*	(0,126)
			7	0,4	0,497	(0,127)	0,621	(0,157)	0,501	(0,139)	0,624	(0,172)	0,493	(0,139)	0,614*	(0,172)	0,494	(0,138)	0,615*	(0,171)
				0,9	0,513	(0,154)	0,639	(0,184)	0,502	(0,162)	0,626	(0,196)	0,500	(0,162)	0,616	(0,195)	0,506	(0,162)	0,622	(0,195)

* test significatif à 5% - LRM : $H_0 : \mu_{\hat{\beta}_{gp}} = \Delta_\theta$ - SM : $H_0 : \mu_{\hat{\beta}_{gp}} = \Delta_S$ ($\Delta_S = 0.38$ pour J=4, $\Delta_S = 0.63$ pour J=7)

Comme attendu, l'estimation des erreurs standards de l'effet groupe diminue avec l'augmentation de la taille d'échantillon. Au contraire, les erreurs standards estimées augmentent avec le coefficient de corrélation de la variable latente. Cette augmentation peut expliquer la perte de puissance observée. En effet, le calcul de la puissance est basé sur un test de Wald du paramètre de l'effet groupe dont la statistique de test est $\hat{\beta}_{gp}/\hat{se}.(\hat{\beta}_{gp})$. L'estimation de l'erreur standard augmente avec la quantité de dropout quel que soit le type de dropout simulé.

6.1.3 Risque de première espèce et puissance de l'effet temps

L'ensemble des résultats présentés dans cette section ont été obtenus à partir de données simulées avec un effet groupe $\Delta_\theta = 0,5$. Les résultats concernant les données simulées sans effet groupe $\Delta_\theta = 0$ ont été présentés dans le chapitre 5 p.77.

TABLE 6.5 – Risque de première espèce de l'effet temps pour les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J), de la corrélation de la variable latente (ρ_θ) et du type de dropout (ρ_{θ_χ}). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour la méthode SM et une structure UN pour la méthode LRM. Effet groupe simulé $\Delta_\theta = 0,5$. Proportion de dropout : $\pi^{(t)} = 20\%$.

				no dropout		MCAR		MNAR			
								$\rho_{\theta\chi} = -0.4$		$\rho_{\theta\chi} = -0.9$	
Δ_θ	N	J	ρ_θ	LRM	SM	LRM	SM	LRM	SM	LRM	SM
0,5	100	4	0,4	0,042	0,038	0,056	0,050	0,049	0,042	0,072$^!$	0,060
			0,9	0,034	0,030	0,067	0,062	0,038	0,038	0,072$^!$	0,064
		7	0,4	0,040	0,036	0,056	0,048	0,081$^!$	0,062	0,100$^!$	0,084$^!$
			0,9	0,036	0,038	0,052	0,048	0,049	0,050	0,064	0,066
	200	4	0,4	0,046	0,042	0,054	0,046	0,066	0,062	0,100$^!$	0,098$^!$
			0,9	0,061	0,054	0,030$^!$	0,034	0,066	0,068	0,070	0,080$^!$
		7	0,4	0,048	0,046	0,058	0,050	0,046	0,046	0,110$^!$	0,108$^!$
			0,9	0,046	0,034	0,044	0,050	0,047	0,030$^!$	0,074$^!$	0,080$^!$

$^!$ l'intervalle de confiance à 95% ne contient pas la valeur attendue 5%.

Le tableau 6.5 présente les estimations du risque de première espèce de l'effet temps pour chaque méthode étudiée et pour certaines valeurs de la taille d'échantillon, du nombre d'items, de la corrélation de la variable latente et du type de dropout. Les résultats complets pour l'ensemble des valeurs de paramètres simulées sont présentés dans les tableaux 5.2 p.80 et D.9 p.184. Les deux méthodes présentent des risques de première espèce comparables.

Les valeurs du risque de première espèce sont maintenues proches de 5% lorsque les données sont complètes ou que le dropout est de type MCAR. Comme dans l'étude précédente, le risque de première espèce a tendance à augmenter avec l'informativité et le taux de dropout lorsque celui-ci est de type MNAR. Dans les pires cas, lorsque $\pi^{(t)} = 20\%$ et $\rho_{\theta\chi} = -0.9$, le risque de première espèce peut atteindre 10% pour les deux méthodes.

TABLE 6.6 – Puissance de l'effet temps pour les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J), de la corrélation de la variable latente (ρ_θ) et du type de dropout (ρ_{θ_χ}). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour la méthode SM et une structure UN pour la méthode LRM. Effet groupe simulé $\Delta_\theta = 0,5$. Proportion de dropout : $\pi^{(t)} = 20\%$.

				no dropout		MCAR		MNAR			
								$\rho_{\theta\chi} = -0.4$		$\rho_{\theta\chi} = -0.9$	
Δ_θ	N	J	ρ_θ	LRM	SM	LRM	SM	LRM	SM	LRM	SM
0,5	100	4	0,4	0,412	0,404	0,324	0,312	0,349	0,360	0,470	0,468
			0,9	0,500	0,428	0,367	0,296	0,466	0,430	0,608	0,570
		7	0,4	0,527	0,498	0,404	0,372	0,528	0,498	0,558	0,568
			0,9	0,753	0,628	0,590	0,480	0,609	0,548	0,744	0,704
	200	4	0,4	0,668	0,662	0,546	0,546	0,670	0,672	0,772	0,782
			0,9	0,804	0,716	0,706	0,630	0,755	0,730	0,866	0,860
		7	0,4	0,822	0,810	0,718	0,716	0,834	0,810	0,876	0,872
			0,9	0,957	0,918	0,846	0,796	0,929	0,908	0,988	0,972

en gris : l'intervalle de confiance à 95% du risque de première espèce ne contient pas la valeur attendue 5%

Le tableau 6.6 présente les estimations de la puissance de l'effet temps pour chaque méthode étudiée et pour certaines valeurs de la taille d'échantillon, du nombre d'items, de la corrélation de la variable latente et du type de dropout. Les résultats complets pour l'ensemble des valeurs de paramètres simulées sont présentés dans les tableaux 5.3 p.81 et D.10 p.185. La puissance de l'effet temps pour la méthode SM est proche de celle pour la méthode LRM lorsque $\rho_\theta = 0,4$. Comme pour l'étude précédente, la puissance pour la méthode LRM semble être légèrement mais systématiquement supérieure à celle pour la méthode LRM lorsque $\rho_\theta = 0,7$ ou $\rho_\theta = 0,9$. La puissance de l'effet temps augmente avec la taille d'échantillon, le nombre d'items et la corrélation de la variable latente.

Comme observé précédemment, la puissance pour les données sujettes à un dropout MCAR semble être inférieure à la puissance des données complètes. Au contraire, la puissance observée pour les données sujettes à un dropout MNAR semble être supérieure à la puissance

des données complètes.

6.1.4 Estimation de l'effet temps

Les graphiques 6.1 et 6.2 présentent l'estimation de l'effet temps entre le premier et le deuxième temps de mesure pour chaque méthode étudiée et pour certaines valeurs de la taille d'échantillon, du nombre d'items et de la corrélation de la variable latente lorsque l'effet groupe simulé vaut $\Delta_\theta = 0,5$. L'effet groupe simulé est approximé à $d_S = 0,38$ pour $J = 4$ et $d_S = 0,63$ pour $J = 7$. Les résultats présentés proviennent de données complètes ($\pi^{(t)} = 0\%$), sujettes à 20% de dropout MCAR ou sujettes à 20% de dropout MNAR ($\rho_{\theta_\chi} = -0,9$). Les résultats complets pour l'ensemble des valeurs de paramètres simulées sont présentés dans les tableaux C.1 p.166, C.2 p.168, D.11 p.186 et D.12 p.188.

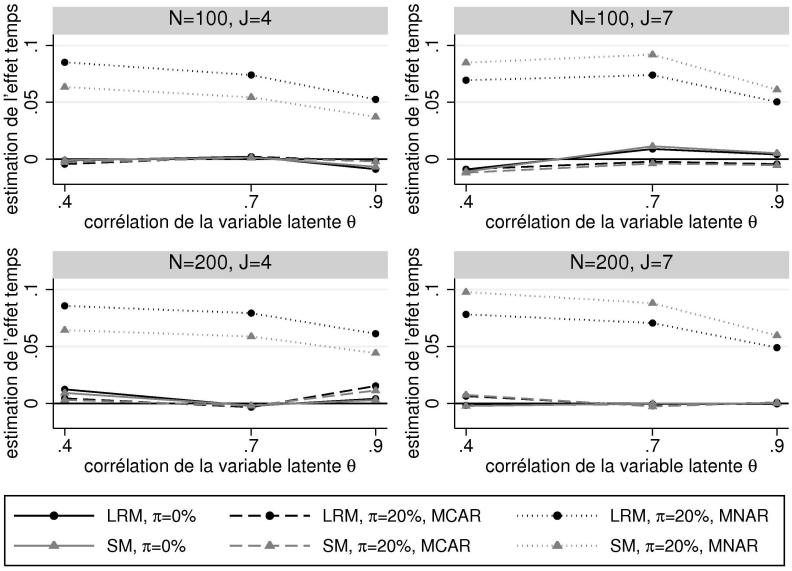

FIGURE 6.1 – Estimations de l'effet temps entre le temps 1 et le temps 2 pour les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J) et de la corrélation de la variable latente (ρ_θ). Données complètes ($\pi^{(t)} = 0\%$), sujettes à 20% de dropout MCAR ou sujettes à 20% de dropout MNAR ($\rho_{\theta_\chi} = -0,9$). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour la méthode SM et une structure UN pour la méthode LRM. Effet temps simulé : $d_\theta = 0$. Effet groupe simulé : $\Delta_\theta = 0,5$.

Les estimations de l'effet temps sont non-biaisées dans la plupart des cas pour des données complètes ou des données sujettes à un dropout MCAR. La plupart des cas pour lesquels les données sont sujettes à un dropout MNAR présentent des estimations biaisées de l'effet

temps. Lorsque les estimations sont biaisées, elles sont sur-estimées pour les deux méthodes. Le nombre de cas présentant des estimations biaisées augmente avec l'informativité et le taux de dropout.

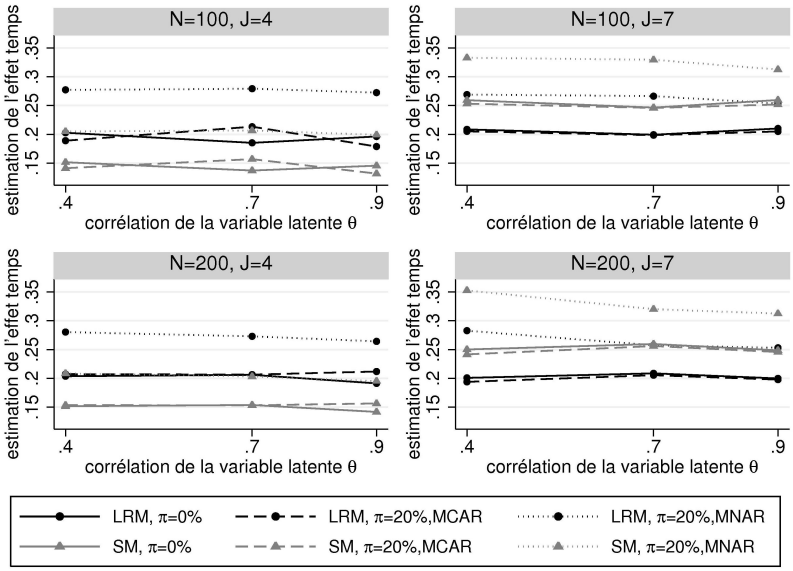

FIGURE 6.2 – Estimations de l'effet temps entre le temps 1 et le temps 2 pour les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J) et de la corrélation de la variable latente (ρ_θ). Données complètes ($\pi^{(t)} = 0\%$), sujettes à 20% de dropout MCAR ou sujettes à 20% de dropout MNAR ($\rho_{\theta_\chi} = -0,9$). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour la méthode SM et une structure UN pour la méthode LRM. Effet temps simulé : $d_\theta = 0,2$. Effet groupe simulé : $\Delta_\theta = 0,5$.

6.2 Discussion

Les résultats obtenus en terme d'effet temps dans cette étude sont comparables aux résultats obtenus dans l'étude précédente. Le risque de première espèce de l'effet temps est correctement maintenu proche de 5% pour les données complètes ou sujettes à du dropout MCAR. En cas de survenue de dropout MNAR, le risque de première espèce a tendance à augmenter avec l'informativité et le taux de dropout et la puissance est supérieure à celle des données complètes. Ces tendances s'expliquent par une sur-estimation de l'effet temps lorsque les données sont sujettes à du dropout non-ignorable. On peut donc conclure de la même façon qu'un effet groupe ait été simulé ou non. Les méthodes SM et LRM ont un comportement comparable et semblent être adaptées à l'analyse de données complètes ou sujettes à du dropout ignorable (MCAR). En revanche, elles doivent être adaptées pour

pouvoir être utilisées sur des données sujettes à du dropout non-ignorable.

Les deux méthodes ont également montré des résultats comparables en terme de risque de première espèce de l'effet groupe. Le risque est maintenu proche de 5% mais a tendance à augmenter en présence de dropout. Le dropout entraîne une perte de puissance quel que soit le type et la quantité de dropout. Comme attendu, la puissance de l'effet groupe augmente avec la taille d'échantillon et le nombre d'items. Au contraire, la puissance diminue avec l'augmentation de la corrélation entre les mesures de la variable latente. Une augmentation de l'erreur standard de l'estimation de l'effet groupe avec la corrélation de la variable latente a été observée alors que la corrélation ne semble pas avoir d'impact sur les estimations de l'effet groupe. Cette augmentation de l'erreur standard peut expliquer la diminution de la puissance en raison de l'expression du test de l'effet groupe. En effet, le test de Wald uitlisé s'exprime comme le ratio de l'estimation de l'effet groupe sur son erreur standard.

L'estimation de l'effet groupe est non-biaisée dans la plupart des cas. Lorsqu'un biais a été observé, celui-ci est faible, de l'ordre de 0,02. En l'absence d'effet groupe, les deux méthodes montrent autant de cas biaisés l'une que l'autre. En revanche, lorsqu'un effet groupe a été simulé et lorsque les données sont sujettes à du dropout MNAR, il semble que la méthode SM présente plus souvent des cas biaisés que la méthode LRM. L'effet groupe est alors sous-estimé pour la méthode SM. Pour la simulation du dropout MNAR, la corrélation entre la variable latente θ et la propension à être en dropout χ est négative. Ainsi, les patients dont le niveau du PRO étudié est le plus bas ont plus de risque de sortir de l'étude que les autres. On peut donc s'attendre à une sur-estimation de plus en plus importante de la moyenne du PRO étudié au fur et à mesure des temps d'évaluation. En présence d'un effet groupe, il semble que la sur-estimation de la moyenne est plus importante pour le groupe de patients dont le niveau moyen du PRO était plus bas (groupe 0 de la figure 3.1 p.54). Ceci pourrait expliquer la sous-estimation de l'effet groupe car il est estimé par la différence de moyennes entre les deux groupes.

Seul un faible impact du dropout a été observé sur l'effet groupe quel que soit le type et la quantité de données manquantes alors que les résultats en terme d'effet temps sont fortement impactés par la présence de dropout, surtout celui de type MNAR. L'effet groupe simulé valait $\Delta_\theta = 0,5$ alors que l'effet temps simulé entre deux temps consécutifs valait $d_\theta = 0,2$. On peut donc s'interroger sur l'impact de la taille d'effet sur les résultats. Les

bonnes performances des méthodes SM et LRM en terme d'effet groupe sont peut-être dues à la taille de l'effet groupe et pourraient être moins bonnes avec une taille d'effet plus petite et donc plus difficile à détecter.

Les deux groupes de patients ont été simulés en faisant l'hypothèse que les deux groupes avaient des niveaux moyens de PRO différents au premier temps d'étude. Cette hypothèse est réaliste dans des études où les patients ne sont pas randomisés. En revanche, dans les essais cliniques, il est généralement requis que les groupes de patients soient comparables à l'inclusion. Il est alors souvent attendu que l'évolution du PRO dans le temps soit différente d'un groupe à l'autre. La présence d'une interaction est alors possible et, dans ce cas, les deux méthodes présentées doivent être adaptées en ajoutant un terme d'interaction entre le temps et le groupe pour pouvoir analyser de type d'étude.

Chapitre 7

Application à deux études de qualité de vie

Les méthodes comparées dans les études de simulation ont été appliquées sur des données réelles. Ces applications visent à mettre en œuvre et à comparer ces méthodes dans deux cas de figure. Dans le premier cas, la dimension étudiée est issue d'un questionnaire de qualité de vie générique et est formée d'items dichotomiques. Cette application se place dans un cadre proche de celui des simulations. Au contraire, la deuxième application se place dans un cadre plus étendu. Les dimensions étudiées sont issues d'un questionnaire de qualité de vie spécifique aux patients atteints de pathologie cancéreuse et composées d'items polytomiques. La méthode LRM, basée sur l'IRT, a donc dû être adaptée par l'utilisation du rating-scale model, extension aux données polytomiques du modèle de Rasch. Les données de ces deux applications sont sujettes à du dropout.

7.1 Hyperparathyroïdie et SF-36

7.1.1 Symptômes non spécifiques et qualité de vie dans l'hyperparathyroïdie primaire modérée

L'hyperparathyroïdie primaire est due à une hypersécrétion de parathormone qui entraîne un excès de calcium dans le sang. Il n'existe pas de traitement médical de l'hyperparathyroïdie primaire. Le seul traitement définitif est un traitement chirurgical. La maladie est souvent découverte avant même l'apparition des signes cliniques dits "classiques". Des symptômes non

spécifiques à la maladie ont été rattachés à l'hyperparathyroïdie primaire : neuropsychiques (troubles de l'humeur, anxiété, dépression, irritabilité, troubles de la mémoire, céphalées), digestifs (nausées, constipation, douleurs abdominales), fatigue physique,... La plupart de ces symptômes régressent après une cure chirurgicale.

Une étude prospective, non randomisée, évaluant des symptômes non spécifiques et la qualité de vie avant et après chirurgie de l'hyperparathyroïdie primaire modérée a été mise en place dans les CHU de Grenoble, Poitiers, Angers, Limoges, Tours, Lille, Marseille et Nantes [18]. Cette étude avait pour but (i) de préciser les signes cliniques des patients atteints d'hyperparathyroïdie primaire modérée, (ii) de connaître le retentissement de cette affection sur leur qualité de vie, (iii) de préciser les symptômes susceptibles de devenir des indications opératoires en raison de leur disparition après guérison et (iv) inversement, les symptômes qui ne devront pas être retenus comme permettant de poser une indication opératoire et enfin, (v) d'évaluer l'évolution de la qualité de vie de ces patients après cure de leur hyperparathyroïdie primaire. Les patients inclus avaient tous un diagnostic d'hyperparathyroïdie primaire et étaient traités par une intervention chirurgicale. Les symptômes non-spécifiques ainsi que la qualité de vie des patients (au moyen du questionnaire SF-36) ont été évalués en période pré-opératoire, à 3 mois et 6 mois après la chirurgie.

7.1.2 Analyse de la dimension "limitations dues à l'état physique" du SF-36

Le cadre de cette étude est proche de celui des simulations. Une dimension contenant 4 items dichotomiques a été analysée avec les méthodes SM, RM, PV et LRM afin de comparer leurs résultats sur des données réelles et d'apprécier la pertinence du choix des paramètres de simulation.

Le questionnaire SF-36 [56] est un questionnaire générique de qualité de vie comprenant 36 items regroupés en 8 dimensions : activité physique, vie et relation avec les autres, douleurs physiques, santé perçue, vitalité, limitations dues à l'état psychique, limitations dues à l'état physique et santé psychique. La dimension "limitations dues à l'état physique" (RP) est formé de 4 items dichotomiques. Les données ont été analysées avec les méthodes SM, RM, PV et LRM. Pour les méthodes SM, RM et PV, les structures de matrice de variance covariance de type AR(1), ARH(1), UN, CS et CSH ont été étudiées. Les résultats avec le modèle ayant le

plus petit AIC seront présentés pour chacune de ces méthodes. Afin d'être comparable avec les résultats des études de simulation, les résultats obtenus avec une matrice de variance covariance de type AR(1) pour les méthodes SM, RM et PV seront également présentés. Une structure de matrice de variance covariance de type UN a été étudiée pour la méthode LRM. Le score d'un individu sur la dimension est la somme des réponses aux items et varie de 0 (niveau de symptôme le plus élevé) à 4 (niveau de symptôme le plus bas). 57 patients ont été évalués en période pré-opératoire, 47 à 3 mois après la chirurgie et 40 à 6 mois après la chirurgie.

TABLE 7.1 – Analyse de la dimension "limitations dues à l'état physique" du SF-36 avec les méthodes Score and Mixed models (SM), Rasch and Mixed models (RM), Plausible Values (PV) et Longitudinal Rasch Mixed model (LRM). Estimations de l'effet temps entre t et t' ($\hat{d}_{tt'}$), variance au temps t ($\hat{\sigma}^2_t$), corrélation entre le temps t et t' ($\hat{\rho}_{tt'}$) pour $t = 1, 2, 3$ et $t \neq t'$ et taille d'effet entre t et t' ($\widehat{ES}_{tt'}$). P-value du test de la présence d'un effet temps global.

	SM		RM		PV		LRM
Structure	AR(1)	CS	AR(1)	CS	AR(1)	CS	UN
AIC	455,5	453,2	586,7	584,3	648,6	644,4	
$\hat{d}_{12}$	1,32*	1,28*	2,23*	2,18*	2,71*	2,71*	3,10*
$\hat{d}_{23}$	-0,45	-0,43	-0,76	-0,73	-0,78	-0,71	-0,94
$\hat{d}_{13}$	0,87*	0,86*	1,47*	1,45*	1,93*	1,99*	2,16*
$\hat{\sigma}^2_1$	2,41	2,39	6,87	6,83	9,80	9,86	9,46
$\hat{\sigma}^2_2$	2,41	2,39	6,87	6,83	9,80	9,86	5,61
$\hat{\sigma}^2_3$	2,41	2,39	6,87	6,83	9,80	9,86	18,86
$\widehat{ES}_{12}$	0,85	0,83	0,85	0,83	0,86	0,86	1,12
$\widehat{ES}_{23}$	-0,29	-0,28	-0,29	-0,28	-0,25	-0,23	-0,27
$\widehat{ES}_{13}$	0,56	0,55	0,56	0,55	0,62	0,63	0,59
$\hat{\rho}_{12}$	0,56	0,52	0,56	0,52	0,31	0,36	0,54
$\hat{\rho}_{23}$	0,56	0,52	0,56	0,52	0,31	0,36	0,61
$\hat{\rho}_{13}$	0,31	0,52	0,31	0,52	0,10	0,36	0,65
p-value	<0,0001	<0,0001	<0,0001	<0,0001	<0,0001	<0,0001	0,002

* indique que le test de Student est significatif à 5% ($H_0 : d_{tt'} = 0$).

Les résultats de l'analyse de la dimension "limitations dues à l'état physique" du SF-36 avec les méthodes SM, RM, PV et LRM sont présentés dans le tableau 7.1. Ce tableau compare les estimations des effets temps, des variances et corrélations ainsi que les tailles

d'effet (effect size - ES) exprimés par le ratio de l'effet temps entre deux temps donnés t et t' sur la racine de la moyenne de la variance de ces deux temps : $\widehat{ES}_{tt'} = \frac{\hat{d}_{tt'}}{\sqrt{\frac{n_t\hat{\sigma}_t^2+n_{t'}\hat{\sigma}_{t'}^2}{n_t+n_{t'}}}}$ avec n_t le nombre d'individus au temps t. Parmi les structures étudiées, la structure de type CS minimisait l'AIC pour les méthodes SM, RM et PV. On peut noter que les résultats obtenus avec une structure de type AR(1) sont proches de ceux obtenus avec une structure de type CS.

Toutes les méthodes ont conclu à la présence d'un effet temps global. L'hypothèse d'égalité des moyennes au seuil de 5% est rejetée par toutes les méthodes. De même, les méthodes ont conclu à la présence d'un effet temps entre le temps 1 et le temps 2 et à l'absence d'effet temps entre le temps 2 et le temps 3. L'augmentation sur la dimension RP entre la période pré-opératoire et le troisième mois après la chirurgie indique une diminution des limitations dues à l'état physique sur cette période. Ensuite, le niveau des limitations est stable entre le troisième et le sixième mois post-opératoire. L'effet temps entre le temps 1 et le temps 3 est significativement différent de 0 au seuil de 5% pour toutes les méthodes. Il y a une augmentation entre la période pré-opératoire et le sixième mois après la chirurgie et donc une diminution des limitations dues à l'état physique sur le temps total de l'étude. Si toutes les méthodes présentent les mêmes conclusions en terme d'effet temps, on observe que, comme pour les études de simulation, les estimations des effets temps sont généralement plus grandes pour la méthode LRM que pour les méthodes RM et PV.

Parallèlement, les estimations des variances sont plus grandes pour la méthode PV que pour la méthode RM ce qui est cohérent avec les études de simulation. En effet, la méthode PV permet de corriger le biais de l'estimation de la variance de la méthode RM. Pour la méthode LRM, les variances estimées sont très différentes d'un temps à l'autre. De plus, la corrélation estimée entre le temps 1 et le temps 3 $\hat{\rho}_{13}$ n'est pas proche du carré de la corrélation entre deux mesures consécutives. La structure de la matrice de variance covariance ne semble donc pas être de type AR(1). Comme pour les études de simulation, les corrélations estimées sont plus petites pour la méthode PV que pour la méthode RM.

Les tailles d'effet estimées, comprises en valeur absolue entre 0,23 et 1,12, sont plus grandes que la taille d'effet des simulations qui vaut 0,2 pour la variable latente. Il se peut que les effets temps sur la dimension "limitations dues à l'état physique" soient plus facilement

détectés en raison de la grande taille d'effet, surtout entre le premier et le deuxième temps d'étude. Le cas de simulation le plus proche des données sur l'hyperparathyroïdie semble être le cas où $N = 100$, $J = 4$ et $\rho_\theta = 0,7$. Dans ce cas, la puissance a été estimée à 37,2%, 12,0%, 14,2% et 42,3% pour les méthodes SM, RM, PV et LRM respectivement. Le taux de dropout des données sur l'hyperparathyroïdie est d'environ 18% au temps 2 et d'environ 15% au temps 3.

7.2 Cancer du sein et EORTC QLQ-C30

7.2.1 Etude longitudinale de la qualité de vie et des stratégies d'adaptation des patientes atteintes de cancer du sein et de leur accompagnant

Une étude longitudinale de la qualité de vie et des stratégies d'adaptation des patientes atteintes de cancer du sein et de leur accompagnant a été menée au Centre régional de lutte contre le cancer René Gauducheau de Nantes. Cette étude portait à la fois sur les patientes, atteintes d'une pathologie cancéreuse au niveau du sein en phase initiale, et sur les personnes dites "accompagnantes". Chaque patiente désignait librement comme accompagnant une personne de son entourage que ce soit son époux, un membre de la famille ou un(e) ami(e). Le principal objectif de cette étude était d'étudier le retentissement de la maladie dans la vie quotidienne de la patiente (qualité de vie) ainsi que la stratégie d'adaptation à la maladie employée (coping). Parallèlement, il s'agissait d'étudier ces mêmes critères chez la personne accompagnant la patiente et d'essayer de comprendre les liens existant entre les deux. A terme, l'objectif pragmatique et préventif était de proposer des soutiens spécifiques, adaptés aux patientes et à leur entourage.

La notion de coping désigne l'ensemble des processus qu'un individu utilise pour maîtriser l'impact d'un évènement perçu comme négatif sur son bien-être physique et psychologique. On distingue plusieurs types de coping dont (i) la focalisation de l'individu sur le problème, (ii) la focalisation sur l'émotion ou (iii) la recherche de soutien social. La focalisation sur le problème correspond à des efforts cognitifs de résolution du problème et à des stratégies comportementales qui réduisent ou gèrent la source du problème. La focalisation sur l'émotion renvoie, elle, aux efforts cognitifs et comportementaux qui gèrent ou réduisent la détresse

émotionnelle. Cela peut se traduire par une restructuration cognitive, une minimisation de la portée de l'évènement, une relativisation ou encore une négation du problème. Dans la recherche de soutien social, il s'agit pour les patients de tentatives effectives pour obtenir une écoute, des informations ou encore une aide matérielle.

Les patientes ont été vues au cours de 3 entretiens : deux ou trois semaines après le diagnostic (t_1), à la fin des traitements de chimiothérapie et/ou radiothérapie (t_2) et six mois après les traitements (t_3). A chaque temps, la qualité de vie des patientes était évaluée avec le Quality of Life Questionnaire C30 (EORTC QLQ-C30) [1], la qualité de vie des accompagnants avec le questionnaire Duke Health Profile (DHP) [84] et les stratégies d'adaptation avec le questionnaire Ways of Coping Checklist (WCC) [21]. Cent patientes et leur accompagnant, leur conjoint pour la plupart, ont été évaluées à t_1, 82 patientes à t_2 et 80 patientes à t_3.

L'étude a permis d'établir une typologie des patientes (Bonnaud-Antignac A, Hardouin JB, Léger J, Dravet F, Sébille V. Quality of life and coping of women treated for a breast cancer and their caregiver : A proximological approach using mixed models. Journal of Clinical Psychology in Medical Settings. *en révision*). La typologie prend en compte les caractéristiques socio-démographiques des patientes et des accompagnants, la qualité de vie des accompagnants ainsi que le coping des patientes et des accompagnants. Elle a été établie au moyen d'une analyse des correspondances multiples suivie d'une classification ascendante hiérarchique. Quatre groupes de patientes ont été identifiés. L'un des groupes ne comprenait que deux patientes dont les caractéristiques étaient très différentes des autres patientes. Ce groupe n'a pas été retenu dans l'analyse présentée ci-dessous.
Le premier groupe (g_1) est composé de 52 patientes. Il est caractérisé par des patientes plus âgées, plus nombreuses à vivre seules et ayant des scores de coping centré sur le problème et de coping centré sur l'émotion moins élevés que l'ensemble de l'échantillon. Les accompagnants sont plus souvent âgés de moins de 45 ans, ont une meilleure santé perçue et ont des scores de coping moins élevés, quelle que soit la stratégie.
Le deuxième groupe (g_2) est composé de 28 patientes. Il est caractérisé par des patientes appartenant principalement à la tranche d'âge 55-64 ans et ayant été moins souvent confrontées à un cancer dans leur entourage. Le coping des patientes de ce groupe est plus centré sur le problème ainsi que sur l'émotion que l'ensemble de l'échantillon. Les accompagnants

sont principalement des hommes, très présents lors de l'annonce du diagnostic et ayant une qualité de vie moins bonne que celle de l'ensemble des accompagnants.
Le troisième groupe (g_3) est composé de 18 patientes. Il est caractérisé par des patientes plus jeunes dont beaucoup sont en activité et qui vivent toutes en couple. Les patientes ont des scores de coping plus élevés que l'ensemble de l'échantillon, quelle que soit la stratégie. Les accompagnants de ces patientes sont également plus jeunes et sont plus souvent en activité que l'ensemble des accompagnants.

Cette étude se place dans un cadre plus étendu que celui des simulations. Les dimensions du questionnaire QLQ-C30 sont composées d'items polytomiques et la population de l'étude est divisée en trois groupes. La comparaison des résultats des différentes méthodes sur ces données va permettre d'évaluer la capacité d'adaptation des méthodes et de soulever de possibles points d'amélioration pour permettre leur utilisation dans un plus large champ d'étude et non pas uniquement dans un contexte identique à celui des études de simulation.

Le questionnaire QLQ-C30 est un questionnaire spécifique de qualité de vie pour les patients atteints de cancer comprenant 30 items regroupés en 15 dimensions : santé globale, fonctionnement physique, limitation fonctionnelle, fonctionnement émotionnel, fonctionnement cognitif, fatigue, nausées, douleur, dyspnée, insomnie, perte d'appétit, constipation, diarrhée et difficultés financières. Les trois dimensions avec le plus grand nombre d'items ont été analysées avec les méthodes SM et LRM : fonctionnement physique (5 items), fonctionnement émotionnel (4 items) et fatigue (3 items). Elles sont composées d'items polytomiques à 4 modalités de réponse. Les méthodes RM et PV n'ont pas été étudiées en raison de leurs faibles performances dans les études de simulation.

7.2.2 Méthodes utilisées

Longitudinal Rating-Scale Model - LRSM Etant donné que les items sont polytomiques, la méthode LRM a été adaptée pour pouvoir analyser ce type d'items. Au lieu d'être basée sur le modèle de Rasch, la méthode Longitudinal Rating-Scale Model (LRSM) est basée sur le rating-scale model (cf 2.3.3) [9], modèle de la famille de Rasch adapté à l'analyse d'items polytomiques. La probabilité de réponse de l'individu i à la modalité positive h

de l'item j au temps t, $\forall g = 1, 2, 3$, s'écrit :

$$P(Y_{i[g]j}^{(t)} = h^{(t)} | \theta_i^{(t)}, \delta_j, \tau_2, ..., \tau_m, (\beta_{gpe})) = \frac{exp(h^{(t)}(\theta_{i[g]}^{(t)} + \beta_{gpe[g]} - \delta_j) - \sum_{p=2}^{h} \tau_p)}{\sum_{l=0}^{m} exp(l(\theta_{i[g]}^{(t)} + \beta_{gpe[g]} - \delta_j) - \sum_{p=2}^{l} \tau_p)} \quad (7.1)$$

$$\boldsymbol{\theta_{i[g]}} = (\theta_{i[g]}^{(1)}, ..., \theta_{i[g]}^{(3)})' \sim N_3 \left(\boldsymbol{\mu}_{i[g]}, \boldsymbol{\Sigma}_i \right)$$

La contrainte d'identifiabilité est la nullité de la somme des moyennes de la variable latente dans le groupe de référence (g_3).

Pour chaque dimension, différentes structures de variance covariance ont été testées : UN, AR(1), ARH(1), CS, CSH ou identité.

$$\boldsymbol{\Sigma} = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix} \text{ AR(1)} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho^2 \\ \sigma_1\sigma_2\rho & \sigma_2^2 & \sigma_2\sigma_3\rho \\ \sigma_1\sigma_3\rho^2 & \sigma_2\sigma_3\rho & \sigma_3^2 \end{pmatrix} \text{ ARH(1)}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma^2 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma^2 \end{pmatrix} \text{ CS} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 & \sigma_2\sigma_3\rho \\ \sigma_1\sigma_3\rho & \sigma_2\sigma_3\rho & \sigma_3^2 \end{pmatrix} \text{ CSH}$$

Pour chaque dimension, le modèle avec le plus petit AIC a été retenu.

Score and Mixed models - SM Les scores sont calculés suivant les recommandations de l'EORTC [32]. Pour chaque dimension, le score d'un individu varie de 0 (niveau de symptôme ou de fonctionnement le plus bas) à 100 (niveau de symptôme ou de fonctionnement le plus élevé). Soit $S_i^{(t)}$ le score standardisé de l'individu i au temps t :

$$S_i^{(t)} = \frac{\frac{1}{J}\sum_j y_{ij}^{(t)} - 1}{\text{étendue}} * 100 \quad (7.2)$$

où l'étendue est la différence entre le score brut maximal (score brut=$\frac{1}{J}\sum_j y_{ij}$) et le score brut minimal de la dimension considérée. Le modèle linéaire mixte utilisé pour expliquer les scores s'écrit :

$$\begin{aligned} &\boldsymbol{S}_i = (S_i^{(1)}, S_i^{(2)}, S_i^{(3)})' = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i \\ &\boldsymbol{b}_i \sim N_q(0, \boldsymbol{D}), \\ &\boldsymbol{e}_i \sim N_{n_i}(0, \boldsymbol{\Sigma}_i), \\ &\boldsymbol{b}_1, ..., \boldsymbol{b}_N, \boldsymbol{e}_1, ..., \boldsymbol{e}_N \textit{ indépendants} \end{aligned} \quad (7.3)$$

Le modèle complet comprend des effets temps, des effets groupes ainsi que les interactions d'ordre 1 temps x groupe comme effets fixes, une constante et une pente aléatoires et une structure de variance covariance pour les effets aléatoires ou variance covariance résiduelle de type UN, AR(1), ARH(1), CS, CSH ou identité. Parmi l'ensemble des modèles, le modèle avec le plus faible AIC a été retenu pour chaque dimension. Les effets fixes du modèle retenu ont ensuite été réduits lorsque le test de l'interaction temps x groupe ne rejetait pas l'hypothèse de la nullité du paramètre associé à l'interaction au seuil de 5%.

Estimation et test des effets Les effet temps et groupe ont été estimés comme précédemment par la différence des moyennes estimées. Les tailles d'effet (effect size - ES) de l'effet temps sont définies comme le ratio de l'effet temps entre deux temps donnés t et t' sur la racine de la moyenne de la variance de ces deux temps : $\widehat{ES}_{tt'} = \frac{\hat{d}_{tt'}}{\sqrt{\frac{n_t\hat{\sigma}^2_t + n_{t'}\hat{\sigma}^2_{t'}}{n_t+n_{t'}}}}$ avec n_t le nombre d'individus au temps t.
La nullité des effets et des tailles d'effet a été testée avec un test de Student. Le test global de l'effet temps ou de l'effet groupe utilise le test de Fisher.

7.2.3 Dimension "fonctionnement physique" du QLQ-C30

Le tableau 7.2 présente les résultats de l'analyse de la dimension "fonctionnement physique" avec les méthodes SM et LRSM. La structure de variance covariance retenue pour la méthode LRSM est de type CS. Le modèle retenu pour la méthode SM comprend pour les effets fixes : les effets temps et les effets groupes et une structure de variance covariance de type CSH. Il ne comprend pas d'interaction temps x groupe ni d'effets aléatoires. Les méthodes SM et LRSM montrent des résultats comparables. Elles concluent toutes deux à la présence d'un effet temps global. L'effet temps entre les temps 1 et 2 est négatif et l'effet temps entre le temps 2 et 3 est positif. Le "fonctionnement physique" s'est dégradé entre le diagnostic (t_1) et la fin des traitements (t_2). Il s'est ensuite amélioré entre la fin des traitements (t_2) et le sixième mois après les traitements (t_3). Néanmoins, l'augmentation sur la deuxième période est de taille plus faible que la diminution de la première période entraînant une diminution globale du niveau de "fonctionnement physique" sur l'ensemble de la période de l'étude. On retrouve les mêmes tendances pour les tailles d'effet estimées. Les tailles d'effet estimées sur le score (méthode SM) sont plus petites que pour la variable latente (méthode LRSM). Les deux méthodes concluent à l'absence d'un effet groupe global. De faibles différences entre

TABLE 7.2 – Estimations des paramètres et des erreurs standards de la dimension "fonctionement physique" du QLQ-C30 avec les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM). Estimations de l'effet temps entre t et t' ($\hat{d}_{tt'}$), variance au temps t ($\hat{\sigma}_t^2$), coefficient de corrélation($\hat{\rho}$), taille de l'effet temps entre t et t' ($\widehat{ES}_{tt'}$), l'effet groupe entre les groupes g et g' ($\hat{d}_{gpe_{gg'}}$) pour $t = 1, 2, 3$ et $g = 1, 2, 3$. Valeur de la statistique de test et p-value du test de la présence d'un effet temps global ou d'un effet groupe global.

Méthode	longitudinal RSM		Score and Mixed Models	
Structure	CS		CSH	
Paramètre	Estimation	Standard erreur	Estimation	Standard erreur
$\hat{\rho}$	0,848	0,052	0,686	0,048
$\hat{\sigma}_1^2$	3,361	0,725	244,780	35,690
$\hat{\sigma}_2^2$	3,361	0,725	417,460	66,505
$\hat{\sigma}_3^2$	3,361	0,725	273,430	43,825
$\hat{d}_{12}$	-1,063*	0,209	-7,862*	1,665
$\hat{d}_{23}$	0,401*	0,198	3,495*	1,710
$\hat{d}_{13}$	-0,662*	0,209	-4,367*	1,424
$\widehat{ES}_{12}$	-0,580*	0,116	$-0,438^{!}$	———
$\widehat{ES}_{23}$	0,219*	0,109	$0,188^{!}$	———
$\widehat{ES}_{13}$	-0,361*	0,114	$-0,272^{!}$	———
$\hat{d}_{gpe_{12}}$	-0,195	0,452	-0,079	3,497
$\hat{d}_{gpe_{23}}$	0,378	0,580	1,559	4,489
$\hat{d}_{gpe_{13}}$	0,183	0,526	1,480	4,057
Test	Statistique	P-value	Statistique	P-value
Effet temps	13,010	<0,0001	12,010	<0,0001
Effet groupe	0,220	0,803	0,080	0,927

* indique que le test de Student est significatif à 5%
! Le test n'a pas pu être effectué avec cette méthode.

chaque groupe deux à deux sont observées et celles-ci ne sont pas significatives au seuil de 5%.

7.2.4 Dimension "fonctionnement émotionnel" du QLQ-C30

Le tableau 7.3 présente les résultats de l'analyse de la dimension "fonctionnement émotionnel" avec les méthodes SM et LRSM. La structure de variance covariance retenue pour la méthode LRSM est de type CSH. Le modèle retenu pour la méthode SM comprend pour les effets fixes : les effets temps et les effets groupes et une structure de variance covariance de type CS. Il ne comprend pas d'interaction temps x groupe ni d'effets aléatoires. Les méthodes SM et LRSM montrent des résultats comparables. Elles concluent toutes deux à la présence d'un effet temps global. Le "fonctionnement émotionnel" s'est amélioré entre le diagnostic

TABLE 7.3 – Estimations des paramètres et des erreurs standards de la dimension "fonctionement émotionnel" du QLQ-C30 avec les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM). Estimations de l'effet temps entre t et t' ($\hat{d}_{tt'}$), variance au temps t ($\hat{\sigma}_t^2$), coefficient de corrélation($\hat{\rho}$), taille de l'effet temps entre t et t' ($\widehat{ES}_{tt'}$), l'effet groupe entre les groupes g et g' ($\hat{d}_{gpe_{gg'}}$) pour $t = 1, 2, 3$ et $g = 1, 2, 3$. Valeur de la statistique de test et p-value du test de la présence d'un effet temps global ou d'un effet groupe global.

Méthode	longitudinal RSM		Score and Mixed Models	
Structure	CSH		CS	
Paramètre	Estimation	Standard erreur	Estimation	Standard erreur
$\hat{\rho}$	0,528	0,078	0,447	———
$\hat{\sigma}_1^2$	3,746	0,863	607,990	70,917
$\hat{\sigma}_2^2$	6,915	1,657	607,990	70,917
$\hat{\sigma}_3^2$	4,476	1,123	607,990	70,917
$\hat{d}_{12}$	0,858*	0,310	5,927*	2,820
$\hat{d}_{23}$	0,409	0,333	3,619	2,946
$\hat{d}_{13}$	1,267*	0,285	9,546*	2,845
$\widehat{ES}_{12}$	0,377*	0,135	0,240!	———
$\widehat{ES}_{23}$	0,171	0,140	0,147!	———
$\widehat{ES}_{13}$	0,628*	0,142	0,387!	———
$\hat{d}_{gpe_{12}}$	-1,333*	0,457	-12,566*	4,773
$\hat{d}_{gpe_{23}}$	0,297	0,576	1,805	6,078
$\hat{d}_{gpe_{13}}$	-1,036	0,525	-10,761	5,483
Test	Statistique	P-value	Statistique	P-value
Effet temps	10,580	<0,0001	5,830	0,004
Effet groupe	4,860	0,010	4,210	0,018

* indique que le test de Student est significatif à 5%
! Le test n'a pas pu être effectué avec cette méthode

(t_1) et la fin des traitements (t_2) avec un effet temps entre les temps 1 et 2 significativement non-nul. Il semble s'être ensuite amélioré dans une moindre mesure entre la fin des traitements (t_2) et le sixième mois après les traitements (t_3) mais l'effet temps entre les temps 2 et 3 n'est pas significatif. On note une augmentation globale du niveau de "fonctionnement émotionnel" sur l'ensemble de la période de l'étude. On retrouve les mêmes tendances pour les tailles d'effet estimées. Les tailles d'effet estimées sur le score (méthode SM) sont plus petites que pour la variable latente (méthode LRSM). Les deux méthodes concluent à la présence d'un effet groupe global. Il existe un effet groupe significatif entre le groupe 1 et le groupe 2 au seuil de 5%. Le groupe 1 a un niveau de "fonctionnement émotionnel" plus élevé que le groupe 2. La différence entre le groupe 1 et le groupe 3 est également importante et à la limite de la significativité.

7.2.5 Dimension "fatigue" du QLQ-C30

TABLE 7.4 – Estimations des paramètres et des erreurs standards de la dimension "fatigue" du QLQ-C30 avec les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM). Estimations de l'effet temps entre t et t' ($\hat{d}_{tt'}$), variance au temps t ($\hat{\sigma}_t^2$), coefficient de corrélation($\hat{\rho}$), taille de l'effet temps entre t et t' ($\widehat{ES}_{tt'}$), l'effet groupe entre les groupes g et g' ($\hat{d}_{gpe_{gg'}}$) pour $t = 1, 2, 3$ et $g = 1, 2, 3$. Valeur de la statistique de test et p-value du test de la présence d'un effet temps global ou d'un effet groupe global.

Méthode	longitudinal RSM		Score and Mixed Models	
Structure	ARH(1)		CSH	
Paramètre	Estimation	Standard erreur	Estimation	Standard erreur
$\hat{\rho}$	0,636	0,068	0,520	0,065
$\hat{\sigma}_1^2$	9,920	2,427	591,200	87,801
$\hat{\sigma}_2^2$	11,546	2,993	770,740	121,950
$\hat{\sigma}_3^2$	6,291	1,621	515,530	82,910
$\hat{d}_{12}$	1,587*	0,390	12,029*	2,837
$\hat{d}_{23}$	-0,758*	0,367	-6,559*	2,847
$\hat{d}_{13}$	0,829*	0,401	5,470*	2,539
$\widehat{ES}_{12}$	0,486*	0,120	0,464!	———
$\widehat{ES}_{23}$	-0,253*	0,123	−0,258!	———
$\widehat{ES}_{13}$	0,288*	0,138	0,232!	———
$\hat{d}_{gpe_{12}}$	1,393*	0,627	10,029*	4,898
$\hat{d}_{gpe_{23}}$	-1,259	0,796	-9,561	6,242
$\hat{d}_{gpe_{13}}$	0,135	0,710	0,468	5,639
Test	Statistique	P-value	Statistique	P-value
Effet temps	8,330	0,001	9,040	0,000
Effet groupe	2,590	0,081	2,260	0,110

* indique que le test de Student est significatif à 5%
! Le test n'a pas pu être effectué avec cette méthode

Le tableau 7.4 présente les résultats de l'analyse de la dimension "fatigue" avec les méthodes SM et LRSM. La structure de variance covariance retenue pour la méthode LRSM est de type ARH(1). Le modèle retenu pour la méthode SM comprend pour les effets fixes : les effets temps et les effets groupes et une structure de variance covariance de type CSH. Il ne comprend pas d'interaction temps x groupe ni d'effets aléatoires. Les méthodes SM et LRSM montrent des résultats comparables. Elles concluent toutes deux à la présence d'un effet temps global. La symptomatologie sur la dimension "fatigue" a augmenté entre le diagnostic (t_1) et la fin des traitements (t_2) avec un effet temps entre les temps 1 et 2 significativement non-nul. Elle a ensuite diminué dans une moindre mesure entre la fin des traitements (t_2) et le sixième mois après les traitements (t_3). Globalement, on note une augmentation significative

des symptômes sur la dimension "fatigue" sur l'ensemble de la période de l'étude. On retrouve les mêmes tendances pour les tailles d'effet estimées. Les tailles d'effet estimées par les deux méthodes sont proches. Les deux méthodes concluent à l'absence d'un effet groupe global. Il existe un effet groupe significatif entre le groupe 1 et le groupe 2 au seuil de 5%. La différence entre le groupe 2 et le groupe 3 est également importante mais n'est pas significative. Les groupes 1 et 3 sont proches en terme de symptômes sur la dimension "fatigue". Le groupe 2 présente plus de symptômes que les groupes 1 et 3 sur la dimension "fatigue".

7.3 Discussion

Ces deux applications ont permis d'utiliser les différentes méthodes d'analyse comparées dans les études de simulation sur des données réelles. Dans l'étude sur l'hyperparathyroïdie, la dimension étudiée était composée d'items dichotomiques et les méthodes ont pu être utilisées sans modifications. Dans l'étude sur le cancer du sein, les dimensions du QLQ C-30 sont composées d'items polytomiques et la méthode LRM basée sur le modèle de Rasch a du être adaptée en développant la méthode LRSM basée sur le rating-scale model, modèle de la famille de Rasch pour les items polytomiques. Cette extension du modèle se trouve confrontée au problème de l'augmentation du nombre de paramètres à estimer. Avec la méthode LRSM, l'introduction d'une interaction temps x groupe ne menait pas à la convergence du modèle quelle que soit la structure de variance covariance choisie. La pertinence de l'ajout d'une interaction dans le modèle n'a donc pas pu être testée. La méthode SM a également été confrontée à des problèmes de convergence des modèles lorsqu'ils contenaient beaucoup d'effets différents. Ce problème risque fort d'être fréquent en santé où les tailles d'échantillon sont plus petites qu'en sciences de l'éducation, discipline dans laquelle les modèles IRT ont initialement été développés. Il convient donc de privilégier la parcimonie dans la construction des modèles ce qui est confirmé par les modèles finalement retenus dans l'étude sur le cancer du sein.

Toute étude de simulation est confrontée au choix des paramètres de simulation. Ces paramètres doivent être le plus proche possible des valeurs fréquemment rencontrées en pratique. Dans les deux applications, les valeurs estimées du coefficient de corrélation pour la variable latente sont proches des valeurs simulées. En revanche, les estimations des variances sont plus élevées que les valeurs simulées et semblent varier dans le temps. De plus, les

structures de variance covariance retenues sont éloignées de la structure AR(1) utilisée pour la simulation des données. Il est envisageable d'étendre l'étude de simulation pour étudier l'impact de la valeur de la variance sur les résultats des différentes méthodes.

Par ailleurs, l'effet temps simulé augmente linéairement dans le temps. Ce n'est pas le cas dans les deux applications présentées. L'hypothèse faite pour les simulations peut se révéler inadéquate dans beaucoup d'études. Il est fréquent que l'évolution d'un PRO soit évaluée avant la phase de traitement des patients puis plusieurs fois après les traitements. Dans ce cas, on peut s'attendre à une dégradation du niveau du PRO sur la première période puis à une stabilité ou une amélioration sur les périodes post-traitement. Toutefois, les méthodes étudiées estiment l'effet temps sans faire l'hypothèse de linéarité de l'effet temps et peuvent donc être utilisées quelle que soit sa forme.

Les tailles d'effet estimées dans les applications étaient généralement plus élevées que la taille d'effet simulée. On peut donc s'attendre à une meilleure puissance des méthodes dans les applications pour des effectifs similaires. Enfin, les modèles utilisées ne prennent pas en compte l'espacement des temps de mesure car le temps est considéré comme une variable discrète. Or, dans les deux applications, les patients ne sont pas évalués à des temps régulièrement espacés. De plus, les temps d'évaluation déterminés dans les protocoles sont approximatifs et les patients peuvent, par exemple, être vus entre le cinquième et le huitième mois après traitement pour une évaluation fixée dans le protocole au sixième mois après traitement. Il semble important de pouvoir prendre en compte cette information en traitant le temps comme une variable continue.

Discussion générale

Dans ce travail, nous avons comparé différentes méthodes basées sur les deux approches existantes pour l'analyse de Patient-Reported Outcomes : la théorie classique des tests et la théorie de réponse aux items. Le principal objectif était de déterminer la méthode la plus adéquate pour analyser des PRO recueillis de manière longitudinale et issus d'un questionnaire validé avec un modèle de Rasch.
Plusieurs méthodes d'analyse ont été comparées dans des situations différentes à travers diverses études de simulation. La première étude a permis de comparer quatre méthodes d'analyse, une basée sur la CTT et trois sur l'IRT, dans le cadre de données complètes et un groupe de patients. Dans les deuxième et troisième études, une méthode basée sur la CTT et une méthode basée sur l'IRT ont été comparées dans le cadre de données complètes ou sujettes à du dropout de type MCAR ou MNAR et portant sur un ou deux groupes de patients.

Méthodes d'analyse à privilégier

Parmi les méthodes basées sur l'IRT, au vu des résultats de la première étude de simulation, seule la méthode utilisant un modèle IRT longitudinal (Longitudinal Rasch Mixed model) a montré des résultats satisfaisants. Les méthodes estimant l'effet temps à partir d'estimations des valeurs individuelles du trait latent (Rasch and Mixed models et Plausible Values) se sont révélées inappropriées pour l'analyse des données dans le cadre de cette étude. En effet, ces méthodes présentaient des puissances peu élevées ainsi que des effets temps sous-estimés au contraire des méthodes SM et LRM.

Dans l'ensemble des études de simulation, les méthodes Longitudinal Rasch Mixed model, basée sur l'IRT, et Score and Mixed models, basée sur la CTT, ont montré des résultats similaires. Il semble donc que l'une ou l'autre approche puisse être utilisée à travers ces

méthodes pour l'analyse de PRO longitudinaux. En pratique, il est probable que la CTT sera privilégiée. La méthode SM est, en effet, simple à mettre en œuvre puisqu'elle repose sur un score observé analysé par un modèle linéaire mixte. De plus, l'interprétation des résultats en terme de score est plus simple et plus parlante pour les cliniciens que l'interprétation en terme de variable latente. Des travaux restent encore à faire dans ce domaine pour rendre plus accessible l'IRT dans le domaine de la santé tant au niveau des modèles utilisés que celui de l'interprétation en terme notamment de taille d'effet mesurée sur une variable latente. Le choix de l'utilisation d'une méthode basée sur un modèle de la famille de Rasch peut être motivé par leurs propriétés psychométriques : l'exhaustivité du score sur le trait latent, l'objectivité spécifique et surtout l'obtention d'une mesure d'intervalle.

La similitude des performances des deux approches CTT et IRT est retrouvée dans les travaux, toutefois encore peu nombreux, de la littérature. Lawson [55] fut un des premiers à comparer de manière empirique les deux approches. Il a observé sur 3 jeux de données différents que les paramètres d'items et d'individus obtenus avec la CTT et le modèle de Rasch étaient très comparables. Fan [31] a ensuite confirmé les résultats de cette première étude par la comparaison de la CTT avec plusieurs modèles IRT (modèle de Rasch, 2-PLM et 3-PLM) dans une analyse transversale de données réelles de tests de mathématiques et de lecture. Il a conclu que les estimations des paramètres d'items et des paramètres des individus étaient très comparables pour les deux approches, en particulier si un modèle de Rasch était utilisé. De plus, il s'est intéressé à l'invariance des paramètres. Dans le contexte de son étude, les paramètres d'items sont invariants de l'échantillon utilisé que ce soit en IRT ou en CTT. L'étude de Fan portait sur 40 échantillons sélectionnés aléatoirement, 80 échantillons sélectionnés par sexe (40 pour chaque sexe) et 80 échantillons sélectionnés par niveau de capacité (40 de bas niveau et 40 de haut niveau), tous composés de 1000 individus, dans une base de 193 000 individus. Cependant, bien qu'intéressantes, ces études comparent les deux approches uniquement sur l'observation de leur comportement sur des données réelles. MacDonald et Paunonen [65] ont mené une étude de simulation dans laquelle sont étudiées la comparabilité, l'invariance et la précision des paramètres selon l'approche utilisée (CTT, modèle de Rasch ou 2-PLM). Cette étude de simulation confirme les résultats des études de Lawson et Fan. Les paramètres de difficulté d'items et d'individus sont comparables et correctement estimés et les paramètres de difficulté d'items sont également invariants pour les deux approches.

Nous pouvons souligner que le cadre des trois études de simulation mises en place dans ce travail était favorable aux deux méthodes. En effet, le choix d'un espacement régulier des paramètres de difficulté d'items a assuré une bonne répartition des scores. On peut ainsi supposer que la distribution des scores était alors proche de la normalité, ce qui n'est pas toujours le cas. La robustesse de la méthode SM pourrait être étudiée. Par exemple, une mauvaise répartition des difficultés des items pourrait entraîner une mauvaise distribution des scores et remettre en cause l'hypothèse de normalité des scores. On peut se demander si la méthode SM serait plus pénalisée que la méthode LRM dans ce cas. De même, les données ont été simulées et analysées avec un modèle de Rasch alors que d'autre modèles IRT existent pour l'analyse d'items dichotomiques. Par exemple, la méthode LRM pourrait être impactée par le choix d'un modèle de Birnbaum pour l'analyse au lieu du modèle de Rasch.

Une analyse de sensibilité pourrait être mise en œuvre pour étudier le comportement des deux méthodes en cas d'écart aux hypothèses du modèle comme, par exemple, en cas de non-indépendance locale. On peut se demander dans quelle mesure les résultats seraient affectés. Par exemple, Glas et Hendrawan [44] ont étudié la robustesse des tests d'hypothèses des paramètres de régression d'un modèle linéaire sur une variable latente, dont le test de Wald. Ils ont montré que le risque de première espéce et la puissance du test en cas de non-indépendance locale pour un seul groupe de patients était très supérieure à ceux obtenus lorsque l'hypothèse d'indépendance locale n'était violée dans aucun groupe. On peut supposer qu'il est peu fréquent que la non-indépendance n'intervienne que dans un seul groupe. En revanche, la violation de l'hypothèse d'indépendance locale a peu d'impact si elle intervient sur l'ensemble des groupes. Il se pourrait donc que la méthode LRM soit peu sensible à la non-indépendance locale.

Gestion des données incomplètes

Le modèle linéaire mixte et le modèle de Rasch ont été utilisés dans ce travail en raison de leur capacité à gérer les données manquantes. Le modèle linéaire mixte a l'avantage d'être un modèle flexible adapté à l'analyse de données de données incomplètes à condition que les données manquantes soient de type MCAR ou MAR. Le modèle de Rasch, contrairement aux modèles de la théorie de réponse aux items de la famille de Lord comme le modèle 2-PLM, possède la propriété d'objectivité spécifique. L'estimation du trait latent est alors

indépendante de l'ensemble d'item utilisé pour la mesure. A condition que le modèle soit valide, les estimations de la variable latente sont consistantes pour tous les individus, même ceux dont les données sont incomplètes [6].

Données MNAR

Dans les études de simulation, les méthodes SM et LRM ont montré un comportement comparable face aux données manquantes. Comme attendu, ces méthodes sont adaptées à l'analyse de données manquantes ignorables. En revanche, les deux méthodes ne peuvent être utilisées lorsque les données manquantes sont non-ignorables et de type MNAR. Le processus de données manquantes doit alors être modélisé conjointement au modèle de mesure. Une piste pour adapter les méthodes SM et LRM à l'analyse de données MNAR pourrait être l'utilisation de modèles de mélange de schémas d'observation ou de modèles de sélection.
Holman et Glas [52] ont étudié la modélisation des données manquantes MNAR en IRT. Leur travail est basé sur l'idée que la probabilité d'avoir une réponse manquante dépend d'une variable latente tout comme la capacité de l'individu. Ils proposent un modèle combinant un modèle IRT multidimensionnel pour la modélisation du processus de données manquantes et un partial-credit model généralisé [79] comme modèle de mesure. Ce principe a été utilisé dans les études de simulations présentées dans ce travail lors de la simulation des données sujettes au dropout. En revanche, la méthode LRM utilisée pour l'analyse de ces données ne modélise pas le processus de données manquantes.
Le modèle d'analyse de données proposé par Holman et Glas est adapté à l'analyse de données transversales. Il serait intéressant de l'étendre à l'analyse de données longitudinales sujettes à des données manquantes MNAR. Le nombre de variables latentes dans le modèle sera important puisqu'aux variables latentes du modèle de mesure (autant que de temps de mesure) s'ajouteront les variables latentes modélisant le processus de données manquantes. Si seul le dropout est modélisé alors une variable latente modélisant le dropout sera incluse dans le modèle pour chaque temps de mesure. Mais si les données manquantes sont intermittentes, il est possible qu'il soit nécessaire d'inclure plusieurs variables latentes pour chaque temps de mesure. Le grand nombre de paramètres à estimer pourrait alors poser des problèmes de convergence.

Données MAR et ignorabilité

L'impact des données manquantes de type MCAR et MNAR a été étudié dans ce travail. Les performances des méthodes SM et LRM pourraient être comparées dans le cadre de données manquantes de type MAR. On peut s'attendre à ce que les méthodes SM et LRM présentent des résultats similaires aux résultats observés sur les données sujettes à du dropout de type MCAR, également ignorable. En effet, les méthodes SM et LRM sont valides pour l'analyse de données MAR si la condition de séparabilité est vérifiée. On peut donc s'attendre à une perte de puissance et des estimations de l'effet temps généralement non-biaisées.

La plupart du temps, l'hypothèse de données manquantes de type MCAR n'est pas satisfaisante. Il faut alors réussir à distinguer si les données manquantes sont de type MAR ou MNAR car cela aura un impact sur le choix du type d'analyse. Il a été montré que pour chaque modèle MNAR, ajusté à un jeu de données incomplètes, il existait un modèle équivalent MAR qui serait aussi bien ajusté à ce jeu de données [74]. Il n'est donc pas possible de distinguer le type de données manquantes à partir des données observées. De plus, si le modèle MNAR et son équivalent MAR s'ajustent de la même façon, ils peuvent produire des estimations différentes.

Lors de l'analyse de données incomplètes supposées MNAR, il est nécessaire de faire des hypothèses invérifiables sur le lien entre le processus de données manquantes et les données non-observées. De plus, les modèles d'analyse de données MNAR sont très sensibles aux hypothèses. Il est alors recommandé d'effectuer une analyse de sensibilité des résultats obtenus [23, 119]. Cette analyse permet d'évaluer la stabilité de l'inférence d'un modèle MNAR en fonction des hypothèses posées et de comparer les résultats sous l'hypothèse de données MNAR à ceux obtenus sous l'hypothèse de données MAR.

Données manquantes intermittentes

Les données des deux applications présentées étaient sujettes à du dropout et les études de simulation ont étudié l'impact du dropout sur les résultats des différentes méthodes d'analyse. Or, les données des applications étaient également sujettes à des données manquantes intermittentes. Il existe plusieurs types de données manquantes intermittentes. Lorsqu'un individu présente des questionnaires complètement non-remplis, on peut classer la non-réponse en trois catégories [24]. Si le patient ne rempli pas le questionnaire à partir d'un temps donné

et le patient sort de l'étude à ce temps, on est alors face à du dropout. Le patient peut présenter des questionnaires manquants jusqu'à un temps donné puis il peut l'avoir rempli à tous les temps suivants lorsqu'un patient a été inclus tardivement dans l'étude. Lorsque le questionnaire est manquant à un temps donné mais que le questionnaire a été rempli au temps suivant alors les données manquantes sont intermittentes.

Les données manquantes intermittentes peuvent également survenir au sein d'un questionnaire lorsque celui-ci n'a été que partiellement rempli [33]. Plusieurs raisons peuvent entraîner la survenue de données manquantes intermittentes sur les items. L'hypothèse d'un item manquant par pur hasard est plausible car le patient a pu simplement ne pas voir la question. Le contenu de l'item peut être en cause s'il est gênant ou tabou, parce qu'il a trait à la sexualité ou à la religion par exemple, et le patient préférera ne pas répondre à l'item. La difficulté de l'item peut également entraîner de la non-réponse. Un item adapté à des personnes en bonne santé peut être trop difficile pour des personnes malades ou l'inverse ce qui pose la question de la pertinence du choix du questionnaire pour la population étudiée. Par exemple, un item sur la capacité du patient à marcher sur une distance donnée peut générer des données manquantes si les patients interrogés ont d'importantes limitations physiques. Enfin, un item peut ne pas être rempli en raison du niveau du critère étudié. En effet, un patient dont la santé est dégradée cessera peut-être de remplir le questionnaire en cours de route à cause de la fatigue qu'il ressent par exemple.

La survenue de données manquantes intermittentes soulève des questions en terme d'analyse des données. Les résultats des analyses avec les méthodes SM et LRM pourraient être biaisés par la présence de données manquantes intermittentes. Tout comme pour le dropout, il est possible que les résultats différent en fonction du type de données manquantes, informatives ou non. La méthode SM peut être adaptée de différentes façons pour analyser ce type de données. Une imputation peut être réalisée au moment du calcul du score des patients. Cependant, plusieurs méthodes d'imputation sont possibles. Il convient de faire un choix éclairé entre ces différentes méthodes puisqu'elles peuvent également entraîner des biais.

De plus, les méthodes SM et LRM se comportaient de façon comparable face à du dropout. Le cas des données manquantes intermittentes pourrait aboutir à des résultats différents. En particulier, en raison de la propriété d'objectivité spécifique du modèle de Rasch, on peut

s'attendre à ce que la méthode LRM gère mieux les données manquantes intermittentes que la méthode SM. En effet, lorsque le nombre d'items manquants pour chaque questionnaire est important, le score risque de ne pouvoir être calculé et l'information même partielle sera perdue pour la méthode SM. La méthode LRM pourrait prendre en compte plus d'information que la méthode SM, sans imputation, puisque le modèle de Rasch utilise l'information disponible au niveau des items. De plus, la propriété d'objectivité spécifique pourrait assurer une estimation correcte du trait latent car cette estimation ne dépend pas de l'ensemble d'items utilisé pour la mesure. On peut également noter que, même en cas d'imputation des valeurs manquantes pour le calcul du score, il se peut que plus d'information ne soit perdue pour le score que pour la variable latente. En général, les recommandations pour l'imputation dans les questionnaires de qualité de vie sont d'imputer avec la méthode PMS (imputation de la moyenne des items remplis de l'individu) si au moins la moitié des items ont été remplis. Cette méthode peut ne pas être la plus adéquate. De plus, l'imputation peut biaiser les résultats si la méthode d'imputation est mal choisie.
L'impact de données manquantes intermittentes sur les deux méthodes pourrait donc être différent. Le choix d'imputer ou non et la méthode d'imputation utilisée pourrait aussi avoir un impact sur les performances de la méthode SM.
L'impact de la survenue de données manquantes intermittentes sur les performances des méthodes SM et LRM sera étudié dans la thèse d'Elodie Dumas au sein de l'EA 4275.

Items polytomiques

L'analyse de plusieurs dimensions du QLQ-C30 a confronté les méthodes SM et LRM à l'analyse d'items polytomiques. En pratique, les échelles couramment employées en santé sont principalement composées d'items polytomiques. Le SF-36 comprend 2 dimensions à items dichotomiques et 7 dimensions à items polytomiques. Toutes les dimensions de la QLQ-C30 et du WHOQOL-BREF sont composées d'items polytomiques. Si l'adaptation des méthodes à l'analyse d'items polytomiques est plutôt aisée, leur utilisation pourrait en revanche se révéler problématique.

Une adaptation naturelle de la méthode LRM, basée sur le modèle de Rasch, est de remplacer ce modèle par une de ses extensions aux items polytomiques, le rating-scale model (RSM) ou le partial credit-model (PCM). Ces modèles comprennent un nombre de paramètres plus

important que le modèle de Rasch qui augmente avec le nombre de modalités de la dimension étudiée. Le rating-scale model est plus parcimonieux que le partial-credit model car seul un paramètre d'ajustement pour chaque modalité positive est estimé, en plus des paramètres de difficulté des items. Cependant, il n'est pas adapté si le nombre de modalités d'un item à l'autre n'est pas constant au sein d'une dimension. Le nombre important de paramètres à estimer pourrait entraîner des problèmes de convergence du modèle, a fortiori si le PCM est utilisé. Pour pouvoir comparer les méthodes SM et LRM dans le cadre d'items polytomiques, il faudra d'abord se pencher sur les problèmes de convergence en IRT. Une solution envisageable serait de fixer les valeurs des paramètres de difficulté des items pour l'estimation de la variable latente. Ceci conduirait à diminuer le nombre de paramètres à estimer. Les valeurs des paramètres de difficulté des items pourrait être issues d'une banque d'items ou de précédentes études. L'impact de ce choix est à étudier tout comme l'impact d'une mauvaise hypothèse sur les valeurs des paramètres d'items sur les résultats de la méthode LRM.

Mesure du changement et response-shift

Lorsque l'évolution d'un PRO est étudiée, on fait souvent l'hypothèse que la perception qu'ont les patients du concept étudié ne va pas se modifier dans le temps. Or, les patients font face à une maladie et à des traitements. Les répercussions que peuvent avoir la maladie et ses traitements dans la vie du patient va généralement l'amener à s'adapter et peut changer sa perception du concept étudié. Dans le cadre d'études longitudinales, l'observation d'une évolution chez un patient peut mêler une réelle évolution à un changement de perception du patient, appelé *response-shift*. Le response-shift est défini comme "un changement dans la signification pour un individu de sa propre évaluation de qualité de vie résultant d'un changement de ses standards de mesure, ses valeurs et sa conception de la qualité de vie" [11]. Ainsi, le changement de référentiel interne d'un patient peut entraîner un changement de sa perception du PRO étudié. Trois composantes du response-shift sont distinguées :

- Le *réétalonnage* de l'échelle (recalibration) concerne un changement de référence vis à vis de l'instrument de mesure. Par exemple, un patient peut évaluer une douleur à 8 sur 10 sur une échelle de douleur (0 étant l'absence de douleur et 10 la pire douleur). Puis le patient peut subir une intervention lourde et douloureuse qu'il évaluera à 10. En raison d'un réétalonnage de l'échelle, il pourra alors réévaluer la première douleur ressentie avant l'intervention à 5 sur 10 au lieu de 8.

- La *redéfinition des priorités* (reprioritization) intervient lors d'un changement de l'importance relative des dimensions. Suite à un événement, le patient peut changer l'ordre d'importance qu'il accorde aux différents domaines constituant l'instrument de mesure tout en considérant toujours ces domaines comme importants. Par exemple, un patient peut initialement considérer que le fonctionnement physique est plus important que la fatigue. Après un problème de santé, ce patient peut redéfinir ses priorités. Il peut alors considérer que la sensation de fatigue est plus importante que le fonctionnement physique, tout en percevant ses deux aspects comme étant toujours importants.

- La *reconceptualisation* (reconceptualization) est une redéfinition du sens du concept mesuré. Des dimensions qui étaient importantes dans la définition du concept ont été supplantées par d'autres dimensions. Par exemple, une femme en bonne santé peut considérer que les aspects qui contribuent le plus à sa qualité de vie sont le fonctionnement physique et la santé mentale. Après un cancer, une reconceptualisation a pu avoir lieu. Cette femme pourra alors considérer que les aspects qui contribuent le plus à sa qualité de vie sont désormais la douleur et la fatigue.

Si la question du response-shift a d'abord été étudiée en management et en sciences de l'éducation, Sprangers et Schwartz [103, 96] ont introduit cette problématique dans l'étude de l'évolution de la qualité de vie en cancérologie dans les années 90. La question s'est alors posée de savoir comment détecter le response-shift et comment le prendre en compte dans les analyses. Dans le contexte dans lequel se place notre travail, les patients peuvent changer leur perception du PRO étudié. On peut se demander si les méthodes SM et LRM permettraient de détecter le response-shift et même de quantifier le response-shift.

Diverses méthodes ont été proposées au niveau de la planification des études pour anticiper le phénomène de response-shift. Parmi elles, le "then test" [104] est la méthode la plus couramment employée. Cette méthode propose de réaliser une évaluation "then test" en plus des évaluations pretest à l'inclusion et posttest pendant le suivi des patients. L'évaluation "then test" est une réévaluation rétrospective du pretest au même moment que le posttest. En pratique, une évaluation du PRO est faite à l'inclusion (pretest). Au deuxième temps d'étude, une évaluation classique du PRO est réalisée (posttest). Au même temps d'étude, il sera demandé au patient de réévaluer rétrospectivement le pretest, cette évaluation est appelée "then test". La différence entre le pretest et le "then test" est censée mesurer le response

shift. Même si cette méthode est couramment utilisée, plusieurs limites à son utilisation sont connues. Les capacités de mémoire des patients doivent être relativement bonnes pour qu'ils puissent réaliser le "then test". De plus, cette méthode mesure uniquement la composante de réétalonnage du response-shift et n'est adaptée qu'aux études comprenant deux temps de mesure. En dépit de ces inconvénients, cette méthode reste utilisée car elle permet d'évaluer le response-shift au niveau de l'individu et la mesure du "then test" serait plus valide que celle du pretest [11] pour mesurer le changement. Dans le cadre de notre travail, l'utilisation du "then-test" n'est pas envisageable car il n'est adapté qu'à des études à deux temps de mesure.

L'utilisation de méthodes statistiques permettant de détecter le response-shift au moment de l'analyse a également été proposée avec notamment l'analyse factorielle, la détection de DIF avec le modèle de Rasch ou l'analyse de courbes de croissance. Mais, ces méthodes présentent l'inconvénient d'évaluer le response-shift au niveau du groupe et non au niveau de l'individu. L'ensemble des méthodes, qu'elles se situent au niveau de la planification ou de l'analyse, ne permettent de détecter que certaines composantes du response-shift. La question du traitement du response-shift reste ouverte et beaucoup de pistes sont à explorer [11]. Il semble que les méthodes SM et LRM ne puissent pas encore intégrer l'étude du response-shift. Cependant, si cette problématique est nouvelle, il est certain qu'elle sera de plus en plus traitée à l'avenir au vu des nombreux travaux de ces dernières années. Les méthodes présentées dans ce travail devront alors être adaptées en conséquence.

Conclusion

L'évaluation de l'évolution des Patient-Reported Outcomes en santé s'est développée ces dernières années. L'analyse de ces données se trouve confrontée au choix entre plusieurs approches. Face au manque de littérature pouvant éclairer ce choix, nous avons comparé la théorie classique des tests et la théorie de réponse aux items à travers différentes études de simulation. Le cadre de ces études était des données de type PRO recueillies de manière longitudinales, potentiellement sujettes à du dropout. L'intérêt portait sur la capacité des méthodes à détecter des effets temps et groupe à travers le risque de première espèce et la puissance et à estimer correctement les effets en évaluant le biais. Ce travail a d'ores et déjà permis d'écarter deux méthodes non satisfaisantes pour l'analyse de PRO longitudinales. Ces méthodes estimaient l'effet temps à partir d'estimations des valeurs individuelles du trait latent (Rasch and Mixed models et Plausible Values) et présentaient des puissances peu élevées ainsi que des effets temps sous-estimés au contraire des méthodes SM et LRM. Les méthodes SM (basée sur la CTT) et LRM (basée sur l'IRT) ont été validées pour l'analyse de données complètes et de données sujettes à du dropout MCAR. Ce travail ouvre de nombreuses perspectives en terme d'étude de l'impact des données manquantes intermittentes ainsi que d'adaptation des méthodes à l'analyse de données MNAR et d'items polytomiques. Plus globalement, l'IRT est très utilisé dans les sciences de l'éducation mais reste encore assez méconnu dans le domaine de la santé où cette théorie reste souvent cantonnée à la construction, la validation ou la réduction de questionnaires. Il semble que des efforts doivent être faits dans l'aide à l'interprétation des résultats et la diffusion de l'IRT dans le domaine de la santé.

Annexes

Annexe A

Article : Comparison of CTT and Rasch-based approaches for the analysis of longitudinal Patient Reported Outcomes

Research Article

Statistics in Medicine

Received 26 October 2009, Accepted 1 November 2010 Published online 28 December 2010 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4153

Comparison of CTT and Rasch-based approaches for the analysis of longitudinal Patient Reported Outcomes

Myriam Blanchin,[a*†] Jean-Benoit Hardouin,[a] Tanguy Le Neel,[a] Gildas Kubis,[a] Claire Blanchard,[b] Eric Mirallié[b] and Véronique Sébille[a]

Health sciences frequently deal with Patient Reported Outcomes (PRO) data for the evaluation of concepts, in particular health-related quality of life, which cannot be directly measured and are often called latent variables. Two approaches are commonly used for the analysis of such data: Classical Test Theory (CTT) and Item Response Theory (IRT). Longitudinal data are often collected to analyze the evolution of an outcome over time. The most adequate strategy to analyze longitudinal latent variables, which can be either based on CTT or IRT models, remains to be identified. This strategy must take into account the latent characteristic of what PROs are intended to measure as well as the specificity of longitudinal designs. A simple and widely used IRT model is the Rasch model. The purpose of our study was to compare CTT and Rasch-based approaches to analyze longitudinal PRO data regarding type I error, power, and time effect estimation bias. Four methods were compared: the Score and Mixed models (SM) method based on the CTT approach, the Rasch and Mixed models (RM), the Plausible Values (PV), and the Longitudinal Rasch model (LRM) methods all based on the Rasch model. All methods have shown comparable results in terms of type I error, all close to 5 per cent. LRM and SM methods presented comparable power and unbiased time effect estimations, whereas RM and PV methods showed low power and biased time effect estimations. This suggests that RM and PV methods should be avoided to analyze longitudinal latent variables. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: Item Response Theory; Classical Test Theory; Patient Reported Outcomes; longitudinal data; simulation study

1. Introduction

Patient Reported Outcomes (PRO) data are widely used in health sciences to evaluate concepts, such as health-related quality of life (HRQoL), pain, fatigue, or anxiety [1], which are often referred to as latent variables because they cannot be directly observed from patients. PRO data are evaluated using the answers of patients to items often grouped into several dimensions in a questionnaire. Two approaches are commonly used for the analysis of such data: Classical Test Theory (CTT) and Item Response Theory (IRT). The CTT is an approach based on the computation of a score usually computed as the sum of the item responses. This score is an estimation of a 'true' score assumed to represent the evaluated outcome (e.g. HRQoL). The observed and true scores are assumed to be linked by a linear relation. In IRT, item responses have a central role. The probability to answer to an item is a function (not necessary linear) of the latent variable which represents the evaluated outcome. IRT models are a large family

[a] *EA 4275 'Biostatistics, Clinical Research and Subjective Measures in Health Sciences', Faculty of Pharmaceutical Sciences, University of Nantes, Nantes, France*

[b] *Department of Digestive and Endocrine Surgery/Institut des Maladies de l'Appareil Digestif, CHU Nantes, Faculty of Medicine, University of Nantes, France*

[*] *Correspondence to: Myriam Blanchin, EA 4275 'Biostatistics, Clinical Research and Subjective Measures in Health Sciences', Faculté de Pharmacie—Université de Nantes, 1, rue Gaston Veil—44035 Nantes Cedex 1, France.*

[†] *E-mail: myriam.blanchin@univ-nantes.fr*

of models relying generally on the same three assumptions: unidimensionality, monotonicity, and local independence. A simple and widely used IRT model is the Rasch model [2] or one-parameter logistic model (1-PLM). The Rasch model is adapted to the analysis of dichotomous items and models the probability of a response to an item through a person parameter (person ability) and an item parameter (item difficulty). Nowadays, a large proportion of scales are developed and validated using models of the Rasch family due to their interesting psychometrics properties including the exhaustivity of the score on the latent trait and the specific objectivity.

Longitudinal data are often collected in order to analyze the evolution of an outcome over time. In this case, the correlation between measurements from each patient over time has to be taken into account. Linear mixed models [3] are commonly used to analyze such data.

In the case of longitudinal PRO data, the latent characteristic of what PRO are intended to measure as well as the specificity of longitudinal designs with repeated measurements should probably both be taken into account to provide reliable analysis. To date, the choice of a statistical strategy for the analysis of such data is usually based on CTT rather than on IRT and seems to more likely rely on the researcher's practice and familiarity with CTT than on scientific grounds. Hence, the most adequate strategy to analyze longitudinal latent variables, which can be either based on the CTT or IRT approach, remains to be identified. The purpose of our study was to compare a CTT- and three Rasch-based approaches to analyze longitudinal PRO data regarding type I error, power and time effect estimation bias. Data from the evaluation of quality of life of patients with primary hyperparathyroidism were used to illustrate simulation results.

2. Methods

2.1. Statistical models

2.1.1. The Rasch model. The Rasch model [2, 4] came from psychometrics and was developed for achievement tests. Later it came to be used in health sciences, in particular for construction, validation, and reduction of questionnaires [5, 6]. In this framework, the responses to the items of a questionnaire are assumed to be the manifestation of a latent variable which cannot be directly observed. The Rasch model proposes to model the relationship between responses to dichotomous items and the latent variable, denoted θ. Let Y_{ij} be the dichotomous variable representing the response of person i $(i=1\dots N)$ to an item j $(j=1\dots J)$.
For a questionnaire containing J dichotomous items, the model can be written as follows:

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

where $y=0$ for a negative response (the most pejorative response) and $y=1$ for a positive response. δ_j is called the difficulty or item parameter and is associated with item j . The personal parameter θ_i is the individual value of the latent trait for patient i and represents the ability of the patient (e.g. HRQoL).

The Rasch model relies on three hypotheses:

- *Unidimensionality*: A unique latent variable explains the response to the items.
- *Monotonicity*: The probability of a positive response to an item is a non-decreasing function of the latent variable.
- *Local independence*: Given an individual, the item responses are independent of one another.

A Rasch model, where the latent trait θ is considered as a random variable and usually has a normal distribution $\mathrm{N}(\mu,\sigma^2)$, is a mixed-effects logistic model [7]. The parameters to be estimated in the model are μ and σ^2 to characterize the distribution of the latent trait θ and the difficulty parameters δ_j $(j=1\dots J)$. They can be jointly estimated using marginal maximum likelihood (MML). The marginal likelihood is expressed as

$$L(\delta_1,\dots,\delta_J,\mu,\sigma^2|\mathbf{y})=\prod_{i=1}^{N}\int\prod_{j=1}^{J}\frac{\exp(y_{ij}(\theta-\delta_j))}{1+\exp(\theta-\delta_j)}G(\theta/\mu,\sigma^2)\,\mathrm{d}\theta \tag{2}$$

with $G(./\mu,\sigma^2)$ the normal distribution function with mean μ and variance σ^2.

In order to ensure the identifiability of the model, one constraint has to be adopted. In general, the mean of the latent trait or the sum of difficulty parameters is assumed to be equal to 0.

2.1.2. Longitudinal mixed Rasch model. For repeated measures data, a longitudinal form of the Rasch model has been developed [8] in the field of educational and psychological testing. The interest of researchers using learning tests was to measure learning ability as well as its evolution. A longitudinal mixed Rasch model for modeling of quality of life evolution was derived from the modeling of learning and change, considering the latent variable θ as a random variable rather than as fixed parameters.

For a questionnaire containing J dichotomous items and measures repeated T times for each person i, the probability of a response to an item j at time t can be written as follows:

$$P(Y_{ij}^{(t)}=y^{(t)}|\theta_i^{(t)};\delta_j)=\frac{\exp(y^{(t)}(\theta_i^{(t)}-\delta_j))}{1+\exp(\theta_i^{(t)}-\delta_j)} \tag{3}$$

Item parameters $\delta_j, j=1\dots J$ are assumed to be constant with time meaning that the characteristics of the questionnaire are assumed not to vary through time.

The marginal likelihood is expressed as

$$L(\delta_1,\dots,\delta_J,\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathbf{y})=\prod_{i=1}^{N}\int_{\mathbb{R}^T}\prod_{t=1}^{T}\prod_{j=1}^{J}\frac{\exp(y_{ij}^{(t)}(\theta^{(t)}-\delta_j))}{1+\exp(\theta^{(t)}-\delta_j)}G(\boldsymbol{\theta}/\boldsymbol{\mu},\boldsymbol{\Sigma})d\boldsymbol{\theta} \tag{4}$$

with $G(./\boldsymbol{\mu},\boldsymbol{\Sigma})$ the multivariate normal distribution function with mean vector $\boldsymbol{\mu}=(\mu_1\dots\mu_T)'$ and covariance matrix $\boldsymbol{\Sigma}$.

A constraint is needed to ensure identifiability of the longitudinal mixed Rasch model. The usual constraint made on the parameters is $\mu_1=0$.

2.1.3. Covariance pattern models. A simple way to analyze repeated measures data where the focus is on the mean responses and their evolution with time is to use a covariance pattern model [9, 10]. In linear mixed models, random effects are used to model individual variation around the mean trajectory when the primary interest is in individual trajectories. Covariance pattern models contain only fixed effects that characterize the mean behavior of the population which is our primary interest. Moreover, this type of mixed model allows to specify a pattern for the correlation between measurements from the same patient.

Let

- n_i be the number of observations on patient i, $i=1\dots N$
- p be the number of parameters
- $\boldsymbol{Y}_i$ be the $(n_i\times 1)$ vector containing the responses for the patient i
- $\boldsymbol{\beta}$ be the $(p\times 1)$ vector of fixed effects parameters
- $\boldsymbol{X}_i$ be the $(n_i\times p)$ design matrix
- $\boldsymbol{e}_i$ be the $(n_i\times 1)$ vector of error terms, characterizing the overall variation and measurement error
- $\boldsymbol{\Sigma}_i$ be the $(n_i\times n_i)$ covariance matrix of error terms
- $M=\sum_{i=1}^{N}n_i$ be the total number of observations.

A covariance pattern model can be written as follows:

$$\begin{aligned}\boldsymbol{Y}_i &= \boldsymbol{X}_i\boldsymbol{\beta}+\boldsymbol{e}_i\\ \text{var}(\boldsymbol{e}_i) &= \boldsymbol{\Sigma}_i\\ \boldsymbol{Y}_i &\sim N_{n_i}(\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Sigma}_i)\end{aligned} \tag{5}$$

The parameters to be estimated in the model are $\boldsymbol{\beta}$ that characterizes the mean and $\boldsymbol{\omega}$ that characterizes $\boldsymbol{\Sigma}_i$. For example, $\boldsymbol{\omega}=(\sigma^2,\rho)$ for an AR(1) structure of $\boldsymbol{\Sigma}_i$. Two methods based on likelihood maximization are used to estimate unknown parameters: Maximum Likelihood (ML) or REstricted Maximum Likelihood (REML) estimation methods. The use of ML estimation leads to unbiased estimate for $\boldsymbol{\beta}$ but $\boldsymbol{\omega}$ is known to be biased when N is not too large. In REML method, the likelihood is modified to include an extra term for correction of the bias on $\boldsymbol{\omega}$ parameter.

Let $\boldsymbol{Y}$ be the $(M\times 1)$ vector summarizing the vectors $\boldsymbol{Y}_i (i=1,\ldots,N)$ into one vector. The joint density of $\boldsymbol{Y}$ is expressed as

$$f(\boldsymbol{y})=\prod_{i=1}^{N}(2\pi)^{-n_i/2}|\boldsymbol{\Sigma}_i|^{-1/2}|\boldsymbol{X}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i|^{-1/2}\exp\{-(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta})/2\} \tag{6}$$

The estimator of $\boldsymbol{\omega}$ resulting from REML estimation is known to be less biased than the estimation based on ML [3].

2.2. *Comparison of methods*

Four methods to analyze longitudinal PRO data have been compared: (i) Score and Mixed models (SM), (ii) Rasch and Mixed models (RM), (iii) Plausible Values (PV), and (iv) Longitudinal Rasch model (LRM).

Score and Mixed models. The SM method, corresponding to the CTT approach, consisted in calculating a score by summing the item responses for each patient. A linear mixed model was then used to explain the evolution of score with time.

Four covariance structures $\boldsymbol{\Sigma}$ are often used with longitudinal data: UN (unstructured), ARH(1) (heterogeneous first-order autoregressive), AR(1) (first-order autoregressive), or CSH (heterogeneous compound symmetry). The unstructured matrix is the most general possible structure. It is used when no hypotheses can be made on the structure of the covariance matrix but lead to estimate an important number of parameters. The ARH(1) structure takes into account the correlation between measures in time. Correlations are assumed to decrease when measures get further apart from each other in time. The use of the AR(1) structure makes an extra hypothesis compared to the ARH(1) structure: the variances are assumed to be equal. On the contrary, the choice of a CSH structure assumes that the variances are not equal but the correlation is constant over time.

The simulated datasets contained only one group and balanced data measured on three different occasions. In the presence of a single group, $n_i=3\ \forall i$ and $\boldsymbol{\Sigma}_i=\boldsymbol{\Sigma}$ could be assumed to be the same for all patients.

The model can be written as

$$\begin{aligned}\boldsymbol{S}_i &= \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{e}_i\\ \mathrm{var}(\boldsymbol{e}_i) &= \boldsymbol{\Sigma}\\ \boldsymbol{S}_i &\sim N_{n_i}(\boldsymbol{X}\boldsymbol{\beta},\boldsymbol{\Sigma})\end{aligned} \tag{7}$$

where $S_i^{(t)}=\sum_j y_{ij}^{(t)}$ for $t=(1,2,3)$ and

$$\boldsymbol{\Sigma}=\begin{pmatrix}\sigma_1^2 & \sigma_{12} & \sigma_{13}\\ \sigma_{12} & \sigma_2^2 & \sigma_{23}\\ \sigma_{13} & \sigma_{23} & \sigma_3^2\end{pmatrix}\quad\text{for UN}\qquad \boldsymbol{\Sigma}=\begin{pmatrix}\sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho^2\\ \sigma_1\sigma_2\rho & \sigma_2^2 & \sigma_2\sigma_3\rho\\ \sigma_1\sigma_3\rho^2 & \sigma_2\sigma_3\rho & \sigma_3^2\end{pmatrix}\quad\text{for ARH(1)}$$

$$\boldsymbol{\Sigma}=\sigma^2\begin{pmatrix}1 & \rho & \rho^2\\ \rho & 1 & \rho\\ \rho^2 & \rho & 1\end{pmatrix}\quad\text{for AR(1)}\qquad \boldsymbol{\Sigma}=\begin{pmatrix}\sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho\\ \sigma_1\sigma_2\rho & \sigma_2^2 & \sigma_2\sigma_3\rho\\ \sigma_1\sigma_3\rho & \sigma_2\sigma_3\rho & \sigma_3^2\end{pmatrix}\quad\text{for CSH}$$

Mean parameters $\boldsymbol{\beta}$ and covariance parameters $\boldsymbol{\omega}$ were estimated using the REML method in order to reduce the bias on covariance parameters. As would be performed on real data, AIC for each covariance matrix structure were compared to choose the adequate structure.

An estimate of $\boldsymbol{\mu}$ could be given by $\hat{\boldsymbol{\mu}}=(\hat{\mu}_1\quad\hat{\mu}_2\quad\hat{\mu}_3)'=\boldsymbol{X}\hat{\boldsymbol{\beta}}$. The time effect between two consecutive measures $(t=(1,2))$ was $d_{t,t+1}=\mu_{t+1}-\mu_t$. The time effect between time 1 and time 2 was estimated as $\hat{d}_{12}=\hat{\mu}_2-\hat{\mu}_1$.

The test of a time effect used an approximate F-test:

$$H_0:\mu_1=\mu_2=\mu_3=\mu\Leftrightarrow\beta_1=\beta_2=\beta_3\Leftrightarrow L\boldsymbol{\beta}=0$$

$$H_1:\exists i\,|\,\mu_i\neq\mu\Leftrightarrow L\boldsymbol{\beta}\neq 0$$

Define

$$\boldsymbol{L}=\begin{pmatrix}1 & -1 & 0\\ 1 & 0 & -1\end{pmatrix}$$

Under H_0, $F_L=(\boldsymbol{L}\hat{\boldsymbol{\beta}})'(\boldsymbol{L}\hat{\boldsymbol{V}}_\beta\boldsymbol{L}')^{-1}\boldsymbol{L}\hat{\boldsymbol{\beta}}/\text{rank}(\boldsymbol{L})$ has approximately an $F_{r,\text{df}}$ distribution where the numerator degrees of freedom, r, equals the rank of $\boldsymbol{L}$, df is the appropriate denominator degrees of freedom and $\hat{\boldsymbol{V}}_\beta=(\sum_{i=1}^{N}\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X})^{-1}$ is the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$.

Rasch and mixed models and plausible values. The Rasch and Mixed models (RM) and the Plausible Values (PV) methods were performed in two steps. In the first step, a mixed Rasch model was applied on the overall sample and an individual value of the latent trait was estimated for each patient at each time point. A linear mixed model was then fitted to explain the evolution of the estimated latent traits with time. A covariance pattern model was used as it was for the score.

First step of RM and PV methods. The parameters μ and σ^2, which characterize the distribution of θ, and δ_j, $j=1\ldots J$ were estimated using MML estimation from a mixed Rasch model [4]. Item parameters δ_j were constant with time.

RM and PV differ in the method used to estimate individual values of the latent trait. The *Bayes Expected A Posteriori* estimator (EAP), used in the first step of the RM method, is a point estimate defined as the mean of the posterior distribution [11]:

$$\text{EAP}(\theta_i)=E(\theta|\boldsymbol{Y},\boldsymbol{\delta},\mu,\sigma^2)=\frac{\int\theta B(\theta)\exp(\theta s_i)G(\theta|\mu,\sigma^2)\,\text{d}\theta}{\int B(\theta)\exp(\theta s_i)G(\theta|\mu,\sigma^2)\,\text{d}\theta}$$

where $\boldsymbol{\delta}$ was the vector of difficulty parameters, s_i the observed raw score for patient i ($s_i=\sum_j y_{ij}$), $B(\theta)=\prod_j[1+\exp(\theta-\delta_j)]^{-1}$, and $G(./\mu,\sigma^2)$ the normal distribution function with mean μ and variance σ^2. Owing to the use of EAP estimates in RM method, all patients with the same total score and hence the same posterior distribution will have the same estimated value of the latent trait $\hat{\theta}$.

The PV method consisted in first estimating a value for the latent variable by plausible value imputation. The plausible value imputation is based on the multiple imputation theory of Rubin [12] and is used in large-scale educational surveys, such as PISA and NAEP [13, 14]. Instead of using the mean of the posterior distribution, a plausible value ($\hat{\theta}$) is randomly drawn from the posterior distribution. This method allows patients with the same total score (and hence the same posterior distribution) to have different plausible values. Usually, several draws of plausible values are used. The same analysis is made on each draw and results of all the analyses are pooled to obtain an estimate of the parameter of interest and an estimate of its variance as in multiple imputation. The multiple draws allow to obtain an estimate of the uncertainty due to the estimate of θ. If this uncertainty has to be taken into account in the analysis but has not to be explicitly estimated, one draw of plausible values is sufficient [15]. Furthermore, Wu [16] has shown that one plausible value could be sufficient to adequately recover population parameters. In health sciences, studies are focused on the evolution of the population and not on the individual trajectories.

Given the item response pattern $\boldsymbol{y}$ and the latent variable θ, $f(\boldsymbol{y}|\theta)$ is the item response probability of the Rasch model also called item response model. Assuming that θ comes from a normal distribution, $g(\theta)\sim\text{N}(\mu,\sigma^2)$ is called the population model. The posterior distribution $h(\theta|\boldsymbol{y})$ of an individual with item response pattern $\boldsymbol{y}$ is defined as

$$h(\theta|\boldsymbol{y})=\frac{f(\boldsymbol{y}|\theta)g(\theta)}{\int f(\boldsymbol{y}|\theta)g(\theta)\,\text{d}\theta} \tag{8}$$

Plausible values are randomly drawn from the posterior distribution with density $h(\theta|\boldsymbol{y})$. As the EAP estimator is defined as the mean of the posterior distribution, we use the EAP estimates and its standard errors to draw the plausible values $\hat{\theta}_i^{(t)}$ for each person i at each time point t from a normal distribution with mean equal to the EAP estimate of θ of the person i and standard error equal to the corresponding estimated standard error.

Second step of RM and PV methods. For both methods, the linear mixed model was expressed as

$$\begin{aligned} \hat{\theta}_i &= X\beta + e_i \\ \text{var}(\boldsymbol{e}_i) &= \boldsymbol{\Sigma} \\ \hat{\theta}_i &\sim N_{n_i}(\boldsymbol{X\beta}, \boldsymbol{\Sigma}) \end{aligned} \tag{9}$$

Mean parameters $\boldsymbol{\beta}$ and covariance parameters $\boldsymbol{\omega}$ were estimated using the REML method. As for the SM method, four structures were investigated for $\boldsymbol{\Sigma}$: UN, ARH(1), AR(1), and CSH.

An estimate of $\boldsymbol{\mu}$ could be given by $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1 \ \ \hat{\mu}_2 \ \ \hat{\mu}_3)' = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. The time effect between time 1 and time 2 was estimated by $\hat{d}_{12} = \hat{\mu}_2 - \hat{\mu}_1$.

The test of a time effect used an approximate F-test such as for the SM method.

Longitudinal mixed Rasch model. The method LRM was based on a longitudinal Rasch model that estimated the time effect, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of the latent trait in the same step.

The longitudinal mixed Rasch model, used to estimate $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and δ_j, $j = 1 \ldots J$, was expressed as

$$P(Y^{(t)} = y^{(t)} | \theta^{(t)}; \delta_j) = \frac{\exp(y^{(t)}(\theta^{(t)} - \delta_j))}{1 + \exp(\theta^{(t)} - \delta_j)} \tag{10}$$

The estimation was based on MML where the marginal likelihood was expressed as

$$L(\delta_1, \ldots, \delta_J, \boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{y}) = \prod_{i=1}^{N} \int_{\mathbb{R}^T} \prod_{t=1}^{T} \prod_{j=1}^{J} \frac{\exp(y_{ij}^{(t)}(\theta^{(t)} - \delta_j))}{1 + \exp(\theta^{(t)} - \delta_j)} G(\boldsymbol{\theta} / \boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathrm{d}\boldsymbol{\theta} \tag{11}$$

with $G(./\boldsymbol{\mu}, \boldsymbol{\Sigma})$ the multivariate normal distribution function with mean vector $\boldsymbol{\mu} = (\mu_1 \cdots \mu_T)'$ and covariance matrix $\boldsymbol{\Sigma}$ of unstructured type.

Constraint of the nullity of latent variable mean at time 1 was used to ensure identifiability. Owing to this constraint, $\hat{\mu}_2$ and $\hat{\mu}_3$ represented, respectively, $\hat{d}_{12}$ and $\hat{d}_{13}$.

The test of a time effect used an approximate Wald test:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu \Leftrightarrow L\boldsymbol{\mu} = 0$$

$$H_1 : \exists i \,|\, \mu_i \neq \mu \Leftrightarrow L\boldsymbol{\mu} \neq 0$$

Define

$$\boldsymbol{L} = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}$$

Under H_0, $T_L = (\boldsymbol{L}\hat{\boldsymbol{\mu}})'(\boldsymbol{L}\hat{\boldsymbol{V}}\boldsymbol{L}')^{-1}\boldsymbol{L}\hat{\boldsymbol{\mu}}$ has approximately a χ^2_r distribution, where $r =$ rank of $\boldsymbol{L}$ and $\hat{\boldsymbol{V}}$ is the estimated covariance matrix.

2.3. *Simulation of data*

Responses of patients to dichotomous items in a repeated measures setting were simulated with a longitudinal mixed Rasch model including three times of assessment according to equation (3). The latent trait vector $\theta = (\theta^{(1)} \ \ \theta^{(2)} \ \ \theta^{(3)})'$ had a multivariate normal distribution with

$$\boldsymbol{\mu} = (\mu_1 \ \ \mu_2 \ \ \mu_3)' \quad \text{and} \quad \boldsymbol{\Sigma} = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix}$$

corresponding to a first-order autoregressive structure for the covariance matrix. The correlation between measures decreased as measures got farther apart from each other in time. The latent trait was considered as having the same variance at each time point ($\sigma^2 = 1$). Three different values for the correlation coefficient of the latent trait between two consecutive times ρ were used to simulate data: $\rho = 0.4$ (small correlation), $\rho = 0.7$, and $\rho = 0.9$ (high correlation).

The time effect between two consecutive measures ($t=(1,2)$) was $d_{t,t+1}=\mu_{t+1}-\mu_t$ and $ES_{t,t+1}$ was the effect size between two consecutive measures. $ES_{t,t+1}=d_{t,t+1}/\sigma\ \forall t=(1,2)$. Under H_0, $ES_{t,t+1}=0$. Under H_1, $ES_{t,t+1}=0.2$.

The data were assumed to come from a 4-item scale or a 7-item scale with dichotomous items. The values of difficulty parameters were $\delta_1=-1$, $\delta_2=-0.5$, $\delta_3=0.5$, $\delta_4=1$ for a 4-item scale and $\delta_1=-1.5$, $\delta_2=-1$, $\delta_3=-0.5$, $\delta_4=0$, $\delta_5=0.5$, $\delta_6=1$, $\delta_7=1.5$ for a 7-item scale. The sample size could be of 100 or 200 individuals. The different values of the number of items, the number of individuals, and the correlation led to consider 24 different cases. Five hundred simulated datasets were generated and analyzed for each case.

Studied criteria. In order to compare the methods to analyze longitudinal PRO data, three criteria were studied: the type I error, the power and the bias of the time effect estimation.

The type I error of the tests was classically computed as the proportion of rejection of H_0 under the null hypothesis. Rejection of H_0 was based on a test of simultaneous equality of mean estimations, i.e. the absence of time effect.

The power calculation used the same tests but calculated the proportion of rejection of H_0 under the alternative hypothesis.

RM, PV, and LRM methods were based on IRT models, hence the calculation of time effect estimation bias was possible. The estimated value of time effect $\hat{d}_{t,t+1}$ was compared with the fixed value $d_{t,t+1}$ used for data simulation. As SM method was based on the classical approach (CTT) and IRT was used to simulate data, the true value of the time effect on the score scale was not known. But since under H_0, no time effect was assumed on the latent variable, no time effect was expected on the score scale under H_0 as well and hence the time effect bias could be calculated under H_0. Under H_1, the true value of time effect was not known on the score scale hence the time effect bias could not be assessed.

The mean of time effect estimation for each case was compared to the true value using a t-test.

Simulations and analyses were performed using SAS 9.1 [17] and Stata 10 [18, 19].

3. Simulation results

For each of the 24 cases, the AIC for the four structures of covariance matrix for mixed models were compared for RM and SM methods.

For the SM, when $\rho=0.4$ or $\rho=0.7$, the AIC was more often minimized by the choice of an AR(1) structure for the covariance matrix. When $\rho=0.9$, the CSH structure more often minimized the AIC of the models than the other structures. The results further presented are from covariance pattern models estimated through the REML method with an AR(1) structure for the covariance matrix.

For the RM method, the AIC was also more often minimized by the choice of an AR(1) structure for the covariance matrix in most of the cases under H_0. The results presented further for the RM method come from analyses with an AR(1) structure.

The LRM method used an unstructured covariance matrix.

3.1. Type I error of the tests

Table I and Figure 1 show the type I error for each method for different values of the parameters: sample size, number of items, and latent variable correlation. All type I errors were close to 5 per cent. All methods showed comparable results whatever the value of the parameters. No ρ, J, or N effects were observed on the type I error values.

All 95 per cent confidence intervals included the target value of 5 per cent. Type I error ranged from 3.6 to 6.0 per cent for the SM method, from 3.4 to 5.8 per cent for the RM method, from 3.8 to 6.2 per cent for the PV method, and from 4.4 to 6.7 per cent for the LRM method.

3.2. Power of the tests

Table I and Figure 2 show the power for each method for different values of the parameters: sample size, number of items, and latent variable correlation. Whatever the value of these parameters, the powers of the LRM method were comparable to those of the SM method. We denoted that the LRM method presented a power a little higher than the SM method whatever the value of correlation coefficient ρ. Moreover, LRM and SM methods achieved much larger power than the RM and PV

Table I. Type I error and power of the tests for Score Mixed model (SM), Rasch Mixed model (RM), Plausible Values (PV), and Longitudinal Rasch Mixed model (LRM) methods for different values of sample size (N), number of items (J), and latent variable correlation (ρ). Results from analyses with an AR(1) structure for the covariance matrix in RM, PV, and SM methods and an unstructured covariance matrix for LRM method.

			SM		RM		PV		LRM	
N	J	ρ	Type I error	Power	Type I error	Power	Type I error	Power	Type I error	Power
100	4	0.4	0.054	0.384	0.052	0.146	0.062	0.166	0.054	0.388
		0.7	0.040	0.372	0.040	0.120	0.042	0.142	0.044	0.423
		0.9	0.038	0.424	0.042	0.156	0.056	0.156	0.048	0.508
	7	0.4	0.036	0.428	0.034	0.108	0.052	0.162	0.044	0.470
		0.7	0.046	0.522	0.048	0.096	0.050	0.150	0.058	0.568
		0.9	0.044	0.564	0.044	0.070	0.054	0.160	0.046	0.688
200	4	0.4	0.060	0.634	0.058	0.280	0.058	0.288	0.052	0.654
		0.7	0.048	0.678	0.048	0.276	0.060	0.304	0.056	0.721
		0.9	0.060	0.756	0.060	0.250	0.052	0.330	0.067	0.826
	7	0.4	0.050	0.810	0.048	0.184	0.042	0.304	0.050	0.822
		0.7	0.036	0.880	0.038	0.154	0.040	0.314	0.048	0.916
		0.9	0.048	0.918	0.048	0.148	0.038	0.306	0.058	0.955

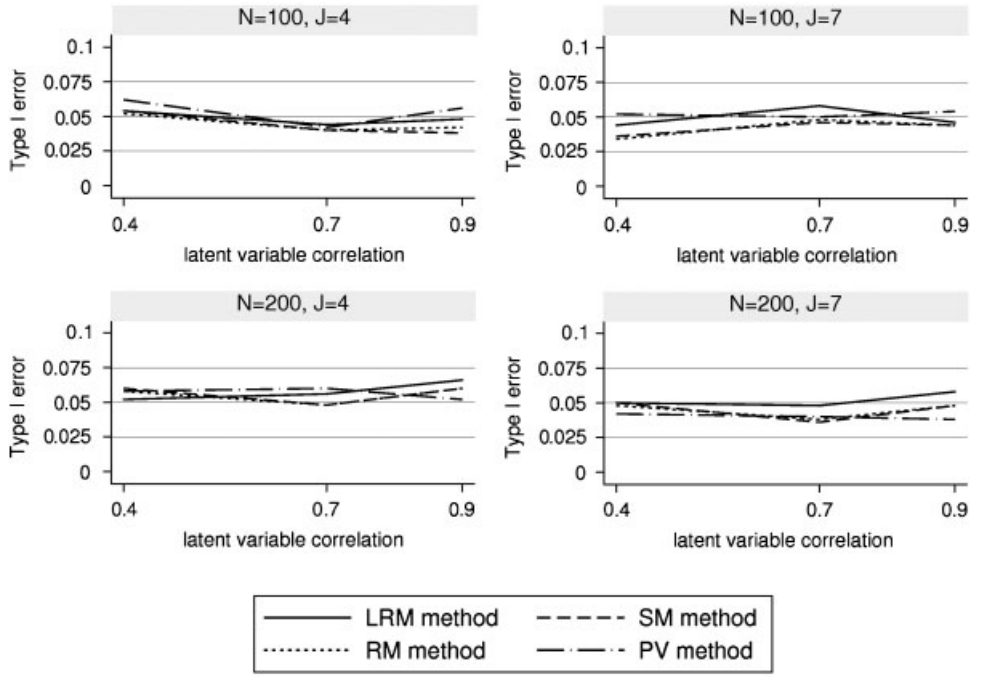

Figure 1. Type I error of the tests for Score Mixed model (SM), Rasch Mixed model (RM), Plausible Values (PV), and Longitudinal Rasch Mixed model (LRM) methods for different values of sample size (N), number of items (J), and latent variable correlation (ρ). Results from analyses with an AR(1) structure for the covariance matrix in RM, PV, and SM methods and an unstructured covariance matrix for LRM method.

methods. Differences in power were the highest when $N=200$ and $J=7$. In this case, LRM and SM powers were all higher than 80 per cent, whereas RM power ranged from 15 to 18 per cent and PV power were close to 30 per cent.

For LRM and SM methods, the power increased with the rise of the correlation between two measures, the number of items or the number of individuals. For example, power of the SM and LRM methods were close to 38 per cent when $N=100$, $J=4$, and $\rho=0.4$. Power was much higher, greater than 90 per cent, when $N=200$, $J=7$, and $\rho=0.9$.

Power from the RM and PV methods was quite stable whatever the values of the parameters. We could note a slight effect of J on the RM method: when the number of items increased, the power decreased

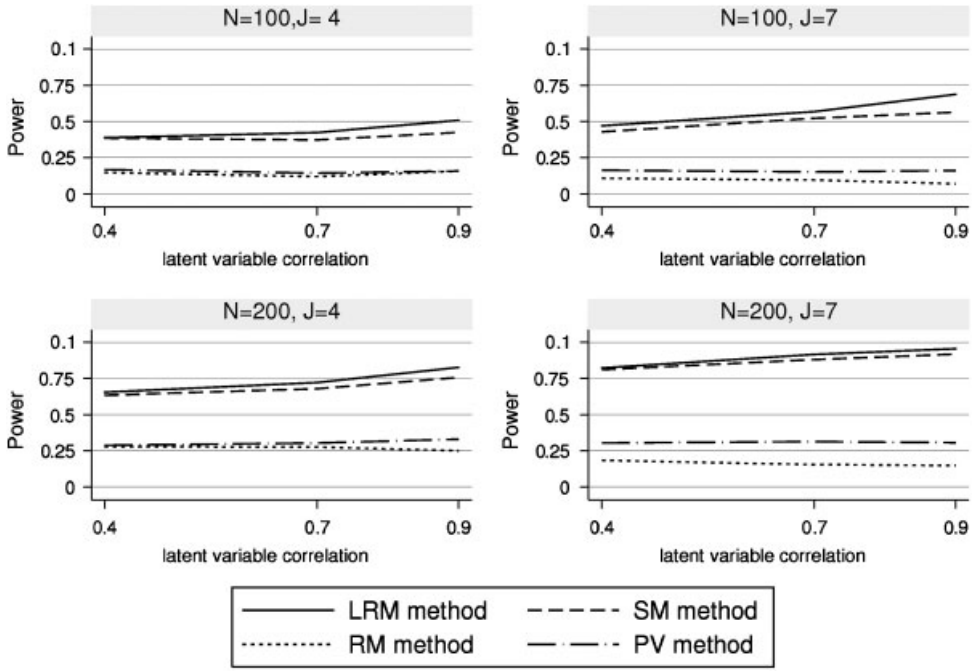

Figure 2. Power of the tests for Score and Mixed models (SM), Rasch and Mixed models (RM), Plausible Values (PV), and Longitudinal Rasch Mixed model (LRM) methods for different values of sample size (N), number of items (J), and latent variable correlation (ρ). Results from analyses with an AR(1) structure for the covariance matrix in RM, PV, and SM methods and an unstructured covariance matrix for LRM method.

by contrast with the LRM and SM methods. A N effect was also shown: the power increased with the sample size.

3.3. Bias on the time effect estimation

Table II shows the time effect estimation between the first and second time of measurement for each method for different values of the parameters: sample size, number of items, and latent variable correlation. For all methods, all cases presented unbiased estimation of time effect between time 1 and time 2 under H_0.

For RM and PV methods, the time effect estimation under H_1 was always biased. For both methods, d_{12} was underestimated in all cases. Table II showed that the estimated time effect was 10 to 20 times less than the true value for RM and 2 to 3 times less than the true value for PV. A J effect could be observed for RM. As the number of items increased, the estimation of time effect decreased and was more biased.

On the contrary, estimations from the LRM method were always unbiased. Remember that the true value of time effect on score was not known. Thus, the bias on time effect estimation could not be assessed for the SM method. The results on time effect estimation between time 2 and time 3 are comparable to the results on time effect estimation between time 1 and time 2 (results not shown).

4. Illustrative example

An analysis was performed on a longitudinal study aimed at evaluating the evolution of nonspecific symptoms and quality of life in primary hyperparathyroidism before and after surgery [20]. The study was multicentric and took place in six academic departments of Endocrine Surgery in France. Patients with primary hyperparathyroidism scheduled for parathyroidectomy were asked to fill out a questionnaire about nonspecific symptoms and a quality of life questionnaire (SF-36). Patients were evaluated during the preoperative period and at 3 and 6 months after surgery.

The SF-36 is a 36-item generic scale made of 8 dimensions: physical functioning, social functioning, bodily pain, general health perceptions, vitality, role limitations due to emotional problems

Table II. Time effect estimation between time 2 and time 1 ($\hat{d}_{12}$) under H_0 and H_1 for Score Mixed model (SM), Rasch Mixed model (RM), Plausible Values (PV), and Longitudinal Rasch Mixed model (LRM) methods for different values of sample size (N), number of items (J), and latent variable correlation (ρ). Results from analyses with an AR(1) structure for the covariance matrix in SM, PV, and RM methods and an unstructured covariance matrix for LRM method. Mean values of $\hat{d}_{12}$ and standard errors (s.e.).

					SM		RM		PV		LRM	
	N	J	ρ	d_{12}	$\hat{d}_{12}$	s.e.	$\hat{d}_{12}$	s.e.	$\hat{d}_{12}$	s.e.	$\hat{d}_{12}$	s.e.
H_0	100	4	0.4	0	0.010	0.007	0.006	0.004	0.008	0.006	0.013	0.009
			0.7	0	−0.003	0.006	−0.002	0.004	0.001	0.006	−0.004	0.008
			0.9	0	−0.007	0.006	−0.005	0.003	−0.003	0.006	−0.010	0.008
		7	0.4	0	0.005	0.009	0.002	0.004	0.000	0.006	0.005	0.007
			0.7	0	0.008	0.009	0.004	0.004	0.008	0.006	0.007	0.007
			0.9	0	0.011	0.007	0.005	0.003	0.005	0.006	0.009	0.006
H_0	200	4	0.4	0	−0.003	0.005	−0.001	0.003	0.001	0.004	−0.003	0.006
			0.7	0	0.007	0.005	0.004	0.003	0.007	0.004	0.009	0.006
			0.9	0	−0.007	0.004	−0.004	0.002	−0.004	0.004	−0.009	0.006
		7	0.4	0	−0.003	0.007	−0.001	0.003	−0.002	0.004	−0.002	0.005
			0.7	0	0.002	0.006	0.001	0.003	0.003	0.004	0.003	0.005
			0.9	0	−0.005	0.005	−0.002	0.002	−0.003	0.004	−0.004	0.004
H_1	100	4	0.4	0.2	0.139	0.007	0.023*	0.002	0.082*	0.006	0.184	0.009
			0.7	0.2	0.138	0.006	0.022*	0.002	0.073*	0.006	0.186	0.008
			0.9	0.2	0.156	0.006	0.022*	0.002	0.084*	0.006	0.211	0.008
		7	0.4	0.2	0.237	0.009	0.009*	0.001	0.080*	0.006	0.189	0.007
			0.7	0.2	0.246	0.008	0.010*	0.001	0.091*	0.006	0.197	0.007
			0.9	0.2	0.250	0.007	0.007*	0.001	0.084*	0.006	0.202	0.006
H_1	200	4	0.4	0.2	0.148	0.005	0.023*	0.001	0.093*	0.004	0.197	0.006
			0.7	0.2	0.155	0.004	0.023*	0.001	0.086*	0.004	0.207	0.005
			0.9	0.2	0.150	0.004	0.022*	0.001	0.087*	0.004	0.202	0.006
		7	0.4	0.2	0.255	0.007	0.011*	0.001	0.089*	0.004	0.203	0.005
			0.7	0.2	0.243	0.006	0.008*	0.001	0.082*	0.004	0.194	0.005
			0.9	0.2	0.254	0.005	0.008*	0.001	0.088*	0.004	0.202	0.004

*The t-test of d_{12} ($H_0: d_{12}=0$ or $d_{12}=0.2$) is significant at 5 per cent.
Under H_1, the time effect bias on the score scale could not be assessed.

(role emotional), role limitations due to physical health problems (role physical), and mental health. The role physical dimension (RP) includes four dichotomous items. The data of the 57 patients from the study were analyzed using the three methods to estimate time effect, variances, and correlations. Item responses were summed to obtain a score on a scale of 0 (lowest symptom level) to 4 (highest symptom level).

The results of the analysis of the RP dimension of the SF-36 with SM, RM, PV, and LRM methods are presented in Table III. SM, RM, and PV methods used a covariance matrix of AR(1) type. LRM method used an unstructured covariance matrix. All the methods have rejected the hypothesis of equality of the means at $\alpha=5$ per cent level. The estimated values of time effect claim for an improvement of quality of life on role physical dimension at 3 months after surgery and a stability between the third and sixth month after surgery. On simulation data, time effect was underestimated for the RM and PV methods under H_1 whereas the time effect was unbiased for the LRM method. Effect sizes estimated on SF-36 data are large (from 0.25 to 1.13), much larger than effects sizes used for simulation data ($d_{t,t+1}=0.2$), especially between the first and second times. The fact that all methods conclude to a time effect, even for the RM and PV methods for which power is low, might be due to the large effect size. A large effect size can be more often detected than a medium one. The correlation coefficients between two measurements were estimated around 0.6 for each method except for the PV method where the correlation coefficient was estimated around 0.3. The case of simulation data where $N=100$, $J=4$, and $\rho=0.7$ is the closest to the SF-36 data. For this simulation case, the power was estimated to 37.2, 12, 14.2, and 42.3 per cent for the SM, RM, PV, and LRM methods, respectively. An effect of the sample size on power was shown. We might argue that power for SF-36 data might be lower than power found in the simulation study that was evaluated for a larger sample size.

Table III. Results of the analysis of the dimension 'Role Physical' of the SF-36 for Score and Mixed models (SM), Rasch and Mixed models (RM), Plausible Values (PV), and Longitudinal Rasch Mixed model (LRM). Estimations of time effect between time t and t' ($\hat{d}_{tt'}$), variance at time t ($\hat{\sigma}_t^2$), and correlation between time t and t' ($\hat{\rho}_{tt'}$) for $t=1,2,3$ and $t\neq t'$. Test statistic and p-value of the test of equality of the means (Fischer test for RM and SM, Wald test for LRM).

	SM	RM	PV	LRM
$\hat{d}_{12}$	1.32	2.23	2.71	3.10
$\hat{d}_{23}$	−0.45	−0.76	−0.78	−0.94
$\hat{d}_{13}$	0.87	1.47	1.93	2.16
$\hat{\sigma}_1^2$	2.41	6.87	9.80	9.46
$\hat{\sigma}_2^2$	2.41	6.87	9.80	5.61
$\hat{\sigma}_3^2$	2.41	6.87	9.80	18.86
$\hat{ES}_{12}$	0.85	0.85	0.86	1.13
$\hat{ES}_{23}$	−0.29	−0.29	−0.25	−0.27
$\hat{ES}_{13}$	0.56	0.56	0.62	0.57
$\hat{\rho_{12}}$	0.56	0.56	0.31	0.54
$\hat{\rho_{23}}$	0.56	0.56	0.31	0.61
$\hat{\rho_{13}}$	0.31	0.31	0.10	0.65
Test statistic	16.53	16.62	11.79	12.4
p-value	<0.0001	<0.0001	<0.0001	0.002

5. Discussion

PRO data are widely used in health sciences, in particular for the evaluation of HRQoL. Such data are often measured several times on the same patients in order to study the evolution of the outcome with time. Four methods to analyze longitudinal PRO data were compared: the SM method based on the CTT approach, the RM, PV, and LRM methods all based on the Rasch model. The four methods have shown comparable results in terms of type I error with type I error rates close to 5 per cent. The LRM and SM methods presented comparable power and unbiased time effect estimations. It has been shown that the rise of the sample size, the questionnaire length or the correlation between measures increased the values of power. This expected rise with these parameters is concordant with the results in Glas *et al.* [15]. The impact of sample size and the number of items on power leads to take with caution the results from studies of longitudinal PRO data with small sample sizes and short questionnaires.

The RM and PV methods presented much lower power as compared with SM and LRM ones. Moreover, it has been shown that RM and PV gave biased time effect estimations under H_1. The large underestimation of time effect under H_1 explains the important loss of power of these methods as compared to the two others that were unbiased. The RM and PV methods seem to be inadequate to analyze longitudinal patient reported outcomes data and should be avoided.

An explanation of the poor performance of both methods might come from the first step of estimation with Bayes *Expected A Posteriori* or plausible values. In the RM method, individual latent traits are estimated based on the EAP estimate. The EAP estimate is a point estimate defined as the mean of the posterior distribution. The mean of the EAP estimates is known to be an unbiased estimate of the population mean. Nevertheless, the variance of the EAPs is an underestimate of the variance population [21]. Other point estimates exist that produce estimates by maximizing the likelihood of observed item responses. For example, the maximum likelihood estimate (MLE) is obtained from joint maximum likelihood estimation. The mean of the MLE estimates is also an unbiased estimate of the population mean and the variance of the MLEs is an overestimate of the population variance. The bias in variance of MLE and EAP is not reduced when the sample size increases but it goes down when the number of items increases [16]. A correction based on the reliability index can be used to eliminate the bias in the EAP case [22]. EAP estimates present the problem of shrinkage toward the mean of prior distribution.

It has been shown that the bias due to shrinkage is minimal with over than 20 items [23]. In this study, the number of items is too small to avoid the bias due to shrinkage. As a consequence of estimating the individual latent trait on the overall sample without time factor, the EAPs are shrunk to zero. This leads to an underestimation of the time effect for the RM method.

In the PV method, plausible values are randomly drawn from the posterior distribution of each individual. In contrast to point estimates, plausible values allow patients with the same total score (and so the same posterior distribution) to have different plausible values. The main difference between the RM method and plausible value imputation is that the latter considers the variability of the estimated value for the latent trait in time effect estimation, whereas the RM method does not take into account the uncertainty related to the latent variable estimation. The underestimation of the time effect is reduced by the use of plausible values instead of EAP estimates but the observed bias is still important and the power is much lower than for the SM and LRM methods. As expected with the PV method, the variance is no longer biased but the correlation coefficient is still underestimated as with the RM method (results not shown). This bias probably affects the estimated covariance matrix of $\hat{\beta}$ in the mixed model step and may explain the poor performance in terms of power because the F-test used for testing time effect and the computation of type I error and power uses this estimated covariance matrix. The poor performance of the 2-step Rasch-based methods (RM and PV) against 1-step Rasch-based method (LRM) pleads for the use of a multivariate form of the Rasch model to account for the particular structure of repeated measures.

The plausible value imputation is widely used in large-scale educational surveys where the number of items and sample size are much larger than in health sciences. In health sciences, Glas *et al.* [15] have shown that, in most of the cases that were studied, a longitudinal IRT model and plausible value imputation methods lead to comparable results in terms of type I error rate and power in the context of two groups with two time points. Nonetheless, the longitudinal IRT model performs better than plausible value imputation method when the number of items is small ($J=5$ or 10) which is the case in our simulation study ($J=4$ or 7). Furthermore, no comparisons were made with a method based on the CTT, which is widely used in practice, and no more than two time points were studied by the authors.

A linear time effect was assumed in this study. This assumption can be inadequate in some cases. For example, in a QOL study where three assessments take place before treatment, during treatment and after treatment, the treatment can have a deleterious effect on quality of life level. We can assume, for instance, that the quality of life decreases between the first and second assessment and increases between the second and third assessment hence leading to a quadratic evolution with time.

Covariance structures such as AR(1) are adequate in a context of equally spaced time, but in practice time of assessments they can be unequally spaced due to the design of the study or problems of recruitment and followup. In the illustrative example, LRM method has shown different variances over time. This indicates that the assumption of constant variances made in this simulation study can be inappropriate for some other studies. Moreover, covariance matrix in the example did not seem to have an AR(1) structure as the correlation between time 1 and time 3 was not close to the square of correlation between consecutive times. Correlation seemed to be constant whatever the time points chosen.

All results from covariance pattern models were presented for an AR(1) structure for the covariance matrix. Regarding the value of AIC for different structures of covariance matrix, the CSH structure most often minimized the AIC of the models for $\rho=0.9$. Investigating the impact on results of misspecification of covariance matrix structure by performing analyses with a CSH structure led to comparable results than the analyses using AR(1) structure in terms of type I error, power and time effect bias (results not shown).

The repeated use of a questionnaire can cause a problem in terms of response-shift. For instance, the assessment of quality of life over time is based on the assumption that the perception that patients have of their own quality of life will not change over time. But patients are faced with a disease and its treatment that may change their perception leading to the phenomenon called response-shift. As patients are adapting to the adverse effects of the disease and its treatments, the repeated measures of quality of life become difficult to compare due to the response-shifts. An observed evolution of a patient's quality of life may confound a true change in the quality of life and the change of patient's perception. As defined by Barclay *et al.* in a recent review of the subject [24], response-shift involves a change in the meaning of an individual's self evaluation of HRQoL as a result of a change in their internal standards, values

and/or concepts of HRQoL. Three components of response-shift are identified: recalibration (a change in the respondent's internal standard of measurement), reprioritization (a change in the importance of component domains constituting the target construct), and reconceptualization (a redefinition of the target construct) [25]. Specific designs like the then-test have been developed to detect response-shift. They have been first used in the area of educational training interventions and then in the area of quality of life, in particular for cancer patients. Treatments of cancer patients can be harmful for quality of life and it has been shown that these patients succeed in adapting to the adverse effects of the disease and its treatments [26]. Since then, the impact of health state changes on an individual's quality of life has gained increased attention in social and medical clinical research. The response-shift that may occur is now considered in studies on evolution of quality of life but the debate on which method to use to detect the response-shift still continues. Some methods are addressing the problem of the response-shift at the design stage of the study, such as the then-test and the individualized methods. Other methods are statistical methods to address response shift, such as factor analysis, growth curve analysis, and Rasch analysis. Among all these possibilities, the then-test is the most commonly used method to measure response-shift. Different components of the response-shift are detected from a method to another. The major point of development remains the quantification of the response-shift. Whereas each method allows to detect it, only the then-test and the factor analysis give a value of change of quality of life adjusted for response-shift effect. Although this simulation study assumed no response-shift, this subject has gained major concern in longitudinal PRO studies and it will be of interest to study the behavior of CTT and Rasch-based methods when response-shift is present.

Finally, many longitudinal studies are faced with the problem of missing observations. In this case, different approaches are often adopted: complete-case analysis, available-data analysis, and imputation. Because the SM method is based on the score computed by summing item responses, in the presence of missing data, the analysis can only be performed through complete-case analysis or imputation approach. In the LRM method, based on item responses, the analysis can also be performed with available-data approach.

Little and Rubin [27] made distinctions between missing value processes. A missing data process is said to be missing completely at random (MCAR) if the missingness is independent of both unobserved and observed data. Data are missing at random (MAR) if the missingness is independent of the unobserved measurements, conditional on the observed data. Otherwise, the missing data process is missing not at random (MNAR). Likelihood-based analyses that ignore the missing data mechanism lead to valid analyses when the missingness is ignorable (MCAR or MAR) [28]. Selection models and pattern-mixture models [29] were proposed to model nonignorable nonresponse. They are an interesting way of dealing with MNAR missing data process by modeling explicitly the missing data mechanism. These models have to be used with caution because untestable assumptions have to be made on the missing data process for selection models and untestable identifying restrictions are used in pattern-mixture models.

Each approach for handling missing data leads to different results and possible bias. It seems important to study the impact of missing data on the performances of methods to analyze longitudinal latent variables. We suspect that there will be a more important loss of information using a CTT-based method than a Rasch-based method because of the necessity to impute for missing data or to use only complete cases in SM method. In the presence of missing data, we expect that the LRM method will present better results than the SM method as it has been shown in the context of sequential analysis of latent variables [30].

This simulation study is based on the assumption that the data follow a Rasch model. The different results on the performance of the methods will probably be affected to different extents if the Rasch model does not correctly fit the data. In this case, we can expect that the CTT approach will perform better than methods based on the Rasch model.

In conclusion, it has been shown that using either the SM or LRM method give comparable and satisfying results. These two methods are adequate for the analysis of longitudinal PRO data following a Rasch model without missing data.

Acknowledgements

This work was supported by the Ligue Nationale Contre le Cancer.

References

1. Gotay CC, Kawamoto CT, Bottomley A, Efficace F. The prognostic significance of patient-reported outcomes in cancer clinical trials. *Journal of Clinical Oncology* 2008; **26**(8):1355–1363.
2. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests* (expanded edn). University of Chicago Press: Chicago, 1980.
3. Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data*. Springer: New York, 2001.
4. Fischer GH, Molenaar IW. *Rasch Models*, *Foundations*, *Recent Developments*, *and Applications*. Springer: New York, 1997.
5. Bjorner JB, Petersen MA, Groenvold M, Aaronson N, Ahlner-Elmqvist M, Arraras JI, Brédart A, Fayers P, Jordhoy M, Sprangers M, Watson M, Young T. Use of item response theory to develop a shortened version of the EORTC QLQ-C30 emotional functioning scale. *Quality of Life Research* 2004; **13**(10):1683–1697.
6. Garcia SF, Cella D, Clauser SB, Flynn KE, Lad T, Lai JS, Reeve BB, Smith AS, Stone AA, Weinfurt K. Standardizing patient-reported outcomes assessment in cancer clinical trials: a patient-reported outcomes measurement information system initiative. *Journal of Clinical Oncology* 2007; **25**(32):5106–5112.
7. Rijmen F, Tuerlinckx F, De Boeck P, Kuppens P. A nonlinear mixed model framework for Item Response Theory. *Psychological Methods* 2003; **8**(2):185–205.
8. Embretson S. A multidimensional latent trait model for measuring learning and change. *Psychometrika* 1991; **56**(3):495–515.
9. Fitzmaurice G, Laird N, Ware SJ. *Applied Longitudinal Analysis*. Wiley: Hoboken, 2004.
10. Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G. *Longitudinal Data Analysis*: *A Handbook of Modern Statistical Methods*. Chapman & Hall/CRC: London, 2008.
11. Hoijtink H, Boomsma A. On person parameter estimation in the dichotomous Rasch model. In *Rasch Models*, Fischer GH, Molenaar IW (eds). Springer: New York, 1997; 53–68.
12. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley-IEEE: New York, 2004.
13. Wu M, Adams RJ. *PISA 2000 Technical Report*. OECD Publications: Paris, 2002.
14. Thomas N. Assessing model sensitivity of the imputation methods used in the national assessment of educational progress. *Journal of Educational and Behavioral Statistics* 2000; **25**(2):351–371.
15. Glas CAW, Geerlings H, van de Laar MAFJ, Taal E. Analysis of longitudinal randomized clinical trials using item response models. *Contemporary Clinical Trials* 2009; **30**(2):158–170.
16. Wu M. The role of plausible values in large-scale surveys. *Studies in Educational Evaluation* 2005; **31**(2–3):114–128.
17. Littell RC, Milliken GA, Stroup WW, Wolfinger R. *SAS System for Mixed Models*. SAS Insitute Inc: Cary, NC, 1996.
18. Hardouin J. Rasch analysis: estimation and tests with Rasch test. *Stata Journal* 2007; **7**(1):22–44.
19. Zheng X, Rabe-Hesketh S. Estimating parameters of dichotomous and ordinal item response models with gllamm. *Stata Journal* 2007; **7**(3):313–333.
20. Caillard C, Sebag F, Mathonnet M, Gibelin H, Brunaud L, Loudot C, Kraimps JL, Hamy A, Bresler L, Charbonel B, Leborgne J, Henry JF, Nguyen JM, Mirallié E. Prospective evaluation of quality of life (SF-36v2) and nonspecific symptoms before and after cure of primary hyperparathyroidism (1-year follow-up). *Surgery* 2007; **141**(2):153–160.
21. Mislevy RJ, Beaton AE, Kaplan B, Sheehan KM. Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement* 1992; **29**:133–161.
22. Adams RJ. Reliability as a measurement design effect. *Studies in Educational Evaluation* 2005; **31**(2–3):162–172.
23. Wainer H, Thissen D. Estimating ability with the wrong model. *Journal of Educational and Behavioral Statistics* 1987; **12**(4):339–368.
24. Barclay-Goddard R, Epstein JD, Mayo NE. Response shift: a brief overview and proposed research priorities. *Quality of Life Research* 2009; **18**(3):335–346.
25. Schwartz CE, Sprangers MAG. Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Social Science and Medicine* 1999; **48**(11):1531–1548.
26. Sprangers MAG. Response-shift bias: a challenge to the assessment of patients' quality of life in cancer clinical trials. *Cancer Treatment Reviews* 1996; **22**(Supplement 1):55–62.
27. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. Wiley: New York, 1987.
28. Troxel AB, Fairclough DL, Curran D, Hahn EA. Statistical analysis of quality of life with missing data in cancer clinical trials. *Statistics in Medicine* 1998; **17**:653–666.
29. Little RJA. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* 1995; **90**(431):1112–1121.
30. Sébille V, Hardouin J, Mesbah M. Sequential analysis of latent variables using mixed-effect latent variable models: impact of non-informative and informative missing data. *Statistics in Medicine* 2007; **26**(27):4889–4904.

Annexe B

Article : Analysis of longitudinal Patient-Reported Outcomes with informative and non-informative dropout : Comparison of CTT and Rasch-based methods.

International Journal of Applied Mathematics & Statistics,
Int. J. Appl. Math. Stat.; Vol. 24; Issue No. SI-11A; Year 2011,
ISSN 0973-1377 (Print), ISSN 0973-7545 (Online)

Analysis of longitudinal Patient-Reported Outcomes with informative and non-informative dropout: Comparison of CTT and Rasch-based methods.

Myriam Blanchin[1], Jean-Benoit Hardouin[1], Tanguy Le Neel[1], Gildas Kubis[1] and Véronique Sébille[1]

[1]EA 4275 "Biostatistics, Clinical Research
and Subjective Measures in Health Sciences"
Faculty of Pharmaceutical Sciences, University of Nantes, France
myriam.blanchin@etu.univ-nantes.fr

ABSTRACT

Patient-reported outcomes (PRO) are more and more used in health sciences to evaluate concepts such as health-related quality of life. These outcomes cannot be directly observed and are often referred to a latent variable. Two psychometric theories exist for the analysis of PRO: the classical test theory, the most common used in practice and the item response theory with its most used model, the Rasch model. In many studies, PRO are collected longitudinally in order to study the evolution of the outcome through time. Missing data are frequently encountered in longitudinal studies and can be potentially informative. This study aimed at comparing Classical Test Theory (CTT) and Rasch-based approaches to analyze longitudinal PRO collected from a scale validated with a Rasch model and studying the impact of dropout, informative or not, on both approaches. Data with informative dropout have shown estimation bias and have to be analyzed with more appropriate methods. For complete data and data with non-informative dropout, a method of analysis based on the Rasch model may be preferred for the analysis of longitudinal PRO collected from a scale validated with a Rasch model due to the generally observed slight gain of power and the psychometric properties of the model.

Keywords: Classical Test Theory, Rasch model, longitudinal data, dropout, informative missing data.

2010 Mathematics Subject Classification: 92B15, 62P10 .

1 Introduction

Patient-reported Outcomes, PRO, have gained major concern in the past years. This type of measures, reported by patients based on their perceptions, includes health-related quality of life (HRQoL), functional well-being, patient satisfaction with treatment,... PRO are broadly used in clinical trials as secondary outcome, especially in chronic diseases like cancer and are sometimes used as primary outcome in some contexts. The evaluated outcome (e.g. HRQoL) is often referred to a latent variable and is measured through the responses of patients to items. The special nature of PRO, which are not directly observable, beg the question of the

analysis of the data. Two psychometric theories exist for the analysis of PRO. The most common used approach is the Classical Test Theory (CTT). In this theory, the observed score, usually computed by summing item responses, is used to estimate the 'true' value of the evaluated outcome. In the second theory, Item Response Theory (IRT), the probability to answer to an item is a function of a latent variable, which represents the evaluated outcome, and item parameters. Among the wide family of IRT models, the Rasch model (Rasch, 1980; Fischer and Molenaar, 1995) is the most commonly used for dichotomous items due to its properties: parameters invariance, specific objectivity and exhaustivity of the score on the latent trait. The Rasch model is now widely used in development and validation of scales in health sciences (Lai et al., 2007; Cella et al., 1996).

Many studies including PRO are studies on chronic illness or follow-up studies after treatment or surgery. Thus, patients are evaluated at different time points to allow the analysis of the evolution of the evaluated PRO. In these longitudinal studies, the measures of each patient are therefore unlikely to be independent such as in cross-sectional studies and the correlation between measurements has to be taken into account in the analysis. One way to deal with correlated data is the use of linear mixed models (Verbeke and Molenberghs, 2000b; Fitzmaurice et al., 2009).

Missing data are frequently encountered in longitudinal studies. For instance, a patient can drop out from the study at a certain time point and so answers to questionnaire are missing for this patient after this time. Intermittent missing data can also occur when some items are not answered in a questionnaire. When the reason for missingness may be related to the evaluated outcome level of the patient, the missing data are said to be informative. Otherwise, they are called ignorable. Missing data, depending on their amount and informativity, may have an impact on the analysis and interpretation of the data. There may be a reduction of the statistical power of the analysis and a bias may be introduced leading to incorrect conclusions.

In practice, when longitudinal data coming from a scale validated with a Rasch model have to be analyzed, many methods can be considered. The researchers tend to use more often the CTT approach, probably more from habit than evidence of suitability. The purpose of this paper is to compare methods either based on CTT or Rasch model to analyze longitudinal latent variables through a simulation study. The impact of missing data in this context has also been studied.

2 Methods

2.1 Longitudinal data analysis

When measures are repeated on the same patients through time, linear mixed models are widely used for the analysis of the data. These models allow to deal with the correlation between measures of longitudinal data by specifying fixed effects (population characteristics), random effects (subject-specific effects) and structure of the variance-covariance matrix

(Verbeke and Molenberghs, 2000b). A general linear mixed model can be written as follows:

$$\begin{aligned}
&\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i,\\
&\boldsymbol{b}_i \sim N_q(0, \boldsymbol{D}),\\
&\boldsymbol{e}_i \sim N_{n_i}(0, \boldsymbol{\Sigma}_i),\\
&\boldsymbol{b}_1, ..., \boldsymbol{b}_q, \boldsymbol{e}_1, ..., \boldsymbol{e}_N \; independent\;,\\
&\boldsymbol{Y}_i \sim N_{n_i}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \boldsymbol{\Sigma}_i),
\end{aligned} \tag{2.1}$$

where $\boldsymbol{Y}_i$ is the response vector for patient i, $i = 1, ..., N$, N is the number of patients, n_i is the number of observations on patient i, p the number of fixed parameters, q the number of random parameters, $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of fixed effects parameters, $\boldsymbol{X}_i$ is the $(n_i \times p)$ design matrix for fixed effects, $\boldsymbol{b}_i$ is a $(q \times 1)$ vector of random effects parameters, $\boldsymbol{Z}_i$ is the $(n_i \times q)$ design matrix for random effects, $\boldsymbol{e}_i$ is a $(n_i \times 1)$ vector of residual components, $\boldsymbol{D}$ is the $(q \times q)$ among-unit covariance matrix and $\boldsymbol{\Sigma}_i$ is the $(n_i \times n_i)$ within-unit covariance matrix.
Two methods of estimation based on the likelihood can be used to estimate the mean parameters $\boldsymbol{\beta}$ and variance components $\boldsymbol{\omega}$ (that contains all variances and covariance parameters found in $V_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \boldsymbol{\Sigma}_i$) of the model: Maximum Likelihood estimation (ML) and Restricted maximum Likelihood estimation (REML). The estimations of variance components obtained with ML are known to be biased for finite samples. The REML estimation is used to correct the bias on ML estimates of variance components (Laird and Ware, 1982). The REML estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ are found by maximizing the so-called REML likelihood function.

$$\begin{aligned}
L_{REML}(\boldsymbol{\beta}, \boldsymbol{\omega}) = &\left|\sum_{i=1}^{N} \boldsymbol{X}_i' V_i^{-1}(\boldsymbol{\omega}) \boldsymbol{X}_i\right|^{-1/2}\\
&\times \prod_{i=1}^{N} (2\pi)^{-n_i/2} |V_i(\boldsymbol{\omega})|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})' V_i^{-1}(\boldsymbol{\omega})(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right\}
\end{aligned} \tag{2.2}$$

2.2 Patient Reported Outcomes analysis

Assume that a questionnaire has J dichotomous items and that measures are repeated T times on the N patients of the study. The response of patient i ($i = 1, ..., N$) to an item j ($j = 1, ..., J$) at time t ($t = 1, ..., T$) is denoted by $Y_{ij}^{(t)}$.

2.2.1 Classical Test Theory approach

The Classical Test Theory (CTT) approach is based on a score usually computed by summing item responses. This approach was the first one developed in psychometrics and has widely spread in PRO analysis because of its simplicity to use and to interpret. The basic model assumes that the observed score is a linear function of the true score and an error term.
A method called **Score and Mixed Models (SM)** and based on the CTT approach was used for the analysis. In the first step, the score of each patient at each time was computed by summing the J item responses of the patient i at time t. A simple linear mixed model was then applied on the scores to investigate whether a time effect was plausible. The SM method was

carried out as follows:

$$\begin{aligned} \boldsymbol{S}_i^{(t)} &= \sum_{j=1}^{J} Y_{ij}^{(t)}, \\ \boldsymbol{S}_i &= (S_i^{(1)}, ..., S_i^{(T)})' = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{e}_i, \\ \boldsymbol{e}_i &\sim N(0, \boldsymbol{\Sigma}_{S,i}), \\ \boldsymbol{S}_i &\sim N_T(\boldsymbol{\mu}_S, \boldsymbol{\Sigma}_{S,i}), \end{aligned} \tag{2.3}$$

with $\boldsymbol{\mu}_S = (\mu_S^{(1)} \cdots \mu_S^{(T)})' = \boldsymbol{X}_i\boldsymbol{\beta}$. The mean parameters β and variance components $\boldsymbol{\omega}$ of the model were estimated using REML estimation in SAS Proc MIXED (Littell et al., 1996).

Three covariance structures $\boldsymbol{\Sigma}_{S,i}$ are often used with longitudinal data : unstructured, first-order autoregressive and heterogeneous compound symmetry denoted by UN, AR(1) and CSH respectively. The unstructured matrix is the most general possible structure. It is used when no hypothesis can be made on the structure of the covariance matrix but leads to estimate an important number of parameters. The AR(1) structure assumed that variances are constant over time and that correlation decreases when measures get further apart from each other in time. On the contrary, the choice of a CSH structure assumes that the variances are not equal but that the correlation is constant over time. For $T = 3$, the three structures of covariance can be written as follows with ρ denoting the correlation coefficient, σ_{ij} the covariance between latent variables at time i and time j ($i \neq j$) and σ_i^2 the variance of the latent variable at time i.

$$\boldsymbol{\Sigma}_{S,i} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{pmatrix} \text{ for UN} \qquad \boldsymbol{\Sigma}_{S,i} = \begin{pmatrix} \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho \\ \sigma^2\rho^2 & \sigma^2\rho & \sigma^2 \end{pmatrix} \text{ for AR(1)}$$

$$\boldsymbol{\Sigma}_{S,i} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 & \sigma_2\sigma_3\rho \\ \sigma_1\sigma_3\rho & \sigma_2\sigma_3\rho & \sigma_3^2 \end{pmatrix} \text{ for CSH}$$

To compare non-nested models for the covariance, one of the information criteria that have been proposed is the Akaike Information Criteria also noted AIC (Akaike, 1974). This tool for model selection aims at comparing models based on their maximized log-likelihood value, ensuring that the retained models show a good fit of data. In order to select the most parsimonious model, the AIC penalizes models for the use of too many parameters. When the likelihood is estimated using REML, the AIC can be expressed as:

$$AIC = -2\hat{l} + 2c \tag{2.4}$$

where $\hat{l}$ is REML maximum log-likelihood and c is the number of covariance parameters. The most parsimonious correct model will be the model with the smallest AIC amongst the models with the same mean structure.

2.2.2 Item Response Theory and the Rasch model

Item Response Theory (IRT) emerged recently in instrument development and data analysis in health outcomes measurements due to its potential advantages over CTT such as parameters invariance (Hambleton, 2000). IRT is a family of models that express the probability of a patient's particular response to an item as a function of characteristics of the patient (latent variable θ) and characteristics of the item. The latent variable is the evaluated outcome and is considered as latent because it is not observable and must be inferred from item responses. Most of IRT models assume the unidimensionality of the construct, that is a unique latent variable explains the item responses. Among the unidimensional IRT model, the most commonly used model is the Rasch model (Rasch, 1980) due to its properties: the exhaustivity of the score on the latent trait and the specific objectivity. The exhaustivity refers to the property that the total score of a person is a sufficient statistic for the unknown latent trait. This means that no additional information is needed to estimate the person parameter θ. Each total score is associated with only one trait level in the Rasch model whatever the pattern of responses. The property of specific objectivity ensures that the difference between two latent traits does not depend on the set of items used to evaluate these traits. It allows to construct shorter versions of questionnaires or several versions of a same questionnaire to adapt the version to the patient's latest variable level.

The Rasch model expresses the probability of a response y ($y = 0$ for a negative response (the most pejorative response) and $y = 1$ for a positive response) of an individual i ($i = 1, ..., N$) to a dichotomous item j ($j = 1, ..., J$) as a logistic function of the individual value of the latent trait θ_i and one item parameter, its difficulty δ_j.

$$P(Y_{ij} = y|\theta_i, \delta_j) = \frac{exp(y(\theta_i - \delta_j))}{1 + exp(\theta_i - \delta_j)} \tag{2.5}$$

The Rasch model has been extended to situations where responses of individuals to items are observed at several points in time (Meiser, 2007). A longitudinal form of the Rasch model was used for the analysis in the method called **Longitudinal Rasch Model (LRM)**.

$$\begin{aligned} &P(Y_{ij}^{(t)} = y^{(t)}|\theta_i^{(t)}, \delta_j) = \frac{exp(y^{(t)}(\theta_i^{(t)} - \delta_j))}{1 + exp(\theta_i^{(t)} - \delta_j)} \\ &\boldsymbol{\theta_i} = (\theta_i^{(1)}, ..., \theta_i^{(T)})' \; iid \; N_T\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}_i\right) \end{aligned} \tag{2.6}$$

This model assumes that the item parameters δ_j remain constant over time. The change in the latent ability $\boldsymbol{\theta}$ may be person-specific, that is the speed or direction of the evolution may be different from a person to another. As $\boldsymbol{\theta}$ is assumed to have a multinormal distribution, this model is of the family of the mixed-effects logistic models. The mean parameters $\boldsymbol{\mu}$ and covariance parameters $\boldsymbol{\Sigma}_i$ of the model are estimated using Marginal Maximum Likelihood (MML) estimation method. The marginal likelihood is expressed as

$$L(\delta_1, .., \delta_J, \boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{y}) = \prod_{i=1}^{N} \int_{\mathbb{R}^T} \prod_{t=1}^{T} \prod_{j=1}^{J} \frac{\exp\left(y_{ij}^{(t)}(\theta_i^{(t)} - \delta_j)\right)}{1 + \exp(\theta_i^{(t)} - \delta_j)} G(\boldsymbol{\theta}_i/\boldsymbol{\mu}, \boldsymbol{\Sigma}_i) d\boldsymbol{\theta}_i \tag{2.7}$$

with $G(./\boldsymbol{\mu}, \boldsymbol{\Sigma}_i)$ the multivariate normal distribution function with mean vector $\boldsymbol{\mu} = (\mu^{(1)} \cdots \mu^{(T)})'$ and an unstructured covariance matrix $\boldsymbol{\Sigma}_i$. $\mu^{(1)}$ is constrained to 0 in order to ensure the identifiability of the model. The parameters of the model were estimated using gllamm in Stata (Zheng and Rabe-Hesketh, 2007).

2.3 Missing data

Missing assessments of PRO are frequently encountered in longitudinal studies. Patients face a disease and/or treatment that have an impact on the evaluated outcome. The reasons for missingness can be totally unrelated to the subject's level of the evaluated outcome (a missed appointment, a move in another town) or may be intimely related to patient's level of evaluated PRO (side effects such as nausea or vomiting).

Missing data are often described as either 'dropout' or 'intermittent'. Dropout occurs when all observations on a subject are obtained until a certain point in time after which all measurements are missing (Diggle and Kenward, 1994). Intermittent missingness occurs when a subject misses an assessment but is later observed.

Little and Rubin (Little and Rubin, 2002) have defined three types of missing data (completely random, random, or not at random) depending on the mechanism that lead to missing data. An observation is said to be missing completely at random (MCAR) if the missingness probability is independent of all previous, current and future assessments. The missing process therefore does not depend on the values of the data, missing or observed. Data are missing at random (MAR) if the missingness probability does not depend on the missing values but only on the observed values. When data are missing not at random (MNAR) the missing data mechanism may depend on the unobserved values. Within the framework of maximum likelihood or Bayesian inference, this mechanism is often termed as 'non-ignorable' or 'informative'. When the data are 'ignorable' (MCAR or MAR), a valid analysis can be obtained through a likelihood-based analysis that ignores the dropout mechanism, provided the parameters describing the measurement process are functionally independent of the parameters describing the dropout process (Verbeke and Molenberghs, 2000b).

2.4 Simulation

The interest of longitudinal studies is in evaluating the evolution of a criteria through time. We define the time effect between time t and time $t+1$ as $d_{t,t+1} = \mu^{(t+1)} - \mu^{(t)}$. The data were simulated with a longitudinal Rasch mixed model assuming that patients were evaluated at three different times ($t = 1, 2, 3$).

$$P(Y_{ij}^{(t)} = y^{(t)} | \theta_i^{(t)}, \delta_j) = \frac{exp(y^{(t)}(\theta_i^{(t)} - \delta_j))}{1 + exp(\theta_i^{(t)} - \delta_j)} \tag{2.8}$$

where the latent trait vector $(\boldsymbol{\theta}_i^{(1)}, \boldsymbol{\theta}_i^{(2)}, \boldsymbol{\theta}_i^{(3)})$', $(i = 1, ..., N)$ had a multivariate normal distribution $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = (-d_\theta, 0, d_\theta)'$ and $\boldsymbol{\Sigma} = \sigma^2 \begin{pmatrix} 1 & \rho_\theta & \rho_\theta^2 \\ \rho_\theta & 1 & \rho_\theta \\ \rho_\theta^2 & \rho_\theta & 1 \end{pmatrix}$.

d_θ is the value of the time effect between two consecutive times, $d_{1,2} = d_{2,3} = d_\theta$. For data simulated without time effect, $d_\theta = 0$. For data simulated with a time effect, $d_\theta = 0.2$. The first-order autoregressive structure adopted for the covariance matrix Σ means that variances are constant with time and that correlation between measures of a same patient decreases with time.

We can expect that some parameters have an impact on the performance of the two methods: datasets with different values for the sample size (N), number of items (J) and correlation of the latent variable (ρ_θ) were simulated. The data were assumed to come from a 4-item scale or a 7-item scale with dichotomous items. The values of difficulty parameters were $\delta_1 = -1$, $\delta_2 = -0.5$, $\delta_3 = 0.5$, $\delta_4 = 1$ for a 4-item scale and $\delta_1 = -1.5$, $\delta_2 = -1$, $\delta_3 = -0.5$, $\delta_4 = 0$, $\delta_5 = 0.5$, $\delta_6 = 1$, $\delta_7 = 1.5$ for a 7-item scale. The sample size could be of 100 or 200 individuals. Three different values for the correlation coefficient of the latent trait between two consecutive times ρ_θ were used: $\rho_\theta = 0.4$ (small correlation), $\rho_\theta = 0.7$, and $\rho_\theta = 0.9$ (high correlation).

To simulate the dropout of patients from the study, a latent variable denoted χ was defined as the dropout propensity. The probability that a patient drops out from the study at time t depends on its dropout propensity. The dropout process was simulated using the following model derived from a 4-parameter logistic model (Sijtsma and Hemker, 2000):

$$P(DO_i^{(t)} = 1|\chi_i^{(t)}, \pi_{min}^{(t)}, \pi_{max}^{(t)}) = \pi_{min}^{(t)} + (\pi_{max}^{(t)} - \pi_{min}^{(t)}) \frac{exp(\chi_i^{(t)})}{1 + exp(\chi_i^{(t)})} \tag{2.9}$$

with $DO_i^{(t)} = 1$ represents the situation where a patient i drops out from the study at time t, $\pi_{min}^{(t)}$ the minimum individual probability of dropout at time t and $\pi_{max}^{(t)}$ the maximum individual probability of dropout at time t. $\pi_{min}^{(t)}$ and $\pi_{max}^{(t)}$ were defined such as the expected proportion of dropout at time t was $\pi^{(t)} = \frac{\pi_{min}^{(t)} + \pi_{max}^{(t)}}{2}$. We assume that data are complete at the first time of evaluation ($\pi^{(1)} = 0$). The dropout of the patients is then linear and π of the remaining patients drop out from the study at each time ($t = 2, 3$).
The dropout propensity χ_i has a multinormal distribution with mean vector (0 0 0)' and a variance covariance matrix equals to $\begin{pmatrix} 1 & \rho_{\theta\chi}^2\rho_\theta & \rho_{\theta\chi}^2\rho_\theta^2 \\ \rho_{\theta\chi}^2\rho_\theta & 1 & \rho_{\theta\chi}^2\rho_\theta \\ \rho_{\theta\chi}^2\rho_\theta^2 & \rho_{\theta\chi}^2\rho_\theta & 1 \end{pmatrix}$.
The correlation between the value of the latent variable at time t, $\theta^{(t)}$, and the dropout propensity at time t, $\chi^{(t)}$, denoted $corr(\theta^{(t)}, \chi^{(t)}) = \rho_{\theta\chi}$ and assumed constant with time, was used to determine the type of missingness of the dropout process. When the value of the latent variable for the outcome does not depend on the dropout propensity ($\rho_{\theta\chi} = 0$), the simulated dropout is of MCAR type following the definition of Little and Rubin. As the two methods are based on likelihood and ignores the dropout mechanism, we can expect that analyses on data with MCAR dropout will be valid. However, it is reasonable to assume that missing data mechanism may often be MNAR in studies including PRO, HRQoL for instance. The simulated dropout is MNAR when $\rho_{\theta\chi} \neq 0$. Furthermore, the patients with worse HRQoL, due to disease progression or increase of side effects, may be most likely to dropout from the study than other patients (Troxel, Fairclough, Curran and Hahn, 1998). So, we assume that $\rho_{\theta\chi} < 0$ to simulate

MNAR dropout.

To study the behaviour of SM and LRM in case of missing data, the proportion of dropout in the simulated datasets could be $\pi^{(t)} = 0\%$ (complete data), 5%, 10% or 20% ($t = 2, 3$). The correlation between the value of the latent variable θ and the dropout propensity were $\rho_{\theta\chi} = 0$ (MCAR dropout), $\rho_{\theta\chi} = -0.4; -0.7; -0.9$ (MNAR dropout with increasing informativity). The different values of the parameters led to consider 312 different cases. Five hundred simulated datasets were generated and analyzed for each case.

To compare the two methods, a test for time effect was defined using an approximate Wald test:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu \Leftrightarrow L\boldsymbol{\mu} = 0$

$H_1 : \exists i | \mu_i \neq \mu \Leftrightarrow L\boldsymbol{\mu} \neq 0$

$$\boldsymbol{L} = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}$$

Under H_0, $T_L = (\boldsymbol{L}\hat{\boldsymbol{\mu}})'(\boldsymbol{L}\hat{\boldsymbol{V}}\boldsymbol{L}')^{-1}\boldsymbol{L}\hat{\boldsymbol{\mu}}$ has approximately a χ^2_r distribution (Verbeke and Molenberghs, 2000a) where $\hat{\boldsymbol{\mu}}$ is the estimate of $\boldsymbol{\mu}$, r is the rank of $\boldsymbol{L}$ and $\hat{\boldsymbol{V}}$ is the estimated covariance matrix.

In order to compare the methods to analyze longitudinal PRO data, three criteria were studied: the type I error, the power and the bias of the time effect estimation. The type I error of the tests were classically computed as the proportion of rejection of H_0 under the null hypothesis. Rejection of H_0 was based on a test of simultaneous equality of mean estimations, i.e. the absence of time effect. This criteria allowed the study of the aptitude of the method to avoid falsely detecting a time effect. The power calculation used the same tests but calculated the proportion of rejection of H_0 under the alternative hypothesis. On the opposite, the power allowed to study the aptitude of the method to correctly detect the presence of time effect. A one-sided McNemar's test for paired data (McNemar, 1947) was used to compare the power observed for each method.

$H_0 : power_{LRM} = power_{SM}$

$H_1 : power_{LRM} > power_{SM}$

The comparison of the estimated time effect to the simulated 'true' time effect gave the bias of time effect and informed about the quality of the parameters estimation. The comparison held for the LRM method as the known true value of time effect had been fixed for the latent variable. For SM method, based on score, the 'true' time effect was not known and was estimated by d_S, the difference of the computed expected score at each time. For $t = 2, 3$,

$$d_S = E\left(S_i^{(t)}\right) - E\left(S_i^{(t-1)}\right) \tag{2.10}$$

with

$$
\begin{aligned}
E\left(S_i^{(t)}\right) &= E\left(\sum_j Y_{ij}^{(t)}\right) \\
&= \sum_j E\left(Y_{ij}^{(t)}\right) \\
&= \sum_j P\left(Y_{ij}^{(t)} = 1\right) \\
&= \sum_j \int_{\mathbb{R}} P\left(Y_{ij}^{(t)} = 1\right) G(\theta_i^{(t)}/\mu^{(t)}, \sigma^2) d\theta_i \\
&= \sum_j \int_{\mathbb{R}} \frac{\exp\left(\theta_i^{(t)} - \delta_j\right)}{1 + \exp(\theta_i^{(t)} - \delta_j)} G(\theta_i^{(t)}/\mu^{(t)}, \sigma^2) d\theta_i
\end{aligned}
\tag{2.11}
$$

where $G(\theta_i^{(t)}/\mu^{(t)}, \sigma^2)$ the normal distribution with mean $\mu^{(t)}$ and variance σ^2. These integrals can be estimated using Gauss-Hermite quadratures. Considering the simulated item parameters, we obtained the following estimations. For a simulated time effect between two consecutive times on the latent variable $d_\theta = 0$, the time effect on the score d_S is also 0. For a simulated time effect on the latent variable $d_\theta = 0.2$, the time effect on the score d_S is 0.15 when $J = 4$ and 0.25 when $J = 7$.

3 Results

To determine which structure of covariance is the most adequate to use for the analysis of data with SM method, we compared the Akaike Information Criteria (AIC) of the three structures of covariance matrix, UN, AR(1) and CSH for each of the 312 cases. When $\rho_\theta = 0.4$ or $\rho_\theta = 0.7$, the AIC was more often minimized by the choice of an AR(1) structure for the covariance matrix. When $\rho_\theta = 0.9$, the CSH structure more often minimized the AIC of the models than the other structures. Results presented further for SM method come from analyses with an AR(1) structure. LRM method used an unstructured covariance matrix.

3.1 Type I error rate and power

Table 1 shows the type I error of the test for time effect for each of the methods depending on the value of all simulation parameters: sample size, number of items, latent variable correlation, proportion of dropout and type of dropout. Both methods, LRM and SM, give comparable values of type I error whatever the value of the simulation parameters. The value of the sample size, the number of items and the latent variable correlation don't seem to have an impact on the values of the type I error since the type I errors only show small variations as the values of these three parameters change. When the data are complete ($\pi = 0$), all type I error are close to the expected value of 5%. In this case, the values range from 4.4% and 3.6% for LRM and SM respectively when N=100, J=7 and $\rho_\theta = 0.4$ to 6.6% and 6.0% for LRM and SM respectively when N=200, J=4 and $\rho_\theta = 0.9$. Whatever the proportion of dropout, data subject to MCAR dropout ($\rho_{\theta\chi} = 0$) show type I errors close to complete data ones. When dropout of

Table 1: Type I error of the tests of time effect for Score Mixed model (SM) and Longitudinal Rasch Mixed model (LRM) methods for different values of sample size (N), number of items (J), latent variable correlation (ρ_θ), proportion of dropout (π) and type of dropout ($\rho_{\theta\chi}$). Results from analyses with an AR(1) structure for the covariance matrix of SM method and an unstructured covariance matrix for LRM method.

				no dropout		MCAR		MNAR					
								$\rho_{\theta\chi}=-0.4$		$\rho_{\theta\chi}=-0.7$		$\rho_{\theta\chi}=-0.9$	
N	J	ρ_θ	π	LRM	SM	LRM	SM	LRM	SM	LRM	SM	LRM	SM
100	4	0.4	0	0.054	0.054								
			0.05			0.054	0.050	0.066	0.066	0.060	0.054	0.038	0.038
			0.1			0.046	0.042	0.050	0.046	0.048	0.052	0.054	0.054
			0.2			0.050	0.048	0.072*	0.064	0.080*	0.078*	0.070	0.066
		0.7	0	0.044	0.040								
			0.05			0.048	0.042	0.040	0.034	0.042	0.044	0.036	0.034
			0.1			0.048	0.038	0.044	0.044	0.034	0.032	0.058	0.052
			0.2			0.068	0.060	0.052	0.060	0.052	0.046	0.076*	0.080*
		0.9	0	0.048	0.038								
			0.05			0.048	0.048	0.032	0.034	0.048	0.050	0.050	0.054
			0.1			0.040	0.046	0.024*	0.030*	0.032	0.038	0.034	0.036
			0.2			0.048	0.046	0.034	0.048	0.050	0.046	0.044	0.038
	7	0.4	0	0.044	0.036								
			0.05			0.054	0.048	0.066	0.058	0.048	0.050	0.040	0.038
			0.1			0.060	0.058	0.054	0.058	0.068	0.064	0.072*	0.066
			0.2			0.046	0.046	0.054	0.046	0.062	0.062	0.088*	0.080*
		0.7	0	0.058	0.046								
			0.05			0.056	0.046	0.038	0.038	0.054	0.052	0.056	0.042
			0.1			0.054	0.040	0.052	0.046	0.050	0.042	0.056	0.036
			0.2			0.052	0.054	0.064	0.044	0.090*	0.084*	0.090*	0.072*
		0.9	0	0.046	0.044								
			0.05			0.049	0.050	0.052	0.038	0.056	0.046	0.036	0.036
			0.1			0.038	0.034	0.046	0.046	0.048	0.040	0.064	0.060
			0.2			0.034	0.034	0.064	0.050	0.054	0.054	0.054	0.068
200	4	0.4	0	0.052	0.060								
			0.05			0.054	0.052	0.040	0.046	0.062	0.060	0.048	0.052
			0.1			0.052	0.044	0.066	0.058	0.080*	0.072*	0.046	0.048
			0.2			0.046	0.050	0.052	0.052	0.074*	0.076*	0.106*	0.104*
		0.7	0	0.056	0.048								
			0.05			0.042	0.038	0.040	0.042	0.080*	0.072*	0.050	0.044
			0.1			0.054	0.050	0.044	0.048	0.062	0.058	0.086*	0.078*
			0.2			0.038	0.040	0.044	0.044	0.050	0.056	0.088*	0.086*
		0.9	0	0.066	0.060								
			0.05			0.044	0.046	0.052	0.044	0.050	0.048	0.032	0.030*
			0.1			0.057	0.050	0.044	0.042	0.057	0.058	0.060	0.052
			0.2			0.046	0.046	0.063	0.052	0.069	0.070	0.074*	0.082*
	7	0.4	0	0.050	0.050								
			0.05			0.056	0.050	0.052	0.052	0.064	0.058	0.064	0.062
			0.1			0.042	0.038	0.064	0.054	0.066	0.052	0.058	0.052
			0.2			0.066	0.070	0.080*	0.072*	0.058	0.054	0.106*	0.100*
		0.7	0	0.048	0.036								
			0.05			0.054	0.054	0.066	0.052	0.046	0.042	0.068	0.066
			0.1			0.056	0.038	0.060	0.046	0.058	0.054	0.064	0.056
			0.2			0.052	0.032	0.084*	0.074*	0.086*	0.084*	0.080*	0.086*
		0.9	0	0.058	0.048								
			0.05			0.056	0.050	0.054	0.046	0.059	0.058	0.060	0.060
			0.1			0.044	0.028*	0.050	0.054	0.038	0.036	0.048	0.044
			0.2			0.053	0.046	0.059	0.040	0.088*	0.080*	0.103*	0.104*

* indicates that the 95% confidence interval of the type I error does not contain the expected value of 5%.

MNAR type occurs, type I errors increase with the proportion of dropout and the value of the correlation between the latent variable and the dropout propensity. As the values of these two parameters rise in absolute value, the number of 95% confidence intervals of the type I error that do not contain the expected 5% also rises. The type I error can reach 10% in the worst cases when $\pi = 20\%$ and $\rho_{\theta\chi} = -0.9$.

Table 2 shows the results of power of the test for time effect for both methods depending on the value of all simulation parameters: sample size, number of items, latent variable correlation, proportion of dropout and type of dropout. As observed for the type I errors, values of power are close to each other for LRM and SM method for fixed values of simulation parameters but the powers for LRM seem to be generally slightly higher than the power for SM when the correlation $\rho_\theta = 0.4$. LRM powers seem to be systematically slightly higher than the SM powers when the correlation $\rho_\theta = 0.7$ or 0.9. In these cases, the McNemar's tests are always significant at 5% concluding that LRM power is higher than SM power. Powers increases with the sample size, number of items and latent variable correlation. For example, powers range from 38.8% and 38.4% for LRM and SM respectively when $N = 100$, $J = 4$, $\rho_\theta = 0.4$ and $\pi = 0$ to 95.5% and 91.8% for LRM and SM respectively when $N = 200$, $J = 7$, $\rho_\theta = 0.9$ and $\pi = 0$. Powers of data presenting MCAR dropout are lower than the corresponding powers of complete data. This loss of power is highest when the proportion of dropout and the latent variable correlation are high, $\pi = 20\%$ and $\rho_\theta = 0.9$. The fall is up to -15.7% and -20.8% for LRM and SM respectively when $N = 100$, $J = 7$, $\rho_\theta = 0.9$ and $\pi = 20\%$. Powers of data with MNAR dropout are higher than powers of complete data. As the proportion of dropout and the informativity of the missing data (as ρ_{θ_χ} decreases to -0.9) increase, the powers for MNAR case get higher than for complete case.

3.2 Time effect estimation

The observed effect of the proportion and the informativity of dropout on type I error and power suggest a bias of the time effect estimation. Figure 1 shows the estimation of the time effect between the two first times of evaluation for different values of sample size, number of items and latent variable correlation when the simulated time effect on the latent trait was null. Figure 1 shows that time effect seems to be well estimated for complete data and data with MCAR dropout since all bias estimations for these cases are close to 0. For data with MNAR dropout, LRM and SM seem to overestimate the time effect. The overestimation is more marked when the latent variable correlation is low. The bias produced by the SM method depends on the number of items J. When $J = 4$, the overestimation for SM is lower than when $J = 7$. As a consequence, the overestimation for SM is lower than for LRM when $J = 4$ and higher than for LRM when $J = 7$.

Figure 2 shows the estimation of the time effect between the two first times of evaluation for different values of sample size, number of items and latent variable correlation when the simulated time effect on the latent trait was equal to 0.2. As described before, the time effect on score was estimated to 0.15 for $J = 4$ and to 0.25 for $J = 7$. On figure 2, the estimations for

Table 2: Power of the tests for Score Mixed model (SM) and Longitudinal Rasch Mixed model (LRM) methods for different values of sample size (N), number of items (J), latent variable correlation (ρ_θ), proportion of dropout (π) and type of dropout ($\rho_{\theta\chi}$). Results from analyses with an AR(1) structure for the covariance matrix of SM method and an unstructured covariance matrix for LRM method.

				no dropout		MCAR		MNAR					
								$\rho_{\theta\chi}=-0.4$		$\rho_{\theta\chi}=-0.7$		$\rho_{\theta\chi}=-0.9$	
N	J	ρ_θ	π	LRM	SM	LRM	SM	LRM	SM	LRM	SM	LRM	SM
100	4	0.4	0	0.388	0.384								
			0.05			0.399	0.400	0.356	0.354	0.383♯	0.360	0.404	0.404
			0.1			0.335	0.320	0.411	0.408	0.439	0.426	0.440	0.426
			0.2			0.325	0.316	0.384	0.376	0.474	0.464	0.513	0.508
		0.7	0	0.423♯	0.372								
			0.05			0.439♯	0.382	0.426♯	0.390	0.457♯	0.400	0.498♯	0.448
			0.1			0.441♯	0.406	0.453♯	0.400	0.480	0.464	0.499♯	0.466
			0.2			0.399♯	0.368	0.399♯	0.378	0.486♯	0.466	0.543♯	0.524
		0.9	0	0.508♯	0.424								
			0.05			0.471♯	0.390	0.505♯	0.432	0.511♯	0.432	0.560♯	0.474
			0.1			0.462♯	0.398	0.427♯	0.366	0.511♯	0.436	0.508♯	0.462
			0.2			0.379♯	0.342	0.477♯	0.434	0.509	0.496	0.598♯	0.578
	7	0.4	0	0.470♯	0.428								
			0.05			0.454♯	0.424	0.488♯	0.468	0.504	0.492	0.518♯	0.496
			0.1			0.482♯	0.454	0.522♯	0.494	0.510♯	0.496	0.556	0.546
			0.2			0.398	0.396	0.500	0.482	0.534♯	0.514	0.600	0.588
		0.7	0	0.568♯	0.522								
			0.05			0.558♯	0.502	0.586♯	0.548	0.598♯	0.534	0.626♯	0.576
			0.1			0.541♯	0.470	0.549♯	0.510	0.623♯	0.574	0.607♯	0.562
			0.2			0.446♯	0.414	0.568♯	0.518	0.655♯	0.632	0.685♯	0.658
		0.9	0	0.688♯	0.564								
			0.05			0.670♯	0.548	0.679♯	0.604	0.723♯	0.622	0.725♯	0.620
			0.1			0.635♯	0.526	0.686♯	0.584	0.743♯	0.654	0.719♯	0.632
			0.2			0.531♯	0.452	0.687♯	0.590	0.745♯	0.674	0.737♯	0.684
200	4	0.4	0	0.654♯	0.634								
			0.05			0.654	0.644	0.682	0.668	0.718♯	0.690	0.678♯	0.658
			0.1			0.631	0.628	0.658	0.646	0.738	0.726	0.718	0.714
			0.2			0.552	0.540	0.694	0.688	0.745	0.736	0.820	0.818
		0.7	0	0.721♯	0.678								
			0.05			0.729♯	0.694	0.747♯	0.700	0.787♯	0.742	0.753♯	0.702
			0.1			0.701♯	0.670	0.752♯	0.724	0.806♯	0.772	0.825♯	0.786
			0.2			0.669♯	0.626	0.745♯	0.710	0.843♯	0.830	0.820♯	0.808
		0.9	0	0.826♯	0.756								
			0.05			0.765♯	0.686	0.807♯	0.736	0.809♯	0.764	0.823♯	0.744
			0.1			0.768♯	0.704	0.797♯	0.740	0.857♯	0.784	0.847♯	0.806
			0.2			0.698♯	0.616	0.781♯	0.744	0.861♯	0.846	0.875	0.872
	7	0.4	0	0.822♯	0.810								
			0.05			0.758	0.750	0.778	0.778	0.826♯	0.812	0.798♯	0.782
			0.1			0.778♯	0.760	0.794	0.788	0.826	0.826	0.846	0.832
			0.2			0.696	0.698	0.812♯	0.786	0.874	0.870	0.896	0.894
		0.7	0	0.916♯	0.880								
			0.05			0.858♯	0.826	0.882♯	0.846	0.892♯	0.860	0.880♯	0.850
			0.1			0.814♯	0.786	0.894♯	0.878	0.904♯	0.870	0.936♯	0.910
			0.2			0.776♯	0.732	0.876♯	0.856	0.900♯	0.890	0.958♯	0.944
		0.9	0	0.955♯	0.918								
			0.05			0.935♯	0.882	0.966♯	0.932	0.946♯	0.910	0.938♯	0.912
			0.1			0.931♯	0.876	0.956♯	0.932	0.946♯	0.932	0.962♯	0.942
			0.2			0.893♯	0.810	0.922♯	0.900	0.956♯	0.942	0.960♯	0.950

♯ indicates that LRM power is significantly higher than SM power at 5% with a McNemar's test.

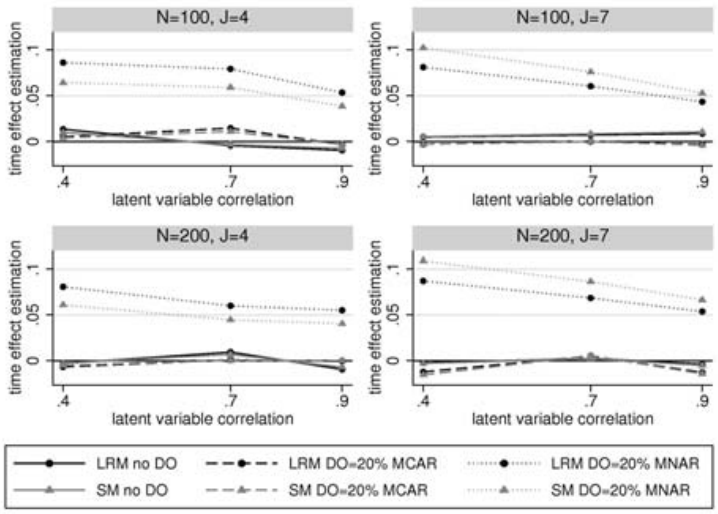

Figure 1: Time effect estimations between time 1 and time 2 for Score and Mixed models (SM) and Longitudinal Rasch Mixed model (LRM) methods for different values of sample size (N), number of items (J) and latent variable correlation (ρ_θ). No time effect simulated ($d_\theta = 0$). Data without dropout (DO), with 20% of MCAR dropout or 20% of MNAR dropout. Analyses performed with an AR(1) structure for the covariance matrix in SM method and an unstructured covariance matrix for LRM method.

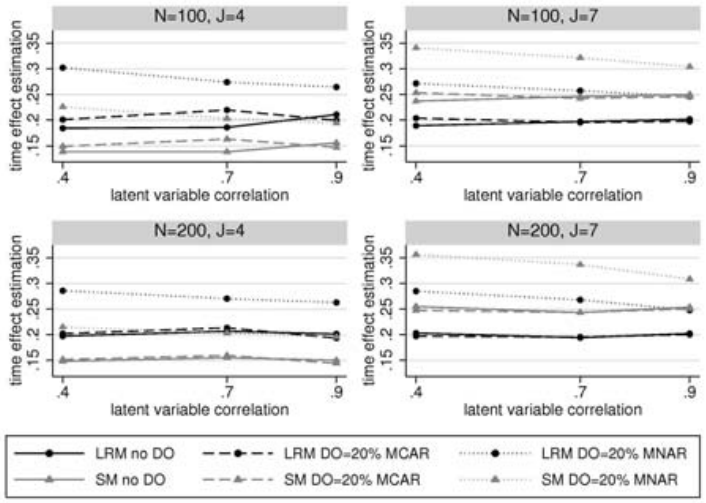

Figure 2: Time effect estimations between time 1 and time 2 for Score and Mixed models (SM) and Longitudinal Rasch Mixed model (LRM) methods for different values of sample size (N), number of items (J) and latent variable correlation (ρ_θ). Time effect simulated ($d_\theta = 0.2$). Data without dropout (DO), with 20% of MCAR dropout or 20% of MNAR dropout. Analyses performed with an AR(1) structure for the covariance matrix in SM method and an unstructured covariance matrix for LRM method.

complete data and data with MCAR dropout are also unbiased whatever the method. LRM and SM methods overestimate the time effect when MNAR dropout occurs compared to estimations on complete data.
The overestimation of the time effect grows with the proportion of dropout (results not shown). When $\pi = 5\%$, the time effect is overestimated for half of the cases with $\rho_\theta = -0.7$ or $\rho_\theta = -0.9$. At 10% of dropout, all the cases show an overestimation of time effect when $\rho_\theta = -0.7$ or $\rho_\theta = -0.9$. At 20% of dropout, all the cases show an overestimation of time effect when $\rho_\theta = -0.4$, $\rho_\theta = -0.7$ or $\rho_\theta = -0.9$.
Results on time effect estimation between time 2 and time 3 are comparable to results on time effect estimation between time 1 and time 2 (results not shown).

4 Discussion

Patient-Reported Outcomes are increasingly used in health sciences. Two methods to analyze longitudinal latent variables following a Rasch model were compared: Score and Mixed models (SM) and Longitudinal Rasch Model (LRM) methods. They have shown comparable results in term of type I error. LRM seemed to generally have a slightly but significantly higher power than SM when the latent variable correlation was medium or high. As expected, powers increased with the sample size, number of items and latent variable correlation.
Before choosing what method to use, statisticians have to keep in mind the underlying hypotheses and properties of each method. Although CTT has the advantage to be simple to use and to interpret, this approach produces ordinal measures (Embretson, 1996). In fact, only ordinal level measurement can be achieved with CTT, that is we can only observe that a person has a higher score than another. But the distance between two scores has no meaning because it depends on the population and the set of items. The Rasch model rests on strong assumptions but presents interesting psychometric properties: the exhaustivity of the score on the latent trait, the specific objectivity and the interval-level measurement. The interval-level measurement property means that the relative distance between two latent traits is maintained across questionnaires of different difficulties. In the Rasch model, the comparisons of the distance between pairs of person that have answered different set of items are possible.

Many assumptions had to be made to perform this simulation study. First, the data were assumed to follow a Rasch model because the purpose of this study was to compare CTT-based and Rasch-based models to determine which approach is the most adequate to analyze longitudinal latent variables when the data actually follow a Rasch model. The results of this study are only valid in this context. The choice of the Rasch model among all possible IRT models comes from its psychometric properties and its wide use in development and validation.
Second, the longitudinal Rasch mixed model used to simulate the data assumes a linear time effect for the evolution of the outcome and item parameters constant with time. It is reasonable to assume a linear time effect for long-term studies of quality of life occurring after treatment completion and where the quality of life is expected to improve. However, many studies include time points before and after treatment. With measurements before the beginning of the treat-

ment, at the end of the treatment and during follow-up period, the quality of life will probably decrease during the treatment period and increase after the completion of treatment leading to a non-linear time effect. Both methods estimate time effect for each period between two consecutive times without assuming that time effect is linear. LRM and SM can be used in the case where time effect decreases first and increases afterwards and we could expect similar results than for linear time effect.

This study aimed at evaluating the impact of type and proportion of dropout on CTT-based and Rasch-based approaches. For complete data and data with MCAR dropout, the type I error rates were well maintained to the expected 5%. As expected, MCAR case have shown close results to complete case due to the use of methods based on likelihood. However, a loss of power as well as an underestimation of the time effect has been observed for MCAR data in certain cases.

In the case where the missing data are non-ignorable, both methods have shown poor results. The type I error rates were not maintained to the expected 5% and could reach 10% in the worst cases, that is when the proportion and the informativity of dropout were high. Powers were also higher than complete case and increased with the proportion and informativity of the dropout as a direct result of the overestimation of the time effect. When MNAR data are encountered, both methods can't be used without taking into account the dropout process in the analysis. Several methods have been proposed in the litterature to handle missing data such as imputation, selection models and pattern-mixture models. Imputation consists in filling in the missing values using observed values. Several simple imputation methods exist based on information on the same subjects or from other subjects. Most of them lead to biased estimates and underestimation of the variability. In multiple imputation (Rubin, 2004), each missing value is replaced by several imputed values. Each imputed dataset is then analysed and the results are combined. This method overcomes the problems encountered with single imputation. Selection and pattern-mixture models are likelihood-based methos for handling missing data. These models combine linear model for the response with a suitable dropout model. Selection models and pattern-mixture models (Little, 1995) were proposed to model nonignorable nonresponse. They are an interesting way of dealing with MNAR missing data process by modeling explicitly the missing data mechanism. These models have to be used carefully because untestable assumptions have to be made on the missing data process for selection models and untestable identifying restrictions are used in pattern-mixture models to ensure their identifiability. For the use of such models, it is recommended to consider estimation coming from many selection/pattern-mixture models with different assumptions rather than only one model.

Missing at random (MAR) mechanism was not examined in this study. The dropout is said to be MAR if the dropout process does not depend on the missing values but only on the observed values. To simulate MAR dropout, the dropout propensity at time t ($\chi^{(t)}$) should be independent of the value of the latent variable at time t ($\theta^{(t)}$) and so $\rho_{\theta\chi}$ should be equal to 0. Furthermore, the dropout propensity at time t should be dependent on the previous values of latent variable already observed. The model could be extended by including it as a covariate. The MAR miss-

ing mechanism is ignorable, such as MCAR missing mechanism, if the separability condition is met (Verbeke and Molenberghs, 2000b). Both methods are likelihood-based analysis and ignores the dropout mechanism, we can expect that LRM and SM will lead to a valid analysis such as for MCAR case, provided that the parameters describing the measurement process are functionally independent of the parameters describing the dropout process.

This article focused on one type of missing data: the dropout. An important further development to this study concerns intermittent missing data. The missing data are intermittent if the patient has not answered all the items of the questionnaire. As dropout, intermittent missing data can lead to a loss of power and possible bias. The causes for intermittent missing data are multiple. For example, the patient may have not seen the question and the value is missing completely at random. The item can bother the patient because its content concerns religion, politics, sexual life. Thus, the patient choose not to answer this particular item and the missing data is informative. The informative dropout seems to be linked with the quality of life level of the patient whereas informative intermittent missing data might be related to the characteristics of the item. The ways to deal with intermittent missing data are complete case analysis, available case analysis, imputation. The complete case analysis only includes measurements that are complete in the analysis. The available case analysis uses as much data as possible to take advantage of all available information. Imputation methods can be simple or multiple. The problem that arises with intermittent missing data in CTT is the computation of the score. The most commonly used questionnaires (SF-36, EORTC QLQ-C30) have general guidelines regarding treatment of missing data. Generally, the score can still be computed if the patient has filled in half or more of the items of the scale. The patients with more missing items can't be used in the analysis. The Rasch model uses all available items in the analysis. In studies with a high proportion of intermittent missing data, we can expect that Rasch-based approach perform better than CTT-based approach because Rasch model may use more information than CTT. Furthermore, the property of specific objectivity of the Rasch model may ensure that the latent variable may be estimated consistently even for patients with missing items. We can expect that the occurrence of ignorable intermittent missing data (MCAR and MAR) will lead to valid analysis because both approaches are based on likelihood. Estimation problems will probably be observed for non-ignorable dropout but maybe in different extent for each approach.

In health sciences, longitudinal studies evaluating PRO often include two or more groups of patients in order to compare the evolution of the outcome between the groups. For example, study of a new strategy of treatment for a cancer may compare the long-term quality of life of patients receiving the new strategy and the long-term quality of life of patients who have received the usual strategy. This study has to be extended to compare both approaches in the context of longitudinal PRO collected in different groups of patients.

The purpose of this study was to compare CTT-based and Rasch-based models to determine which approach is the most adequate to analyze longitudinal latent variables when the

data were actually collected from a scale validated with a Rasch model. Data with informative dropout have shown parameter estimation problems. They have to be analyzed with appropriate methods taking into account of the dropout process such as selection models or pattern-mixture models. For complete data and data with non-informative dropout, the method of analysis based on the Rasch model may be preferred than the method based on CTT due to the generally observed slight gain of power and the psychometric properties of the Rasch model.

Acknowledgment

This work was supported by the Ligue Nationale Contre le Cancer.

References

Akaike, H. 1974. A new look at the statistical model identification, *IEEE Transactions on Automatic Control* **19**: 716–723.

Cella, D. F., Dineen, K., Arnason, B., Reder, A., Webster, K. A., Karabatsos, G., Chang, C., Lloyd, S., Steward, J. and Stefoski, D. 1996. Validation of the functional assessment of multiple sclerosis quality of life instrument, *Neurology* **47**(1): 129–139.

Diggle, P. and Kenward, M. G. 1994. Informative Drop-Out in longitudinal data analysis, *Applied Statistics* **43**(1): 49–93.

Embretson, S. E. 1996. The new rules of measurement, *Psychological Assessment* **8**(4): 341–349.

Fischer, G. H. and Molenaar, I. W. 1995. *Rasch models*, Springer.

Fitzmaurice, G. M., Davidian, M., Verbeke, G. and Molenberghs, G. 2009. *Longitudinal data analysis*, Chapman and Hall/CRC.

Hambleton, R. K. 2000. Emergence of item response modeling in instrument development and data analysis, *Medical Care* **38**(9 Suppl II): 60–65.

Lai, J., Cella, D., Kupst, M. J., Holm, S., Kelly, M. E., Bode, R. K. and Goldman, S. 2007. Measuring fatigue for children with cancer: Development and validation of the pediatric functional assessment of chronic illness Therapy-Fatigue (pedsFACIT-F), *Journal of Pediatric Hematology/Oncology* **29**(7): 471–479.

Laird, N. M. and Ware, J. H. 1982. Random-Effects models for longitudinal data, *Biometrics* **38**(4): 963–974.

Littell, R. C., Milliken, G. A., Stroup, W. W. and Wolfinger, R. D. 1996. *SAS system for mixed models*, SAS Institute, Inc., Cary, NC.

Little, R. J. A. 1995. Modeling the Drop-Out mechanism in Repeated-Measures studies, *Journal of the American Statistical Association* **90**(431): 1112–1121.

Little, R. J. A. and Rubin, D. B. 2002. *Statistical analysis with missing data*, Wiley.

McNemar, Q. 1947. Note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika* **12**(2): 153–157.

Meiser, T. 2007. Rasch models for longitudinal data, *in* M. von Davier and C. H. Carstensen (eds), *Multivariate and Mixture Distribution Rasch Models*, Springer, New York, pp. 191–199.

Rasch, G. 1980. *Probabilistic models for some intelligence and attainment tests*, University of Chicago Press.

Rubin, D. B. 2004. *Multiple imputation for nonresponse in surveys*, Wiley-IEEE.

Sijtsma, K. and Hemker, B. T. 2000. A taxonomy of IRT models for ordering persons and items using simple sum scores, *Journal of Educational and Behavioral Statistics* **25**(4): 391–415.

Troxel, A. B., Fairclough, D. L., Curran, D. and Hahn, E. A. 1998. Statistical analysis of quality of life with missing data in cancer clinical trials, *Statistics in Medicine* **17**(5-7): 653–66.

Verbeke, G. and Molenberghs, G. 2000a. Inference for the marginal model, *Linear Mixed Models for Longitudinal Data*, Springer, New York, pp. 55–76.

Verbeke, G. and Molenberghs, G. 2000b. *Linear Mixed Models for Longitudinal Data*, corrected edn, Springer.

Zheng, X. and Rabe-Hesketh, S. 2007. Estimating parameters of dichotomous and ordinal item response models with gllamm, *Stata Journal* **7**(3): 313–333.

Annexe C

Comparaison de méthodes d'analyse de données subjectives longitudinales sujettes au dropout : Résultats complets

TABLE C.1 – Estimations de l'effet temps entre le temps 1 et le temps 2 ($\hat{d}_{12}$) pour les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J), de la corrélation de la variable latente (ρ_θ), de la proportion de dropout ($\pi^{(t)}$) et du type de dropout ($\rho_{\theta\chi}$). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour la méthode SM et une structure UN pour la méthode LRM. Valeurs moyennes de $\hat{d}_{12}$ et écarts-types (s.d.). Effet temps simulé : $d_\theta = 0$. $d_S = 0$.

				no dropout				MCAR				MNAR											
												$\rho_{\theta\chi} = -0.4$				$\rho_{\theta\chi} = -0.7$				$\rho_{\theta\chi} = -0.9$			
				LRM		SM		LRM		SM		LRM		SM		LRM		SM		LRM		SM	
N	J	ρ_θ	$\pi^{(t)}$	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)
100	4	0,4	0	0,013	(0,200)	0,010	(0,149)																
			0,05					-0,009	(0,210)	-0,007	(0,156)	0,014	(0,199)	0,011	(0,150)	0,008	(0,193)	0,006	(0,144)	0,019*	(0,198)	0,014*	(0,149)
			0,1					-0,008	(0,205)	-0,005	(0,153)	0,016	(0,206)	0,013	(0,154)	0,033*	(0,205)	0,026*	(0,153)	0,026*	(0,203)	0,020*	(0,151)
			0,2					0,005	(0,217)	0,004	(0,160)	0,052*	(0,213)	0,038*	(0,159)	0,065*	(0,214)	0,049*	(0,159)	0,086*	(0,210)	0,064*	(0,156)
		0,7	0	-0,004	(0,188)	-0,003	(0,139)																
			0,05					0,011	(0,172)	0,007	(0,128)	0,002	(0,190)	0,001	(0,142)	0,007	(0,185)	0,005	(0,139)	0,022*	(0,179)	0,017*	(0,133)
			0,1					-0,003	(0,186)	-0,002	(0,138)	0,008	(0,193)	0,007	(0,144)	0,030*	(0,181)	0,023*	(0,135)	0,028*	(0,182)	0,019*	(0,136)
			0,2					0,015	(0,197)	0,011	(0,145)	0,035*	(0,207)	0,026*	(0,153)	0,069*	(0,190)	0,051*	(0,141)	0,079*	(0,209)	0,059*	(0,154)
		0,9	0	-0,010	(0,167)	-0,007	(0,124)																
			0,05					0,001	(0,178)	0,000	(0,133)	0,014	(0,170)	0,010	(0,126)	0,007	(0,176)	0,005	(0,130)	0,017*	(0,172)	0,012*	(0,127)
			0,1					0,001	(0,178)	0,000	(0,129)	0,008	(0,171)	0,007	(0,126)	0,016*	(0,172)	0,012*	(0,128)	0,023*	(0,189)	0,017*	(0,139)
			0,2					-0,004	(0,193)	-0,003	(0,141)	0,042*	(0,197)	0,031*	(0,146)	0,040*	(0,184)	0,028*	(0,136)	0,054*	(0,181)	0,039*	(0,133)
	7	0,4	0	0,005	(0,165)	0,005	(0,206)																
			0,05					0,004	(0,169)	0,006	(0,211)	0,016*	(0,174)	0,020*	(0,219)	0,011	(0,162)	0,014	(0,201)	0,001	(0,166)	0,001	(0,208)
			0,1					0,014	(0,182)	0,017	(0,226)	0,012	(0,168)	0,015	(0,210)	0,022*	(0,171)	0,029*	(0,213)	0,039*	(0,172)	0,049*	(0,214)
			0,2					-0,003	(0,168)	-0,003	(0,211)	0,038*	(0,173)	0,047*	(0,215)	0,080*	(0,178)	0,098*	(0,221)	0,081*	(0,186)	0,102*	(0,232)
		0,7	0	0,007	(0,155)	0,008	(0,192)																
			0,05					-0,004	(0,153)	-0,004	(0,191)	0,008	(0,140)	0,010	(0,174)	0,016*	(0,149)	0,021*	(0,186)	0,012	(0,151)	0,015	(0,187)
			0,1					-0,010	(0,155)	-0,012	(0,193)	0,006	(0,160)	0,006	(0,199)	0,013	(0,148)	0,017*	(0,185)	0,031*	(0,162)	0,038*	(0,201)
			0,2					0,000	(0,158)	0,000	(0,196)	0,046*	(0,165)	0,057*	(0,207)	0,054*	(0,165)	0,068*	(0,205)	0,061*	(0,161)	0,076*	(0,200)
		0,9	0	0,009	(0,127)	0,011	(0,157)																
			0,05					-0,011	(0,135)	-0,012	(0,168)	-0,005	(0,140)	-0,007	(0,175)	0,013*	(0,141)	0,016*	(0,176)	-0,002	(0,132)	-0,004	(0,164)
			0,1					0,004	(0,139)	0,006	(0,173)	-0,009	(0,140)	-0,012	(0,175)	0,018*	(0,134)	0,023*	(0,167)	0,014*	(0,131)	0,018*	(0,163)
			0,2					-0,003	(0,140)	-0,004	(0,172)	0,020*	(0,145)	0,023*	(0,180)	0,045*	(0,143)	0,055*	(0,177)	0,043*	(0,150)	0,053*	(0,185)

Suite page suivante...

Table C.1 – Suite

				no dropout				MCAR				MNAR									
												$\rho_{\theta\chi} = -0.4$				$\rho_{\theta\chi} = -0.7$				$\rho_{\theta\chi} = -0.9$			
				LRM		SM		LRM		SM		LRM		SM		LRM		SM		LRM		SM	
N	J	ρ_θ	$\pi^{(t)}$	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)
200	4	0,4	0	-0,003	(0,140)	-0,003	(0,105)																
			0,05					0,005	(0,142)	0,003	(0,107)	0,003	(0,145)	0,002	(0,108)	0,008	(0,148)	0,006	(0,110)	0,019*	(0,137)	0,014*	(0,104)
			0,1					0,004	(0,141)	0,003	(0,106)	0,022*	(0,140)	0,016*	(0,106)	0,020*	(0,149)	0,015*	(0,112)	0,045*	(0,144)	0,033*	(0,108)
			0,2					-0,007	(0,142)	-0,005	(0,106)	0,035*	(0,147)	0,026*	(0,110)	0,052*	(0,152)	0,039*	(0,114)	0,081*	(0,156)	0,061*	(0,117)
		0,7	0	0,009	(0,136)	0,007	(0,102)																
			0,05					0,003	(0,126)	0,002	(0,095)	0,001	(0,133)	0,002	(0,100)	0,011	(0,135)	0,008	(0,101)	0,019*	(0,123)	0,015*	(0,092)
			0,1					0,003	(0,137)	0,002	(0,103)	0,013*	(0,132)	0,010*	(0,099)	0,023*	(0,133)	0,017*	(0,099)	0,023*	(0,140)	0,018*	(0,104)
			0,2					0,001	(0,135)	0,000	(0,101)	0,033*	(0,139)	0,025*	(0,104)	0,048*	(0,145)	0,036*	(0,109)	0,060*	(0,136)	0,045*	(0,102)
		0,9	0	-0,009	(0,125)	-0,007	(0,093)																
			0,05					0,003	(0,130)	0,002	(0,096)	0,001	(0,121)	0,001	(0,091)	0,013*	(0,120)	0,010*	(0,089)	0,010	(0,116)	0,008*	(0,087)
			0,1					0,002	(0,129)	0,001	(0,096)	0,002	(0,123)	0,001	(0,091)	0,024*	(0,128)	0,018*	(0,096)	0,019*	(0,121)	0,014*	(0,090)
			0,2					0,000	(0,134)	0,000	(0,100)	0,030*	(0,128)	0,022*	(0,095)	0,044*	(0,132)	0,032*	(0,099)	0,055*	(0,128)	0,041*	(0,095)
	7	0,4	0	-0,002	(0,121)	-0,003	(0,151)																
			0,05					-0,002	(0,125)	-0,002	(0,156)	-0,001	(0,123)	-0,001	(0,154)	0,017*	(0,120)	0,021*	(0,149)	0,019*	(0,120)	0,023*	(0,150)
			0,1					0,006	(0,119)	0,007	(0,149)	0,007	(0,116)	0,010	(0,146)	0,029*	(0,123)	0,036*	(0,154)	0,029*	(0,120)	0,037*	(0,150)
			0,2					-0,012*	(0,130)	-0,015*	(0,163)	0,038*	(0,126)	0,047*	(0,158)	0,063*	(0,124)	0,079*	(0,157)	0,087*	(0,120)	0,109*	(0,150)
		0,7	0	0,003	(0,103)	0,002	(0,129)																
			0,05					-0,005	(0,110)	-0,006	(0,138)	-0,001	(0,112)	-0,002	(0,140)	0,010*	(0,109)	0,013*	(0,137)	0,006	(0,114)	0,007	(0,143)
			0,1					0,001	(0,109)	0,002	(0,137)	0,016*	(0,117)	0,020*	(0,145)	0,018*	(0,107)	0,023*	(0,133)	0,027*	(0,112)	0,034*	(0,139)
			0,2					0,005	(0,117)	0,006	(0,145)	0,033*	(0,117)	0,041*	(0,146)	0,054*	(0,116)	0,068*	(0,146)	0,069*	(0,109)	0,086*	(0,137)
		0,9	0	-0,004	(0,095)	-0,005	(0,118)																
			0,05					-0,004	(0,092)	-0,005	(0,114)	-0,002	(0,095)	-0,002	(0,118)	0,010*	(0,093)	0,012*	(0,117)	0,006	(0,102)	0,007	(0,127)
			0,1					0,000	(0,097)	0,000	(0,121)	0,008	(0,103)	0,009	(0,128)	0,017*	(0,097)	0,020*	(0,121)	0,019*	(0,102)	0,024*	(0,126)
			0,2					-0,013*	(0,105)	-0,014*	(0,131)	0,025*	(0,098)	0,031*	(0,123)	0,042*	(0,105)	0,052*	(0,131)	0,054*	(0,107)	0,066*	(0,133)

* indique que le test de Student est significatif à 5% ($H_0 : \mu_{\hat{d}_{12}} = 0$)

TABLE C.2 – Estimations de l'effet temps entre le temps 1 et le temps 2 ($\hat{d}_{12}$) pour les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J), de la corrélation de la variable latente (ρ_θ), de la proportion de dropout ($\pi^{(t)}$) et du type de dropout (ρ_{θ_χ}). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour la méthode SM et une structure UN pour la méthode LRM. Valeurs moyennes de $\hat{d}_{12}$ et écarts-types (s.d.). Effet temps simulé : $d_\theta = 0,2$. $d_S = 0.15$ pour J=4, $d_S = 0.25$ pour J=7.

				no dropout				MCAR				MNAR											
												$\rho_{\theta\chi} = -0.4$				$\rho_{\theta\chi} = -0.7$				$\rho_{\theta\chi} = -0.9$			
				LRM		SM		LRM		SM		LRM		SM		LRM		SM		LRM		SM	
N	J	ρ_θ	$\pi^{(t)}$	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)
100	4	0,4	0	0,184	(0,208)	0,139	(0,156)																
			0,05					0,198	(0,206)	0,149	(0,154)	0,199	(0,205)	0,148	(0,151)	0,210	(0,194)	0,156	(0,143)	0,215	(0,194)	0,162	(0,145)
			0,1					0,185	(0,215)	0,137	(0,159)	0,217	(0,197)	0,162	(0,147)	0,236*	(0,207)	0,174*	(0,154)	0,235*	(0,206)	0,175*	(0,153)
			0,2					0,201	(0,210)	0,149	(0,156)	0,233*	(0,219)	0,173*	(0,162)	0,283*	(0,210)	0,210*	(0,155)	0,302*	(0,205)	0,226*	(0,154)
		0,7	0	0,186	(0,184)	0,138	(0,136)																
			0,05					0,210	(0,189)	0,157	(0,141)	0,195	(0,199)	0,145	(0,147)	0,221*	(0,194)	0,165*	(0,144)	0,223*	(0,191)	0,166*	(0,141)
			0,1					0,207	(0,200)	0,153	(0,147)	0,212	(0,183)	0,158	(0,136)	0,245*	(0,197)	0,183*	(0,147)	0,241*	(0,201)	0,180*	(0,148)
			0,2					0,220*	(0,194)	0,163*	(0,144)	0,222*	(0,202)	0,166*	(0,149)	0,255*	(0,186)	0,188*	(0,137)	0,274*	(0,209)	0,203*	(0,154)
		0,9	0	0,211	(0,171)	0,156	(0,126)																
			0,05					0,194	(0,175)	0,143	(0,128)	0,202	(0,170)	0,15	(0,125)	0,216	(0,184)	0,160	(0,136)	0,226*	(0,167)	0,167*	(0,123)
			0,1					0,211	(0,192)	0,157	(0,139)	0,205	(0,178)	0,152	(0,130)	0,228*	(0,177)	0,169*	(0,132)	0,223*	(0,182)	0,164*	(0,134)
			0,2					0,200	(0,193)	0,147	(0,142)	0,231*	(0,183)	0,170*	(0,135)	0,260*	(0,182)	0,191*	(0,133)	0,265*	(0,197)	0,194*	(0,144)
	7	0,4	0	0,189	(0,165)	0,237	(0,205)																
			0,05					0,199	(0,170)	0,249	(0,212)	0,208	(0,176)	0,26	(0,220)	0,222*	(0,183)	0,277*	(0,228)	0,222*	(0,165)	0,276*	(0,205)
			0,1					0,201	(0,184)	0,252	(0,228)	0,222*	(0,171)	0,276*	(0,211)	0,230*	(0,183)	0,288*	(0,229)	0,239*	(0,170)	0,298*	(0,213)
			0,2					0,204	(0,186)	0,253	(0,230)	0,246*	(0,171)	0,307*	(0,211)	0,247*	(0,190)	0,307*	(0,237)	0,272*	(0,191)	0,341*	(0,237)
		0,7	0	0,197	(0,147)	0,246	(0,182)																
			0,05					0,205	(0,155)	0,257	(0,194)	0,204	(0,151)	0,254	(0,188)	0,207	(0,148)	0,259	(0,184)	0,213	(0,153)	0,266	(0,190)
			0,1					0,196	(0,150)	0,246	(0,187)	0,206	(0,157)	0,258	(0,195)	0,220*	(0,158)	0,276*	(0,197)	0,228*	(0,155)	0,285*	(0,192)
			0,2					0,196	(0,165)	0,243	(0,205)	0,229*	(0,153)	0,286*	(0,190)	0,248*	(0,161)	0,310*	(0,201)	0,257*	(0,160)	0,322*	(0,199)
		0,9	0	0,202	(0,130)	0,25	(0,161)																
			0,05					0,194	(0,135)	0,241	(0,168)	0,209	(0,136)	0,26	(0,166)	0,214*	(0,138)	0,266*	(0,171)	0,208	(0,139)	0,259	(0,170)
			0,1					0,194	(0,139)	0,243	(0,172)	0,204	(0,135)	0,255	(0,166)	0,229*	(0,146)	0,284*	(0,182)	0,218*	(0,137)	0,271*	(0,170)

Suite page suivante...

Table C.2 – Suite

				no dropout				MCAR				MNAR											
												$\rho_{\theta\chi} = -0.4$				$\rho_{\theta\chi} = -0.7$				$\rho_{\theta\chi} = -0.9$			
				LRM		SM		LRM		SM		LRM		SM		LRM		SM		LRM		SM	
N	J	ρ_θ	$\pi^{(t)}$	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)
			0,2					0,198	(0,147)	0,246	(0,182)	0,228*	(0,147)	0,283*	(0,182)	0,244*	(0,151)	0,302*	(0,186)	0,246*	(0,148)	0,304*	(0,183)
200	4	0,4	0	0,197	(0,141)	0,148	(0,105)																
			0,05					0,196	(0,156)	0,147	(0,117)	0,213*	(0,140)	0,159	(0,104)	0,211	(0,143)	0,159	(0,107)	0,217*	(0,144)	0,163*	(0,108)
			0,1					0,202	(0,142)	0,152	(0,106)	0,205	(0,139)	0,154	(0,104)	0,225*	(0,142)	0,168*	(0,107)	0,232*	(0,150)	0,175*	(0,113)
			0,2					0,202	(0,144)	0,152	(0,108)	0,231*	(0,139)	0,173*	(0,105)	0,261*	(0,155)	0,196*	(0,115)	0,286*	(0,147)	0,215*	(0,111)
		0,7	0	0,207	(0,119)	0,155	(0,089)																
			0,05					0,201	(0,132)	0,151	(0,098)	0,211	(0,132)	0,157	(0,098)	0,203	(0,132)	0,153	(0,099)	0,214*	(0,135)	0,161*	(0,101)
			0,1					0,203	(0,136)	0,152	(0,102)	0,217*	(0,138)	0,163*	(0,103)	0,234*	(0,133)	0,176*	(0,100)	0,239*	(0,132)	0,179*	(0,099)
			0,2					0,213*	(0,145)	0,160*	(0,108)	0,236*	(0,132)	0,176*	(0,098)	0,255*	(0,134)	0,190*	(0,099)	0,270*	(0,146)	0,203*	(0,109)
		0,9	0	0,202	(0,126)	0,15	(0,094)																
			0,05					0,195	(0,121)	0,145	(0,090)	0,207	(0,123)	0,155	(0,092)	0,21	(0,125)	0,156	(0,093)	0,211*	(0,122)	0,158	(0,091)
			0,1					0,199	(0,126)	0,149	(0,094)	0,207	(0,124)	0,154	(0,092)	0,227*	(0,124)	0,169*	(0,092)	0,229*	(0,137)	0,171*	(0,102)
			0,2					0,193	(0,126)	0,144	(0,094)	0,229*	(0,131)	0,170*	(0,098)	0,249*	(0,133)	0,185*	(0,098)	0,263*	(0,132)	0,196*	(0,099)
	7	0,4	0	0,203	(0,123)	0,255	(0,154)																
			0,05					0,198	(0,120)	0,248	(0,150)	0,207	(0,120)	0,259	(0,151)	0,207	(0,115)	0,259	(0,144)	0,211*	(0,117)	0,264*	(0,147)
			0,1					0,194	(0,130)	0,243	(0,163)	0,218*	(0,131)	0,272*	(0,162)	0,227*	(0,123)	0,283*	(0,152)	0,234*	(0,130)	0,293*	(0,162)
			0,2					0,197	(0,124)	0,247	(0,155)	0,237*	(0,128)	0,297*	(0,160)	0,268*	(0,129)	0,335*	(0,160)	0,285*	(0,132)	0,356*	(0,165)
		0,7	0	0,194	(0,105)	0,243	(0,133)																
			0,05					0,208	(0,106)	0,26	(0,131)	0,205	(0,109)	0,257	(0,136)	0,219*	(0,106)	0,274*	(0,131)	0,208	(0,106)	0,261	(0,132)
			0,1					0,199	(0,110)	0,249	(0,138)	0,211*	(0,110)	0,263*	(0,137)	0,228*	(0,114)	0,285*	(0,144)	0,230*	(0,108)	0,288*	(0,136)
			0,2					0,195	(0,116)	0,244	(0,144)	0,234*	(0,116)	0,292*	(0,144)	0,244*	(0,117)	0,305*	(0,146)	0,268*	(0,113)	0,337*	(0,141)
		0,9	0	0,202	(0,096)	0,254	(0,120)																
			0,05					0,200	(0,101)	0,251	(0,126)	0,215*	(0,090)	0,269*	(0,112)	0,209*	(0,099)	0,262*	(0,124)	0,209*	(0,095)	0,261*	(0,118)
			0,1					0,207	(0,100)	0,258	(0,125)	0,208	(0,102)	0,259	(0,127)	0,218*	(0,102)	0,273*	(0,127)	0,220*	(0,093)	0,275*	(0,115)
			0,2					0,201	(0,103)	0,251	(0,127)	0,222*	(0,102)	0,276*	(0,127	0,235*	(0,105)	0,293*	(0,131)	0,248*	(0,100)	0,309*	(0,125)

* indique que le test de Student est significatif à 5% - LRM : $H_0 : \mu_{\hat{d}_{12}} = d_\theta$ - SM : $H_0 : \mu_{\hat{d}_{12}} = d_S$ ($d_S = 0.15$ pour J=4, $d_S = 0.25$ pour J=7)

Annexe D

Comparaison de méthodes d'analyse des effets temps et groupe pour données subjectives longitudinales sujettes au dropout : Résultats complets

TABLE D.1 – Risque de première espèce de l'effet groupe pour les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J), de la corrélation de la variable latente (ρ_θ), de la proportion de dropout $\pi^{(t)}$ et du type de dropout (ρ_{θ_χ}). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour la méthode SM et une structure UN pour la méthode LRM. Effet temps simulé : $d_\theta = 0$.

				no dropout		MCAR		MNAR					
								$\rho_{\theta\chi} = -0.4$		$\rho_{\theta\chi} = -0.7$		$\rho_{\theta\chi} = -0.9$	
N	J	ρ_θ	$\pi^{(t)}$	LRM	SM	LRM	SM	LRM	SM	LRM	SM	LRM	SM
100	4	0,4	0	0,050	0,040								
			0,05			0,058	0,062	0,058	0,056	0,062	0,060	0,038	0,038
			0,1			0,062	0,060	0,048	0,050	0,064	0,060	0,042	0,046
			0,2			0,066	0,068	0,070	0,060	0,054	0,048	0,050	0,052
		0,7	0	0,054	0,060								
			0,05			0,052	0,060	0,050	0,052	0,054	0,058	0,074!	0,078!
			0,1			0,062	0,056	0,052	0,052	0,036	0,042	0,050	0,050
			0,2			0,048	0,060	0,066	0,072!	0,046	0,048	0,068	0,068
		0,9	0	0,058	0,064								
			0,05			0,040	0,054	0,064	0,082!	0,066	0,066	0,067	0,070
			0,1			0,056	0,062	0,056	0,066	0,046	0,044	0,074!	0,080!
			0,2			0,070	0,072!	0,072!	0,072!	0,042	0,042	0,034	0,044
	7	0,4	0	0,068	0,064								
			0,05			0,054	0,052	0,064	0,062	0,060	0,064	0,060	0,052
			0,1			0,050	0,042	0,062	0,064	0,080!	0,066	0,056	0,058
			0,2			0,066	0,052	0,064	0,060	0,042	0,050	0,072!	0,068
		0,7	0	0,056	0,056								
			0,05			0,054	0,058	0,064	0,054	0,042	0,050	0,032	0,034
			0,1			0,046	0,054	0,048	0,046	0,058	0,070	0,078!	0,070
			0,2			0,050	0,044	0,054	0,052	0,074!	0,070	0,060	0,056
		0,9	0	0,060	0,054								
			0,05			0,065	0,072!	0,072!	0,080!	0,064	0,060	0,074!	0,054
			0,1			0,079!	0,082!	0,082!	0,090!	0,076!	0,066	0,066	0,072!
			0,2			0,076!	0,054	0,070	0,052	0,050	0,046	0,074!	0,064
200	4	0,4	0	0,048	0,046								
			0,05			0,052	0,050	0,064	0,066	0,050	0,050	0,044	0,050
			0,1			0,074!	0,072!	0,040	0,040	0,054	0,050	0,052	0,054
			0,2			0,048	0,046	0,044	0,050	0,062	0,068	0,040	0,042
		0,7	0	0,042	0,052								
			0,05			0,060	0,058	0,046	0,054	0,068	0,074!	0,042	0,044
			0,1			0,056	0,066	0,066	0,070	0,050	0,062	0,062	0,064
			0,2			0,056	0,064	0,046	0,044	0,052	0,044	0,058	0,066
		0,9	0	0,063	0,062								
			0,05			0,038	0,058	0,058	0,070	0,056	0,062	0,060	0,068
			0,1			0,053	0,046	0,038	0,050	0,063	0,070	0,036	0,046
			0,2			0,054	0,064	0,055	0,066	0,040	0,056	0,062	0,076!
	7	0,4	0	0,054	0,048								
			0,05			0,048	0,044	0,056	0,060	0,042	0,042	0,048	0,050
			0,1			0,044	0,050	0,050	0,050	0,058	0,054	0,058	0,064
			0,2			0,068	0,068	0,044	0,042	0,062	0,050	0,032	0,042
		0,7	0	0,058	0,058								
			0,05			0,062	0,064	0,076!	0,084!	0,050	0,050	0,070	0,072!
			0,1			0,056	0,064	0,044	0,040	0,058	0,052	0,058	0,064
			0,2			0,046	0,056	0,058	0,064	0,052	0,052	0,058	0,062
		0,9	0	0,044	0,054								
			0,05			0,060	0,060	0,036	0,036	0,045	0,060	0,054	0,060
			0,1			0,057	0,068	0,054	0,064	0,058	0,052	0,058	0,058
			0,2			0,053	0,044	0,077!	0,076!	0,054	0,048	0,060	0,052

! l'intervalle de confiance à 95% ne contient pas la valeur attendue 5%.

TABLE D.2 – Risque de première espèce de l'effet groupe pour les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J), de la corrélation de la variable latente (ρ_θ), de la proportion de dropout $\pi^{(t)}$ et du type de dropout (ρ_{θ_χ}). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour la méthode SM et une structure UN pour la méthode LRM. Effet temps simulé : $d_\theta = 0.2$.

				no dropout		MCAR		MNAR					
								$\rho_{\theta_\chi} = -0.4$		$\rho_{\theta_\chi} = -0.7$		$\rho_{\theta_\chi} = -0.9$	
N	J	ρ_θ	$\pi^{(t)}$	LRM	SM	LRM	SM	LRM	SM	LRM	SM	LRM	SM
100	4	0,4	0	0,060	0,058								
			0,05			0,056	0,060	0,062	0,060	0,064	0,056	0,058	0,048
			0,1			0,048	0,060	0,074!	0,078!	0,046	0,048	0,064	0,056
			0,2			0,074!	0,056	0,060	0,062	0,050	0,050	0,060	0,052
		0,7	0	0,074!	0,066								
			0,05			0,054	0,048	0,050	0,046	0,058	0,066	0,058	0,062
			0,1			0,056	0,052	0,066	0,062	0,056	0,062	0,054	0,060
			0,2			0,040	0,044	0,058	0,062	0,060	0,050	0,048	0,050
		0,9	0	0,053	0,054								
			0,05			0,052	0,062	0,088!	0,088!	0,060	0,068	0,070	0,072!
			0,1			0,072!	0,076!	0,065	0,062	0,058	0,068	0,102!	0,098!
			0,2			0,062	0,060	0,048	0,056	0,056	0,066	0,066	0,070
	7	0,4	0	0,048	0,046								
			0,05			0,042	0,042	0,056	0,052	0,054	0,048	0,070	0,068
			0,1			0,076!	0,070	0,062	0,056	0,056	0,060	0,068	0,068
			0,2			0,066	0,056	0,048	0,042	0,064	0,056	0,062	0,068
		0,7	0	0,080!	0,084!								
			0,05			0,042	0,032	0,064	0,060	0,060	0,064	0,052	0,052
			0,1			0,044	0,048	0,080!	0,066	0,046	0,054	0,052	0,050
			0,2			0,060	0,054	0,054	0,056	0,056	0,054	0,054	0,054
		0,9	0	0,070	0,062								
			0,05			0,068	0,068	0,064	0,070	0,060	0,060	0,048	0,038
			0,1			0,075!	0,066	0,054	0,066	0,045	0,044	0,076!	0,082!
			0,2			0,066	0,056	0,062	0,052	0,058	0,054	0,062	0,070
200	4	0,4	0	0,052	0,048								
			0,05			0,042	0,044	0,056	0,058	0,056	0,052	0,026!	0,034
			0,1			0,044	0,044	0,042	0,042	0,040	0,036	0,068	0,072!
			0,2			0,058	0,056	0,064	0,052	0,058	0,060	0,054	0,062
		0,7	0	0,048	0,054								
			0,05			0,052	0,052	0,060	0,066	0,048	0,054	0,036	0,054
			0,1			0,040	0,050	0,048	0,054	0,056	0,050	0,044	0,062
			0,2			0,066	0,058	0,036	0,044	0,050	0,052	0,070	0,060
		0,9	0	0,057	0,064								
			0,05			0,056	0,074!	0,060	0,078!	0,056	0,066	0,066	0,058
			0,1			0,054	0,062	0,058	0,070	0,048	0,058	0,058	0,066
			0,2			0,043	0,070	0,060	0,074!	0,046	0,064	0,054	0,064
	7	0,4	0	0,056	0,056								
			0,05			0,060	0,050	0,062	0,066	0,046	0,054	0,048	0,052
			0,1			0,054	0,046	0,058	0,056	0,048	0,044	0,048	0,058
			0,2			0,036	0,028!	0,044	0,040	0,046	0,050	0,046	0,052
		0,7	0	0,070	0,068								
			0,05			0,058	0,064	0,060	0,066	0,054	0,062	0,050	0,062
			0,1			0,048	0,050	0,040	0,048	0,048	0,054	0,042	0,036
			0,2			0,048	0,050	0,050	0,052	0,054	0,060	0,062	0,058
		0,9	0	0,048	0,058								
			0,05			0,059	0,070	0,051	0,058	0,046	0,056	0,040	0,042
			0,1			0,079!	0,078!	0,056	0,060	0,060	0,066	0,064	0,064
			0,2			0,067	0,056	0,054	0,062	0,075!	0,078!	0,046	0,058

! l'intervalle de confiance à 95% ne contient pas la valeur attendue 5%.

TABLE D.3 – Puissance de l'effet groupe pour les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J), de la corrélation de la variable latente (ρ_θ), de la proportion de dropout $\pi^{(t)}$ et du type de dropout (ρ_{θ_χ}). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour la méthode SM et une structure UN pour la méthode LRM. Effet temps simulé : $d_\theta = 0$.

				no dropout		MCAR		MNAR					
								$\rho_{\theta\chi} = -0.4$		$\rho_{\theta\chi} = -0.7$		$\rho_{\theta\chi} = -0.9$	
N	J	ρ_θ	$\pi^{(t)}$	LRM	SM	LRM	SM	LRM	SM	LRM	SM	LRM	SM
100	4	0,4	0	0,729	0,734								
			0,05			0,653	0,656	0,693	0,702	0,676	0,678	0,685	0,694
			0,1			0,682	0,682	0,698	0,706	0,662	0,658	0,681	0,684
			0,2			0,636	0,626	0,609	0,622	0,654	0,650	0,606	0,604
		0,7	0	0,643	0,668								
			0,05			0,621	0,640	0,602	0,614	0,604	0,600	0,613	0,650
			0,1			0,618	0,654	0,634	0,652	0,566	0,586	0,574	0,590
			0,2			0,516	0,530	0,572	0,586	0,525	0,544	0,598	0,610
		0,9	0	0,579	0,598								
			0,05			0,525	0,544	0,546	0,576	0,500	0,550	0,548	0,592
			0,1			0,553	0,576	0,566	0,584	0,514	0,546	0,541	0,580
			0,2			0,540	0,562	0,504	0,524	0,545	0,576	0,499	0,532
	7	0,4	0	0,780	0,782								
			0,05			0,780	0,790	0,784	0,798	0,796	0,810	0,756	0,758
			0,1			0,730	0,750	0,758	0,744	0,748	0,744	0,724	0,718
			0,2			0,750	0,746	0,731	0,720	0,714	0,708	0,708	0,702
		0,7	0	0,683	0,688								
			0,05			0,670	0,678	0,649	0,662	0,659	0,694	0,699	0,716
			0,1			0,696	0,716	0,680	0,698	0,650	0,658	0,661	0,652
			0,2			0,616	0,640	0,632	0,660	0,630	0,668	0,697	0,706
		0,9	0	0,623	0,658								
			0,05			0,610	0,638	0,621	0,648	0,616	0,624	0,661	0,666
			0,1			0,643	0,664	0,634	0,642	0,632	0,656	0,563	0,570
			0,2			0,614	0,622	0,551	0,576	0,606	0,626	0,599	0,634
200	4	0,4	0	0,960	0,972								
			0,05			0,940	0,944	0,928	0,936	0,936	0,946	0,936	0,942
			0,1			0,916	0,922	0,934	0,938	0,920	0,912	0,924	0,924
			0,2			0,904	0,912	0,872	0,882	0,906	0,906	0,908	0,910
		0,7	0	0,910	0,912								
			0,05			0,900	0,900	0,881	0,888	0,864	0,876	0,895	0,904
			0,1			0,882	0,888	0,884	0,894	0,892	0,904	0,884	0,896
			0,2			0,838	0,842	0,846	0,858	0,864	0,876	0,845	0,862
		0,9	0	0,887	0,898								
			0,05			0,831	0,854	0,835	0,852	0,865	0,878	0,847	0,874
			0,1			0,819	0,846	0,872	0,880	0,842	0,860	0,834	0,852
			0,2			0,842	0,858	0,777	0,802	0,810	0,820	0,806	0,824
	7	0,4	0	0,968	0,976								
			0,05			0,962	0,964	0,982	0,986	0,966	0,962	0,970	0,970
			0,1			0,962	0,964	0,970	0,970	0,974	0,978	0,948	0,952
			0,2			0,960	0,960	0,929	0,938	0,939	0,948	0,956	0,948
		0,7	0	0,928	0,944								
			0,05			0,912	0,922	0,940	0,944	0,938	0,950	0,940	0,946
			0,1			0,914	0,920	0,926	0,946	0,932	0,934	0,914	0,912
			0,2			0,920	0,928	0,916	0,922	0,902	0,912	0,904	0,922
		0,9	0	0,897	0,912								
			0,05			0,868	0,902	0,879	0,886	0,895	0,912	0,896	0,912
			0,1			0,867	0,890	0,883	0,904	0,846	0,858	0,881	0,894
			0,2			0,869	0,880	0,859	0,866	0,868	0,872	0,867	0,894

en gris : l'intervalle de confiance à 95% du risque de première espèce ne contient pas la valeur attendue 5%

TABLE D.4 – Puissance de l'effet groupe pour les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J), de la corrélation de la variable latente (ρ_θ), de la proportion de dropout $\pi^{(t)}$ et du type de dropout (ρ_{θ_χ}). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour la méthode SM et une structure UN pour la méthode LRM. Effet temps simulé : $d_\theta = 0,2$.

				no dropout		MCAR		MNAR					
								$\rho_{\theta\chi} = -0.4$		$\rho_{\theta\chi} = -0.7$		$\rho_{\theta\chi} = -0.9$	
N	J	ρ_θ	$\pi^{(t)}$	LRM	SM	LRM	SM	LRM	SM	LRM	SM	LRM	SM
100	4	0,4	0	0,718	0,722								
			0,05			0,690	0,684	0,670	0,668	0,683	0,704	0,680	0,680
			0,1			0,684	0,688	0,646	0,666	0,683	0,706	0,684	0,684
			0,2			0,624	0,634	0,633	0,636	0,632	0,634	0,624	0,642
		0,7	0	0,625	0,652								
			0,05			0,567	0,576	0,569	0,578	0,648	0,660	0,615	0,630
			0,1			0,639	0,654	0,588	0,616	0,588	0,608	0,612	0,630
			0,2			0,564	0,584	0,595	0,594	0,568	0,586	0,578	0,582
		0,9	0	0,582	0,610								
			0,05			0,592	0,620	0,565	0,590	0,548	0,582	0,558	0,594
			0,1			0,561	0,582	0,540	0,572	0,536	0,564	0,537	0,568
			0,2			0,525	0,534	0,554	0,576	0,492	0,516	0,570	0,584
	7	0,4	0	0,784	0,792								
			0,05			0,752	0,762	0,778	0,788	0,772	0,772	0,778	0,784
			0,1			0,788	0,792	0,730	0,748	0,758	0,756	0,772	0,770
			0,2			0,722	0,718	0,716	0,718	0,758	0,770	0,754	0,742
		0,7	0	0,736	0,758								
			0,05			0,708	0,718	0,672	0,686	0,669	0,710	0,698	0,728
			0,1			0,637	0,670	0,659	0,674	0,688	0,704	0,668	0,678
			0,2			0,648	0,668	0,620	0,632	0,638	0,666	0,658	0,652
		0,9	0	0,610	0,628								
			0,05			0,616	0,626	0,620	0,636	0,614	0,618	0,614	0,650
			0,1			0,589	0,636	0,626	0,632	0,630	0,648	0,626	0,626
			0,2			0,600	0,608	0,582	0,592	0,579	0,610	0,640	0,638
200	4	0,4	0	0,908	0,916								
			0,05			0,936	0,932	0,926	0,934	0,942	0,938	0,916	0,926
			0,1			0,926	0,938	0,904	0,914	0,930	0,930	0,928	0,928
			0,2			0,916	0,918	0,911	0,920	0,901	0,896	0,888	0,896
		0,7	0	0,918	0,924								
			0,05			0,843	0,850	0,882	0,898	0,896	0,900	0,876	0,890
			0,1			0,878	0,886	0,863	0,882	0,896	0,902	0,868	0,874
			0,2			0,866	0,872	0,872	0,878	0,871	0,874	0,844	0,864
		0,9	0	0,851	0,876								
			0,05			0,853	0,862	0,834	0,854	0,851	0,858	0,876	0,882
			0,1			0,846	0,868	0,841	0,848	0,834	0,864	0,816	0,858
			0,2			0,815	0,816	0,809	0,840	0,817	0,844	0,810	0,818
	7	0,4	0	0,974	0,980								
			0,05			0,972	0,972	0,960	0,962	0,960	0,964	0,948	0,956
			0,1			0,966	0,964	0,976	0,974	0,962	0,960	0,964	0,962
			0,2			0,952	0,952	0,935	0,944	0,962	0,962	0,944	0,946
		0,7	0	0,942	0,954								
			0,05			0,920	0,928	0,952	0,950	0,936	0,944	0,926	0,930
			0,1			0,918	0,930	0,918	0,936	0,940	0,946	0,928	0,942
			0,2			0,902	0,902	0,903	0,904	0,919	0,920	0,912	0,918
		0,9	0	0,911	0,926								
			0,05			0,889	0,892	0,903	0,912	0,895	0,910	0,903	0,920
			0,1			0,881	0,904	0,897	0,902	0,888	0,912	0,887	0,912
			0,2			0,844	0,880	0,860	0,892	0,874	0,878	0,861	0,880

en gris : l'intervalle de confiance à 95% du risque de première espèce ne contient pas la valeur attendue 5%

TABLE D.5 – Estimations de l'effet groupe ($\hat{\beta}_{gp}$) et de son erreur standard (s.e.) pour les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J), de la corrélation de la variable latente (ρ_θ), de la proportion ($\pi^{(t)}$) et du type de dropout (ρ_{θ_χ}). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour la méthode SM et une structure UN pour la méthode LRM. Effet temps simulé : $d_\theta = 0$. Effet groupe simulé : $\Delta_\theta = 0$. $\Delta_S = 0$.

				no dropout				MCAR				MNAR											
												$\rho_{\theta\chi} = -0.4$				$\rho_{\theta\chi} = -0.7$				$\rho_{\theta\chi} = -0.9$			
				LRM		SM		LRM		SM		LRM		SM		LRM		SM		LRM		SM	
N	J	ρ_θ	$\pi^{(t)}$	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)
100	4	0,4	0	-0,021*	(0,200)	-0,015*	(0,149)																
			0,05					0,012	(0,205)	0,007	(0,153)	-0,014	(0,204)	-0,010	(0,153)	0,006	(0,204)	0,005	(0,152)	0,013	(0,204)	0,011	(0,153)
			0,1					0,005	(0,207)	0,005	(0,155)	0,003	(0,209)	0,002	(0,156)	-0,004	(0,207)	-0,003	(0,155)	0,008	(0,207)	0,003	(0,156)
			0,2					-0,015	(0,220)	-0,010	(0,164)	-0,014	(0,216)	-0,011	(0,162)	0,000	(0,218)	0,001	(0,163)	0,019	(0,218)	0,013	(0,163)
		0,7	0	0,010	(0,220)	0,007	(0,161)																
			0,05					-0,009	(0,224)	-0,007	(0,164)	0,021*	(0,223)	0,016*	(0,164)	0,000	(0,222)	0,000	(0,163)	0,003	(0,224)	0,001	(0,164)
			0,1					0,004	(0,228)	0,005	(0,167)	0,015	(0,229)	0,010	(0,168)	0,002	(0,227)	0,004	(0,167)	-0,008	(0,229)	-0,005	(0,168)
			0,2					-0,020	(0,237)	-0,014	(0,174)	0,024*	(0,238)	0,015	(0,174)	0,007	(0,237)	0,002	(0,174)	-0,014	(0,237)	-0,008	(0,174)
		0,9	0	0,008	(0,237)	0,006	(0,169)																
			0,05					0,006	(0,239)	0,004	(0,172)	0,003	(0,239)	0,001	(0,172)	0,015	(0,240)	0,014	(0,172)	0,010	(0,239)	0,008	(0,172)
			0,1					-0,022	(0,243)	-0,014	(0,174)	-0,008	(0,243)	-0,008	(0,175)	0,009	(0,244)	0,005	(0,175)	0,021	(0,243)	0,017	(0,175)
			0,2					-0,020	(0,251)	-0,010	(0,181)	-0,004	(0,249)	-0,004	(0,180)	-0,017	(0,249)	-0,012	(0,180)	-0,007	(0,252)	-0,008	(0,181)
	7	0,4	0	-0,010	(0,181)	-0,010	(0,225)																
			0,05					0,010	(0,184)	0,012	(0,230)	0,001	(0,184)	0,002	(0,230)	0,025*	(0,184)	0,033*	(0,230)	-0,002	(0,183)	0,001	(0,230)
			0,1					0,001	(0,187)	-0,003	(0,235)	0,000	(0,187)	0,000	(0,235)	-0,003	(0,188)	-0,002	(0,234)	0,017	(0,188)	0,025*	(0,234)
			0,2					0,013	(0,195)	0,015	(0,244)	0,022*	(0,195)	0,031*	(0,245)	-0,009	(0,196)	-0,011	(0,245)	0,001	(0,194)	-0,002	(0,244)
		0,7	0	-0,002	(0,200)	-0,004	(0,246)																
			0,05					0,000	(0,204)	-0,002	(0,251)	0,008	(0,206)	0,008	(0,253)	-0,010	(0,205)	-0,012	(0,252)	0,011	(0,207)	0,013	(0,253)
			0,1					-0,003	(0,208)	-0,001	(0,257)	-0,006	(0,209)	-0,004	(0,257)	0,018	(0,209)	0,028*	(0,258)	0,007	(0,209)	0,005	(0,257)
			0,2					0,002	(0,215)	0,000	(0,265)	-0,004	(0,215)	-0,005	(0,266)	-0,003	(0,214)	-0,004	(0,264)	-0,026*	(0,214)	-0,028*	(0,265)
		0,9	0	-0,018	(0,218)	-0,021	(0,263)																
			0,05					0,007	(0,220)	0,012	(0,266)	0,000	(0,220)	-0,005	(0,267)	0,010	(0,221)	0,012	(0,267)	0,003	(0,219)	-0,001	(0,267)
			0,1					-0,007	(0,222)	-0,005	(0,271)	-0,010	(0,222)	-0,011	(0,270)	0,012	(0,222)	0,016	(0,270)	-0,012	(0,222)	-0,014	(0,270)
			0,2					-0,017	(0,228)	-0,009	(0,278)	-0,001	(0,228)	-0,002	(0,279)	-0,018	(0,229)	-0,026*	(0,278)	-0,024*	(0,228)	-0,028*	(0,278)

Suite page suivante...

TABLE D.5 – Suite

					no dropout		MCAR				MNAR							
									$\rho_{\theta\chi} = -0.4$				$\rho_{\theta\chi} = -0.7$				$\rho_{\theta\chi} = -0.9$	
					LRM	SM	LRM	SM	LRM	SM	LRM	SM	LRM	SM	LRM	SM
200	4	0,4	0		0,002 (0,141)	0,001 (0,105)												
				0,05			0,003 (0,144)	0,002 (0,107)	0,009 (0,143)	0,007 (0,107)	0,004 (0,144)	0,003 (0,107)	-0,003 (0,143)	-0,002 (0,107)				
				0,1			-0,003 (0,148)	-0,002 (0,110)	-0,008 (0,147)	-0,006 (0,110)	0,008 (0,147)	0,005 (0,110)	0,006 (0,148)	0,004 (0,110)				
				0,2			0,007 (0,153)	0,005 (0,115)	0,006 (0,154)	0,004 (0,115)	0,004 (0,154)	0,003 (0,115)	0,013 (0,153)	0,009 (0,115)				
		0,7	0		0,001 (0,155)	0,000 (0,113)												
				0,05			0,013 (0,158)	0,008 (0,115)	0,008 (0,157)	0,005 (0,115)	0,009 (0,158)	0,006 (0,116)	-0,002 (0,157)	-0,003 (0,115)				
				0,1			-0,004 (0,160)	-0,001 (0,118)	-0,010 (0,161)	-0,007 (0,118)	0,013 (0,161)	0,008 (0,118)	0,010 (0,160)	0,008 (0,118)				
				0,2			-0,011 (0,166)	-0,009 (0,122)	-0,003 (0,166)	-0,001 (0,123)	0,000 (0,165)	0,000 (0,122)	-0,006 (0,166)	-0,003 (0,122)				
		0,9	0		0,007 (0,166)	0,005 (0,119)												
				0,05			0,001 (0,169)	0,000 (0,121)	0,007 (0,168)	0,005 (0,121)	-0,015 (0,169)	-0,011 (0,122)	-0,004 (0,169)	-0,004 (0,121)				
				0,1			0,003 (0,171)	0,002 (0,123)	0,003 (0,171)	0,002 (0,123)	-0,002 (0,171)	0,001 (0,123)	0,000 (0,171)	0,001 (0,123)				
				0,2			0,002 (0,176)	0,002 (0,128)	-0,013 (0,176)	-0,009 (0,128)	0,005 (0,176)	0,004 (0,127)	0,022* (0,175)	0,016* (0,127)				
	7	0,4	0		-0,005 (0,127)	-0,007 (0,158)												
				0,05			-0,004 (0,130)	-0,005 (0,162)	-0,002 (0,130)	-0,003 (0,162)	-0,005 (0,130)	-0,004 (0,162)	-0,007 (0,130)	-0,010 (0,162)				
				0,1			-0,001 (0,133)	-0,001 (0,166)	-0,010 (0,132)	-0,013 (0,166)	0,006 (0,133)	0,007 (0,166)	-0,007 (0,133)	-0,009 (0,166)				
				0,2			-0,014* (0,138)	-0,018* (0,173)	-0,006 (0,139)	-0,006 (0,173)	0,009 (0,138)	0,011 (0,172)	0,000 (0,138)	-0,001 (0,172)				
		0,7	0		0,008 (0,142)	0,010 (0,174)												
				0,05			0,004 (0,145)	0,005 (0,178)	0,006 (0,144)	0,008 (0,177)	-0,004 (0,145)	-0,004 (0,178)	-0,004 (0,145)	-0,005 (0,177)				
				0,1			0,008 (0,147)	0,010 (0,180)	-0,001 (0,148)	0,000 (0,182)	-0,010 (0,147)	-0,011 (0,181)	-0,005 (0,147)	-0,007 (0,180)				
				0,2			0,008 (0,152)	0,010 (0,188)	-0,004 (0,152)	-0,004 (0,188)	-0,004 (0,152)	-0,006 (0,187)	0,010 (0,151)	0,012 (0,186)				
		0,9	0		0,010 (0,153)	0,013 (0,185)												
				0,05			0,004 (0,155)	0,008 (0,188)	-0,009 (0,156)	-0,010 (0,189)	-0,011 (0,156)	-0,006 (0,189)	-0,005 (0,155)	-0,009 (0,188)				
				0,1			-0,019* (0,157)	-0,018* (0,191)	0,001 (0,157)	0,001 (0,191)	-0,001 (0,157)	0,005 (0,191)	0,002 (0,158)	0,003 (0,191)				
				0,2			-0,004 (0,162)	-0,001 (0,197)	0,010 (0,161)	0,011 (0,197)	-0,017* (0,161)	-0,018* (0,197)	0,002 (0,161)	0,007 (0,196)				

* test significatif à 5% - $H_0 : \mu_{\hat{\beta}_{gp}} = 0$

TABLE D.6 – Estimations de l'effet groupe ($\hat{\beta}_{gp}$) et de son erreur standard (s.e.) pour les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J), de la corrélation de la variable latente (ρ_θ), de la proportion ($\pi^{(t)}$) et du type de dropout (ρ_{θ_χ}). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour la méthode SM et une structure UN pour la méthode LRM. Effet temps simulé : $d_\theta = 0, 2$. Effet groupe simulé : $\Delta_\theta = 0$. $\Delta_S = 0$

				no dropout				MCAR				MNAR									
												$\rho_{\theta\chi} = -0.4$				$\rho_{\theta\chi} = -0.7$				$\rho_{\theta\chi} = -0.9$			
				LRM		SM		LRM		SM		LRM		SM		LRM		SM		LRM		SM	
N	J	ρ_θ	$\pi^{(t)}$	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)
100	4	0,4	0	0,003	(0,200)	0,002	(0,149)																
			0,05					-0,013	(0,202)	-0,010	(0,151)	0,006	(0,206)	0,003	(0,153)	-0,011	(0,204)	-0,011	(0,152)	-0,005	(0,203)	-0,004	(0,152)
			0,1					0,009	(0,210)	0,008	(0,156)	-0,010	(0,208)	-0,006	(0,155)	-0,003	(0,211)	-0,002	(0,156)	-0,010	(0,207)	-0,009	(0,155)
			0,2					-0,007	(0,218)	-0,005	(0,163)	0,009	(0,219)	0,006	(0,162)	-0,001	(0,216)	-0,002	(0,161)	0,008	(0,217)	0,007	(0,162)
		0,7	0	0,011	(0,219)	0,009	(0,160)																
			0,05					0,008	(0,224)	0,009	(0,163)	0,004	(0,223)	0,003	(0,163)	0,005	(0,224)	0,003	(0,163)	0,020*	(0,223)	0,015*	(0,163)
			0,1					0,023*	(0,229)	0,017*	(0,167)	0,004	(0,228)	0,002	(0,166)	-0,008	(0,229)	-0,007	(0,167)	-0,001	(0,228)	0,000	(0,166)
			0,2					0,013	(0,234)	0,009	(0,172)	-0,009	(0,235)	-0,005	(0,173)	-0,001	(0,237)	-0,002	(0,173)	0,011	(0,237)	0,005	(0,172)
		0,9	0	0,016	(0,235)	0,012	(0,168)																
			0,05					0,014	(0,241)	0,008	(0,172)	0,001	(0,240)	0,004	(0,171)	-0,011	(0,241)	-0,008	(0,172)	-0,009	(0,240)	-0,008	(0,171)
			0,1					-0,014	(0,244)	-0,011	(0,174)	-0,001	(0,244)	0,001	(0,175)	0,016	(0,243)	0,007	(0,174)	-0,004	(0,244)	-0,002	(0,174)
			0,2					0,007	(0,253)	0,004	(0,181)	0,006	(0,251)	0,004	(0,180)	0,010	(0,250)	0,008	(0,179)	0,008	(0,250)	0,005	(0,179)
	7	0,4	0	0,002	(0,179)	0,001	(0,224)																
			0,05					-0,010	(0,185)	-0,013	(0,230)	-0,003	(0,184)	-0,002	(0,229)	0,005	(0,183)	0,008	(0,229)	0,010	(0,185)	0,013	(0,230)
			0,1					0,013	(0,188)	0,014	(0,235)	0,007	(0,188)	0,006	(0,234)	-0,008	(0,186)	-0,007	(0,232)	-0,003	(0,188)	-0,003	(0,234)
			0,2					-0,010	(0,196)	-0,017	(0,244)	0,001	(0,195)	-0,001	(0,243)	-0,001	(0,195)	0,000	(0,243)	-0,006	(0,194)	-0,007	(0,242)
		0,7	0	-0,015	(0,202)	-0,021	(0,247)																
			0,05					0,006	(0,205)	0,008	(0,252)	-0,001	(0,205)	-0,005	(0,251)	-0,004	(0,205)	-0,008	(0,250)	-0,004	(0,205)	-0,003	(0,251)
			0,1					-0,001	(0,209)	-0,001	(0,256)	0,010	(0,206)	0,010	(0,254)	0,002	(0,207)	0,002	(0,254)	-0,009	(0,209)	-0,017	(0,256)
			0,2					-0,01	(0,216)	-0,014	(0,265)	-0,005	(0,215)	-0,009	(0,264)	-0,024*	(0,215)	-0,027*	(0,264)	0,012	(0,214)	0,014	(0,263)
		0,9	0	-0,009	(0,218)	-0,014	(0,263)																
			0,05					0,016	(0,218)	0,021	(0,265)	0,014	(0,221)	0,019	(0,266)	-0,012	(0,219)	-0,013	(0,266)	-0,012	(0,219)	-0,014	(0,265)
			0,1					-0,015	(0,225)	-0,007	(0,271)	0,004	(0,222)	0,011	(0,269)	-0,003	(0,221)	-0,003	(0,269)	0,005	(0,223)	0,009	(0,270)
			0,2					0,002	(0,230)	0,003	(0,280)	0,004	(0,228)	0,004	(0,278)	0,016	(0,229)	0,015	(0,278)	-0,017	(0,225)	-0,021	(0,274)

Suite page suivante...

Table D.6 – Suite

				no dropout		MCAR		MNAR									
								$\rho_{\theta\chi} = -0.4$				$\rho_{\theta\chi} = -0.7$				$\rho_{\theta\chi} = -0.9$	
				LRM	SM	LRM	SM	LRM	SM	LRM	SM	LRM	SM	LRM	SM
200	4	0,4	0	-0,007 (0,140)	-0,005 (0,105)												
			0,05			-0,004 (0,144)	-0,003 (0,107)	0,002 (0,144)	0,002 (0,107)	-0,010 (0,144)	-0,008 (0,107)	-0,004 (0,144)	-0,003 (0,107)				
			0,1			-0,002 (0,147)	-0,002 (0,110)	0,009 (0,147)	0,006 (0,110)	0,008 (0,147)	0,007 (0,109)	0,005 (0,147)	0,005 (0,109)				
			0,2			0,002 (0,154)	0,002 (0,115)	0,018* (0,154)	0,014* (0,114)	-0,006 (0,153)	-0,004 (0,114)	0,001 (0,153)	0,000 (0,114)				
		0,7	0	-0,005 (0,156)	-0,005 (0,113)												
			0,05			-0,003 (0,158)	-0,002 (0,115)	0,004 (0,158)	0,005 (0,115)	-0,012 (0,158)	-0,008 (0,115)	-0,002 (0,157)	-0,002 (0,115)				
			0,1			0,008 (0,161)	0,005 (0,118)	-0,001 (0,161)	-0,002 (0,118)	-0,002 (0,160)	-0,002 (0,117)	0,002 (0,160)	0,002 (0,117)				
			0,2			-0,009 (0,167)	-0,006 (0,122)	-0,007 (0,166)	-0,006 (0,122)	0,001 (0,167)	0,000 (0,122)	0,007 (0,166)	0,006 (0,121)				
		0,9	0	-0,001 (0,165)	-0,001 (0,118)												
			0,05			0,012 (0,168)	0,009 (0,121)	-0,004 (0,168)	-0,004 (0,121)	-0,003 (0,168)	-0,003 (0,121)	0,016* (0,168)	0,012* (0,121)				
			0,1			0,011 (0,171)	0,007 (0,123)	0,009 (0,171)	0,005 (0,123)	0,020* (0,172)	0,015* (0,123)	-0,009 (0,171)	-0,006 (0,122)				
			0,2			-0,004 (0,176)	-0,002 (0,127)	-0,002 (0,176)	-0,002 (0,127)	0,005 (0,175)	0,003 (0,127)	0,010 (0,175)	0,008 (0,126)				
	7	0,4	0	0,002 (0,127)	0,003 (0,158)												
			0,05			-0,001 (0,130)	-0,002 (0,162)	-0,012* (0,130)	-0,015* (0,162)	-0,001 (0,130)	-0,001 (0,161)	0,005 (0,129)	0,007 (0,161)				
			0,1			0,004 (0,132)	0,005 (0,165)	-0,001 (0,132)	-0,001 (0,165)	0,005 (0,133)	0,006 (0,165)	0,001 (0,132)	0,002 (0,165)				
			0,2			0,003 (0,138)	0,004 (0,172)	0,008 (0,138)	0,010 (0,172)	-0,001 (0,139)	0,002 (0,172)	-0,003 (0,138)	-0,004 (0,172)				
		0,7	0	0,011 (0,142)	0,014 (0,174)												
			0,05			0,001 (0,145)	0,003 (0,177)	0,014* (0,145)	0,014 (0,177)	0,006 (0,145)	0,010 (0,177)	-0,001 (0,144)	-0,002 (0,176)				
			0,1			-0,005 (0,146)	-0,005 (0,180)	-0,002 (0,148)	-0,003 (0,181)	0,004 (0,147)	0,004 (0,180)	-0,001 (0,147)	-0,003 (0,179)				
			0,2			0,002 (0,152)	0,003 (0,187)	0,002 (0,152)	0,002 (0,187)	0,006 (0,151)	0,008 (0,186)	0,000 (0,151)	0,000 (0,185)				
		0,9	0	0,011 (0,153)	0,012 (0,185)												
			0,05			-0,001 (0,156)	-0,004 (0,188)	0,000 (0,155)	0,000 (0,188)	0,007 (0,155)	0,005 (0,187)	0,006 (0,155)	0,008 (0,187)				
			0,1			0,015* (0,157)	0,017 (0,191)	-0,012 (0,157)	-0,016 (0,190)	-0,005 (0,157)	-0,004 (0,190)	0,002 (0,157)	0,002 (0,190)				
			0,2			0,010 (0,162)	0,011 (0,197)	0,002 (0,162)	0,002 (0,197)	-0,003 (0,161)	-0,006 (0,196)	-0,010 (0,160)	-0,015 (0,195)				

* test significatif à 5% - $H_0 : \mu_{\hat{\beta}_{gp}} = 0$

TABLE D.7 – Estimations de l'effet groupe ($\hat{\beta}_{gp}$) et de son erreur standard (s.e.) pour les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J), de la corrélation de la variable latente (ρ_θ), de la proportion ($\pi^{(t)}$) et du type de dropout (ρ_{θ_χ}). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour la méthode SM et une structure UN pour la méthode LRM. Effet temps simulé : $d_\theta = 0$. Effet groupe simulé : $\Delta_\theta = 0,5$. $\Delta_S = 0.38$ pour J=4, $\Delta_S = 0.63$ pour J=7.

				no dropout				MCAR				MNAR											
												$\rho_{\theta\chi} = -0.4$				$\rho_{\theta\chi} = -0.7$				$\rho_{\theta\chi} = -0.9$			
				LRM		SM		LRM		SM		LRM		SM		LRM		SM		LRM		SM	
N	J	ρ_θ	$\pi^{(t)}$	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)
100	4	0,4	0	0,508	(0,201)	0,380	(0,149)																
			0,05					0,490	(0,205)	0,366*	(0,151)	0,515	(0,205)	0,386	(0,152)	0,499	(0,204)	0,374	(0,151)	0,491	(0,204)	0,368	(0,151)
			0,1					0,508	(0,209)	0,381	(0,155)	0,514	(0,209)	0,384	(0,155)	0,502	(0,208)	0,374	(0,155)	0,503	(0,210)	0,374	(0,155)
			0,2					0,511	(0,219)	0,381	(0,162)	0,500	(0,221)	0,373	(0,163)	0,516	(0,219)	0,385	(0,162)	0,484	(0,219)	0,361*	(0,161)
		0,7	0	0,504	(0,220)	0,380	(0,159)																
			0,05					0,503	(0,225)	0,375	(0,163)	0,500	(0,226)	0,370	(0,164)	0,491	(0,225)	0,363*	(0,163)	0,510	(0,225)	0,380	(0,163)
			0,1					0,518	(0,230)	0,387	(0,167)	0,519	(0,229)	0,387	(0,166)	0,492	(0,229)	0,368	(0,166)	0,496	(0,229)	0,370	(0,166)
			0,2					0,488	(0,237)	0,363*	(0,173)	0,514	(0,236)	0,381	(0,172)	0,493	(0,239)	0,364*	(0,173)	0,521	(0,237)	0,385	(0,173)
		0,9	0	0,506	(0,235)	0,375	(0,167)																
			0,05					0,494	(0,239)	0,365	(0,170)	0,496	(0,240)	0,369	(0,171)	0,491	(0,241)	0,363*	(0,171)	0,503	(0,242)	0,371	(0,171)
			0,1					0,510	(0,245)	0,375	(0,174)	0,515	(0,244)	0,379	(0,174)	0,511	(0,246)	0,377	(0,174)	0,513	(0,245)	0,379	(0,174)
			0,2					0,506	(0,254)	0,372	(0,180)	0,507	(0,252)	0,375	(0,180)	0,514	(0,251)	0,379	(0,180)	0,500	(0,251)	0,370	(0,180)
	7	0,4	0	0,492	(0,181)	0,618	(0,224)																
			0,05					0,497	(0,184)	0,622	(0,228)	0,501	(0,183)	0,627	(0,227)	0,502	(0,185)	0,626	(0,229)	0,500	(0,184)	0,623	(0,229)
			0,1					0,504	(0,189)	0,627	(0,235)	0,513	(0,189)	0,638	(0,234)	0,500	(0,189)	0,621	(0,234)	0,492	(0,188)	0,614	(0,234)
			0,2					0,510	(0,196)	0,635	(0,243)	0,510	(0,197)	0,633	(0,244)	0,492	(0,196)	0,614	(0,244)	0,491	(0,196)	0,610	(0,243)
		0,7	0	0,506	(0,203)	0,633	(0,247)																
			0,05					0,503	(0,206)	0,629	(0,251)	0,489	(0,205)	0,610	(0,250)	0,492	(0,205)	0,622	(0,250)	0,508	(0,206)	0,632	(0,251)
			0,1					0,513	(0,209)	0,640	(0,255)	0,503	(0,210)	0,628	(0,256)	0,499	(0,208)	0,624	(0,254)	0,501	(0,209)	0,623	(0,255)
			0,2					0,498	(0,217)	0,621	(0,265)	0,502	(0,217)	0,628	(0,264)	0,507	(0,216)	0,635	(0,264)	0,525*	(0,215)	0,649	(0,263)
		0,9	0	0,503	(0,219)	0,627	(0,261)																
			0,05					0,508	(0,220)	0,633	(0,264)	0,505	(0,221)	0,632	(0,265)	0,512	(0,221)	0,621	(0,266)	0,518	(0,219)	0,645	(0,264)
			0,1					0,514	(0,222)	0,643	(0,269)	0,506	(0,224)	0,626	(0,270)	0,509	(0,222)	0,632	(0,269)	0,487	(0,222)	0,606	(0,269)
			0,2					0,513	(0,227)	0,633	(0,276)	0,493	(0,228)	0,608	(0,276)	0,516	(0,229)	0,637	(0,277)	0,501	(0,227)	0,625	(0,276)

Suite page suivante...

TABLE D.7 – Suite

				no dropout		MCAR		MNAR							
								$\rho_{\theta\chi} = -0.4$		$\rho_{\theta\chi} = -0.7$		$\rho_{\theta\chi} = -0.9$			
				LRM	SM	LRM	SM	LRM	SM	LRM	SM	LRM	SM		
200	4	0,4	0	0,511	(0,141)	0,384	(0,104)								
			0,05					0,494	(0,144)	0,371*	(0,107)	0,510	(0,145)	0,382	(0,107)
			0,1					0,505	(0,149)	0,378	(0,110)	0,500	(0,148)	0,375	(0,109)
			0,2					0,496	(0,154)	0,373	(0,114)	0,484*	(0,154)	0,364*	(0,114)
		0,7	0	0,514	(0,156)	0,384	(0,113)								
			0,05					0,506	(0,159)	0,379	(0,115)	0,494	(0,159)	0,371	(0,115)
			0,1					0,510	(0,161)	0,381	(0,117)	0,510	(0,162)	0,380	(0,118)
			0,2					0,492	(0,167)	0,369*	(0,122)	0,502	(0,168)	0,375	(0,122)
		0,9	0	0,521*	(0,167)	0,388	(0,119)								
			0,05					0,492	(0,169)	0,369	(0,120)	0,506	(0,169)	0,374	(0,121)
			0,1					0,498	(0,171)	0,372	(0,122)	0,506	(0,171)	0,376	(0,122)
			0,2					0,506	(0,176)	0,376	(0,127)	0,484	(0,177)	0,363	(0,127)
	7	0,4	0	0,494	(0,128)	0,618	(0,158)								
			0,05					0,498	(0,131)	0,624	(0,161)	0,506	(0,131)	0,632	(0,161)
			0,1					0,492	(0,133)	0,614*	(0,165)	0,506	(0,134)	0,633	(0,165)
			0,2					0,499	(0,139)	0,624	(0,172)	0,496	(0,138)	0,619	(0,172)
		0,7	0	0,507	(0,142)	0,636	(0,173)								
			0,05					0,497	(0,145)	0,621	(0,177)	0,507	(0,146)	0,635	(0,177)
			0,1					0,495	(0,148)	0,618	(0,180)	0,510	(0,147)	0,639	(0,179)
			0,2					0,500	(0,153)	0,627	(0,186)	0,504	(0,153)	0,627	(0,186)
		0,9	0	0,507	(0,154)	0,632	(0,185)								
			0,05					0,495	(0,156)	0,618	(0,188)	0,506	(0,155)	0,628	(0,187)
			0,1					0,489	(0,157)	0,613	(0,190)	0,501	(0,157)	0,628	(0,190)
			0,2					0,504	(0,161)	0,626	(0,195)	0,507	(0,161)	0,631	(0,195)

				LRM	SM	LRM	SM
				$\rho_{\theta\chi} = -0.7$		$\rho_{\theta\chi} = -0.9$	
200	4	0,4	0	0,504	(0,145)	0,377	(0,107)
			0,05	0,484*	(0,148)	0,363*	(0,109)
			0,1	0,507	(0,155)	0,376	(0,114)
		0,7	0				
			0,05	0,500	(0,158)	0,373	(0,115)
			0,1	0,514	(0,161)	0,386	(0,117)
			0,2	0,504	(0,167)	0,378	(0,122)
		0,9	0				
			0,05	0,510	(0,169)	0,380	(0,121)
			0,1	0,505	(0,171)	0,377	(0,123)
			0,2	0,503	(0,176)	0,373	(0,127)
	7	0,4	0				
			0,05	0,499	(0,130)	0,625	(0,161)
			0,1	0,513*	(0,134)	0,640	(0,165)
			0,2	0,497	(0,139)	0,618	(0,172)
		0,7	0				
			0,05	0,509	(0,145)	0,640	(0,176)
			0,1	0,502	(0,147)	0,629	(0,180)
			0,2	0,501	(0,152)	0,628	(0,186)
		0,9	0				
			0,05	0,512	(0,155)	0,639	(0,187)
			0,1	0,484*	(0,157)	0,603*	(0,190)
			0,2	0,492	(0,161)	0,612*	(0,195)

				LRM	SM	LRM	SM
				$\rho_{\theta\chi} = -0.7$		$\rho_{\theta\chi} = -0.9$	
200	4	0,4	0	0,496	(0,145)	0,371	(0,107)
			0,05	0,501	(0,148)	0,376	(0,110)
			0,1	0,509	(0,154)	0,382	(0,114)
		0,7	0				
			0,05	0,509	(0,159)	0,382	(0,115)
			0,1	0,513	(0,162)	0,385	(0,117)
			0,2	0,501	(0,167)	0,374	(0,122)
		0,9	0				
			0,05	0,500	(0,169)	0,373	(0,120)
			0,1	0,505	(0,172)	0,377	(0,123)
			0,2	0,509	(0,176)	0,379	(0,127)
	7	0,4	0				
			0,05	0,500	(0,131)	0,625	(0,161)
			0,1	0,497	(0,133)	0,622	(0,165)
			0,2	0,507	(0,139)	0,632	(0,172)
		0,7	0				
			0,05	0,511	(0,146)	0,640	(0,177)
			0,1	0,493	(0,147)	0,614	(0,179)
			0,2	0,487	(0,152)	0,609*	(0,186)
		0,9	0				
			0,05	0,502	(0,156)	0,630	(0,187)
			0,1	0,501	(0,157)	0,628	(0,190)
			0,2	0,493	(0,161)	0,619	(0,196)

* test significatif à 5% - LRM : $H_0 : \mu_{\hat{\beta}_{gp}} = \Delta_\theta$ - SM : $H_0 : \mu_{\hat{\beta}_{gp}} = \Delta_S$ ($\Delta_S = 0.38$ pour J=4, $\Delta_S = 0.63$ pour J=7)

TABLE D.8 – Estimations de l'effet groupe ($\hat{\beta}_{gp}$) et de son erreur standard (s.e.) pour les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J), de la corrélation de la variable latente (ρ_θ), de la proportion ($\pi^{(t)}$) et du type de dropout (ρ_{θ_χ}). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour la méthode SM et une structure UN pour la méthode LRM. Effet temps simulé : $d_\theta = 0,2$. Effet groupe simulé : $\Delta_\theta = 0,5$. $\Delta_S = 0.38$ pour J=4, $\Delta_S = 0.63$ pour J=7.

				no dropout				MCAR				MNAR									
												$\rho_{\theta\chi} = -0.4$				$\rho_{\theta\chi} = -0.7$				$\rho_{\theta\chi} = -0.9$			
				LRM		SM		LRM		SM		LRM		SM		LRM		SM		LRM		SM	
N	J	ρ_θ	$\pi^{(t)}$	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)	$\hat{\beta}_{gp}$	(s.e.)
100	4	0,4	0	0,516	(0,200)	0,385	(0,148)																
			0,05					0,491	(0,206)	0,365*	(0,151)	0,492	(0,205)	0,366*	(0,151)	0,501	(0,205)	0,373	(0,151)	0,503	(0,205)	0,374	(0,151)
			0,1					0,503	(0,209)	0,374	(0,154)	0,490	(0,209)	0,365*	(0,154)	0,511	(0,209)	0,383	(0,154)	0,508	(0,209)	0,381	(0,154)
			0,2					0,507	(0,219)	0,379	(0,162)	0,512	(0,220)	0,379	(0,162)	0,505	(0,220)	0,373	(0,162)	0,498	(0,219)	0,371	(0,161)
		0,7	0	0,503	(0,221)	0,374	(0,159)																
			0,05					0,483	(0,225)	0,358*	(0,163)	0,490	(0,224)	0,361*	(0,162)	0,522*	(0,224)	0,389	(0,162)	0,508	(0,225)	0,376	(0,162)
			0,1					0,518	(0,229)	0,383	(0,166)	0,497	(0,229)	0,370	(0,166)	0,495	(0,229)	0,370	(0,166)	0,513	(0,231)	0,377	(0,166)
			0,2					0,511	(0,237)	0,376	(0,172)	0,512	(0,238)	0,382	(0,172)	0,512	(0,238)	0,380	(0,172)	0,501	(0,235)	0,370	(0,170)
		0,9	0	0,511	(0,238)	0,377	(0,168)																
			0,05					0,526*	(0,240)	0,386	(0,170)	0,494	(0,241)	0,362*	(0,170)	0,493	(0,240)	0,367	(0,170)	0,489	(0,241)	0,361*	(0,170)
			0,1					0,518	(0,245)	0,382	(0,173)	0,507	(0,246)	0,374	(0,174)	0,499	(0,245)	0,370	(0,173)	0,501	(0,244)	0,370	(0,173)
			0,2					0,514	(0,251)	0,381	(0,179)	0,516	(0,250)	0,378	(0,179)	0,493	(0,253)	0,360*	(0,179)	0,529*	(0,252)	0,384	(0,178)
	7	0,4	0	0,497	(0,181)	0,619	(0,224)																
			0,05					0,493	(0,184)	0,613	(0,228)	0,510	(0,184)	0,635	(0,228)	0,508	(0,184)	0,631	(0,228)	0,498	(0,184)	0,620	(0,228)
			0,1					0,509	(0,189)	0,636	(0,234)	0,492	(0,187)	0,614	(0,232)	0,500	(0,189)	0,619	(0,233)	0,510	(0,188)	0,635	(0,232)
			0,2					0,505	(0,196)	0,627	(0,243)	0,506	(0,196)	0,629	(0,243)	0,506	(0,197)	0,629	(0,243)	0,507	(0,195)	0,629	(0,241)
		0,7	0	0,523*	(0,202)	0,650	(0,245)																
			0,05					0,516	(0,205)	0,640	(0,249)	0,498	(0,205)	0,621	(0,249)	0,491	(0,207)	0,607*	(0,251)	0,510	(0,206)	0,637	(0,250)
			0,1					0,488	(0,208)	0,611	(0,254)	0,499	(0,209)	0,620	(0,254)	0,514	(0,209)	0,636	(0,254)	0,512	(0,209)	0,634	(0,254)
			0,2					0,501	(0,215)	0,627	(0,263)	0,495	(0,216)	0,614	(0,262)	0,507	(0,215)	0,625	(0,262)	0,498	(0,215)	0,619	(0,262)
		0,9	0	0,503	(0,219)	0,622	(0,262)																
			0,05					0,495	(0,220)	0,606	(0,264)	0,489	(0,221)	0,605*	(0,265)	0,498	(0,219)	0,617	(0,264)	0,501	(0,220)	0,624	(0,265)
			0,1					0,498	(0,222)	0,630	(0,268)	0,516	(0,223)	0,633	(0,267)	0,517	(0,223)	0,642	(0,269)	0,502	(0,223)	0,620	(0,268)
			0,2					0,507	(0,229)	0,628	(0,277)	0,508	(0,230)	0,620	(0,277)	0,493	(0,227)	0,614	(0,275)	0,514	(0,227)	0,631	(0,274)

Suite page suivante...

TABLE D.8 – Suite

					no dropout		MCAR		MNAR							
									$\rho_{\theta\chi} = -0.4$		$\rho_{\theta\chi} = -0.7$		$\rho_{\theta\chi} = -0.9$			
					LRM	SM	LRM	SM	LRM	SM	LRM	SM	LRM	SM		
200	4	0,4	0	0,499	(0,142)	0,373	(0,104)									
			0,05					0,499	(0,144)	0,373	(0,106)	0,507	(0,145)	0,378	(0,107)	0,501
			0,1					0,499	(0,148)	0,374	(0,109)	0,496	(0,148)	0,372	(0,109)	0,495
			0,2					0,523*	(0,155)	0,390	(0,114)	0,513	(0,154)	0,383	(0,114)	0,502
		0,7	0	0,503	(0,156)	0,375	(0,113)									
			0,05					0,499	(0,159)	0,373	(0,115)	0,503	(0,158)	0,377	(0,114)	0,509
			0,1					0,502	(0,162)	0,376	(0,117)	0,497	(0,162)	0,372	(0,117)	0,514
			0,2					0,510	(0,168)	0,380	(0,122)	0,498	(0,167)	0,374	(0,121)	0,509
		0,9	0	0,509	(0,167)	0,378	(0,118)									
			0,05					0,499	(0,170)	0,372	(0,120)	0,499	(0,169)	0,373	(0,120)	0,504
			0,1					0,520*	(0,172)	0,385	(0,122)	0,502	(0,171)	0,371	(0,122)	0,497
			0,2					0,500	(0,177)	0,369	(0,127)	0,503	(0,177)	0,373	(0,126)	0,506
	7	0,4	0	0,497	(0,127)	0,621	(0,157)									
			0,05					0,510	(0,130)	0,637	(0,161)	0,505	(0,131)	0,630	(0,161)	0,499
			0,1					0,505	(0,134)	0,629	(0,165)	0,512*	(0,133)	0,639	(0,164)	0,500
			0,2					0,501	(0,139)	0,624	(0,172)	0,493	(0,139)	0,614*	(0,172)	0,508
		0,7	0	0,508	(0,143)	0,634	(0,173)									
			0,05					0,500	(0,145)	0,628	(0,176)	0,498	(0,146)	0,622	(0,177)	0,509
			0,1					0,508	(0,147)	0,633	(0,179)	0,496	(0,147)	0,620	(0,179)	0,511
			0,2					0,497	(0,151)	0,622	(0,185)	0,496	(0,152)	0,621	(0,186)	0,499
		0,9	0	0,513	(0,154)	0,639	(0,184)									
			0,05					0,490	(0,155)	0,607*	(0,186)	0,512	(0,155)	0,638	(0,187)	0,496
			0,1					0,488	(0,157)	0,610*	(0,190)	0,499	(0,158)	0,619	(0,190)	0,504
			0,2					0,502	(0,162)	0,626	(0,196)	0,500	(0,162)	0,616	(0,195)	0,509

					$\rho_{\theta\chi} = -0.7$		$\rho_{\theta\chi} = -0.9$	
					LRM	SM	LRM	SM
200	4	0,4	0,05		(0,144)	0,374	(0,107)	0,495
			0,1		(0,147)	0,370*	(0,109)	0,497
			0,2		(0,154)	0,376	(0,113)	0,498
		0,7	0,05		(0,159)	0,381	(0,115)	0,499
			0,1		(0,161)	0,385	(0,117)	0,504
			0,2		(0,167)	0,379	(0,121)	0,497
		0,9	0,05		(0,170)	0,373	(0,120)	0,509
			0,1		(0,171)	0,368*	(0,122)	0,501
			0,2		(0,176)	0,377	(0,126)	0,494
	7	0,4	0,05		(0,130)	0,623	(0,160)	0,497
			0,1		(0,133)	0,623	(0,164)	0,504
			0,2		(0,138)	0,633	(0,171)	0,494
		0,7	0,05		(0,145)	0,633	(0,176)	0,501
			0,1		(0,148)	0,634	(0,179)	0,501
			0,2		(0,152)	0,622	(0,185)	0,503
		0,9	0,05		(0,155)	0,612*	(0,186)	0,508
			0,1		(0,158)	0,629	(0,189)	0,498
			0,2		(0,161)	0,632	(0,195)	0,506

					$\rho_{\theta\chi} = -0.7$		$\rho_{\theta\chi} = -0.9$	
					LRM	SM	LRM	SM
200	4	0,4	0,05		(0,145)	0,370*	(0,107)	(0,107)
			0,1		(0,148)	0,370*	(0,109)	(0,109)
			0,2		(0,154)	0,371	(0,113)	(0,113)
		0,7	0,05		(0,159)	0,372	(0,115)	(0,115)
			0,1		(0,161)	0,375	(0,117)	(0,117)
			0,2		(0,167)	0,370	(0,121)	(0,121)
		0,9	0,05		(0,170)	0,378	(0,120)	(0,120)
			0,1		(0,172)	0,373	(0,122)	(0,122)
			0,2		(0,177)	0,364*	(0,126)	(0,126)
	7	0,4	0,05		(0,130)	0,621	(0,160)	(0,160)
			0,1		(0,133)	0,628	(0,164)	(0,164)
			0,2		(0,138)	0,615*	(0,171)	(0,171)
		0,7	0,05		(0,146)	0,627	(0,176)	(0,176)
			0,1		(0,148)	0,622	(0,179)	(0,179)
			0,2		(0,152)	0,627	(0,185)	(0,185)
		0,9	0,05		(0,156)	0,631	(0,187)	(0,187)
			0,1		(0,157)	0,620	(0,188)	(0,188)
			0,2		(0,162)	0,622	(0,195)	(0,195)

* test significatif à 5% - LRM : $H_0 : \mu_{\hat{\beta}_{gp}} = \Delta_\theta$ - SM : $H_0 : \mu_{\hat{\beta}_{gp}} = \Delta_S$ ($\Delta_S = 0.38$ pour J=4, $\Delta_S = 0.63$ pour J=7)

TABLE D.9 – Risque de première espèce de l'effet temps pour les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J), de la corrélation de la variable latente (ρ_θ), de la proportion de dropout $\pi^{(t)}$ et du type de dropout (ρ_{θ_χ}). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour la méthode SM et une structure UN pour la méthode LRM. Effet groupe simulé : $\Delta_\theta = 0.5$.

				no dropout		MCAR		MNAR					
								$\rho_{\theta\chi} = -0.4$		$\rho_{\theta\chi} = -0.7$		$\rho_{\theta\chi} = -0.9$	
N	J	ρ_θ	$\pi^{(t)}$	LRM	SM	LRM	SM	LRM	SM	LRM	SM	LRM	SM
100	4	0,4	0	0,042	0,038								
			0,05			0,050	0,044	0,044	0,046	0,054	0,054	0,070	0,070
			0,1			0,058	0,052	0,058	0,054	0,050	0,048	0,050	0,058
			0,2			0,056	0,050	0,049	0,042	0,063	0,062	0,072[!]	0,060
		0,7	0	0,066	0,050								
			0,05			0,042	0,034	0,046	0,040	0,044	0,044	0,071	0,068
			0,1			0,042	0,044	0,040	0,036	0,070	0,068	0,066	0,062
			0,2			0,044	0,042	0,059	0,058	0,062	0,062	0,062	0,074[!]
		0,9	0	0,034	0,030								
			0,05			0,040	0,036	0,048	0,054	0,044	0,048	0,046	0,050
			0,1			0,058	0,052	0,052	0,036	0,068	0,056	0,042	0,044
			0,2			0,067	0,062	0,038	0,038	0,042	0,032	0,072[!]	0,064
	7	0,4	0	0,040	0,036								
			0,05			0,066	0,054	0,066	0,048	0,052	0,048	0,062	0,056
			0,1			0,048	0,040	0,064	0,056	0,056	0,052	0,050	0,048
			0,2			0,056	0,048	0,081[!]	0,062	0,065	0,056	0,100[!]	0,084[!]
		0,7	0	0,038	0,038								
			0,05			0,054	0,050	0,060	0,054	0,048	0,060	0,048	0,050
			0,1			0,054	0,050	0,056	0,050	0,062	0,052	0,058	0,042
			0,2			0,070	0,062	0,047	0,044	0,071	0,064	0,066	0,070
		0,9	0	0,036	0,038								
			0,05			0,048	0,044	0,034	0,040	0,062	0,060	0,054	0,048
			0,1			0,048	0,038	0,052	0,046	0,063	0,062	0,052	0,052
			0,2			0,052	0,048	0,049	0,050	0,050	0,058	0,064	0,066
200	4	0,4	0	0,046	0,042								
			0,05			0,054	0,046	0,044	0,044	0,064	0,062	0,078[!]	0,070
			0,1			0,064	0,068	0,052	0,052	0,062	0,062	0,078[!]	0,076[!]
			0,2			0,054	0,046	0,066	0,062	0,085[!]	0,080[!]	0,100[!]	0,098[!]
		0,7	0	0,058	0,058								
			0,05			0,058	0,052	0,042	0,034	0,038	0,038	0,038	0,032
			0,1			0,038	0,036	0,070	0,062	0,060	0,054	0,060	0,050
			0,2			0,052	0,048	0,072[!]	0,076[!]	0,079[!]	0,074[!]	0,102[!]	0,098[!]
		0,9	0	0,061	0,054								
			0,05			0,044	0,060	0,026[!]	0,036	0,034	0,026[!]	0,060	0,054
			0,1			0,040	0,032	0,054	0,054	0,070	0,052	0,061	0,058
			0,2			0,030[!]	0,034	0,066	0,068	0,104[!]	0,108[!]	0,070	0,080[!]
	7	0,4	0	0,048	0,046								
			0,05			0,044	0,044	0,042	0,042	0,064	0,068	0,048	0,042
			0,1			0,034	0,032	0,056	0,048	0,056	0,052	0,074[!]	0,070
			0,2			0,058	0,050	0,046	0,046	0,088[!]	0,080[!]	0,110[!]	0,108[!]
		0,7	0	0,042	0,038								
			0,05			0,040	0,030	0,058	0,060	0,066	0,060	0,046	0,040
			0,1			0,048	0,042	0,058	0,058	0,058	0,048	0,066	0,062
			0,2			0,058	0,046	0,088[!]	0,072[!]	0,057	0,050	0,106[!]	0,096[!]
		0,9	0	0,046	0,034								
			0,05			0,047	0,046	0,042	0,034	0,034	0,030[!]	0,050	0,046
			0,1			0,052	0,046	0,054	0,062	0,057	0,056	0,054	0,046
			0,2			0,044	0,050	0,047	0,030[!]	0,079[!]	0,084[!]	0,074[!]	0,080[!]

[!] l'intervalle de confiance à 95% ne contient pas la valeur attendue 5%.

TABLE D.10 – Puissance de l'effet temps pour les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J), de la corrélation de la variable latente (ρ_θ), de la proportion de dropout $\pi^{(t)}$ et du type de dropout ($\rho_{\theta\chi}$). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour la méthode SM et une structure UN pour la méthode LRM. Effet groupe simulé : $\Delta_\theta = 0.5$.

				no dropout		MCAR		MNAR					
								$\rho_{\theta\chi} = -0.4$		$\rho_{\theta\chi} = -0.7$		$\rho_{\theta\chi} = -0.9$	
N	J	ρ_θ	$\pi^{(t)}$	LRM	SM	LRM	SM	LRM	SM	LRM	SM	LRM	SM
100	4	0,4	0	0,412	0,404								
			0,05			0,366	0,364	0,416	0,404	0,383	0,372	0,426	0,410
			0,1			0,344	0,326	0,386	0,386	0,423	0,402	0,422	0,428
			0,2			0,324	0,312	0,349	0,360	0,439	0,444	0,470	0,468
		0,7	0	0,435	0,390								
			0,05			0,411	0,376	0,469	0,416	0,431	0,380	0,441	0,400
			0,1			0,367	0,334	0,446	0,410	0,480	0,438	0,512	0,484
			0,2			0,336	0,310	0,428	0,392	0,485	0,462	0,550	0,538
		0,9	0	0,500	0,428								
			0,05			0,493	0,400	0,462	0,380	0,514	0,442	0,508	0,420
			0,1			0,451	0,396	0,524	0,440	0,538	0,472	0,543	0,502
			0,2			0,367	0,296	0,466	0,430	0,470	0,436	0,608	0,570
	7	0,4	0	0,527	0,498								
			0,05			0,488	0,478	0,496	0,478	0,554	0,532	0,484	0,470
			0,1			0,490	0,466	0,480	0,468	0,528	0,510	0,544	0,528
			0,2			0,404	0,372	0,528	0,498	0,562	0,560	0,558	0,568
		0,7	0	0,590	0,522								
			0,05			0,580	0,516	0,610	0,548	0,582	0,528	0,592	0,534
			0,1			0,577	0,522	0,610	0,538	0,648	0,586	0,640	0,592
			0,2			0,514	0,470	0,573	0,516	0,649	0,618	0,684	0,648
		0,9	0	0,753	0,628								
			0,05			0,652	0,554	0,691	0,588	0,711	0,632	0,729	0,622
			0,1			0,629	0,534	0,670	0,574	0,706	0,608	0,727	0,654
			0,2			0,590	0,480	0,609	0,548	0,729	0,660	0,744	0,704
200	4	0,4	0	0,668	0,662								
			0,05			0,622	0,616	0,704	0,680	0,696	0,694	0,686	0,690
			0,1			0,638	0,618	0,642	0,632	0,676	0,666	0,768	0,742
			0,2			0,546	0,546	0,670	0,672	0,749	0,724	0,772	0,782
		0,7	0	0,772	0,724								
			0,05			0,709	0,676	0,696	0,662	0,741	0,706	0,795	0,760
			0,1			0,685	0,644	0,715	0,684	0,784	0,740	0,810	0,796
			0,2			0,633	0,598	0,728	0,688	0,858	0,834	0,866	0,846
		0,9	0	0,804	0,716								
			0,05			0,795	0,714	0,834	0,770	0,835	0,746	0,841	0,790
			0,1			0,766	0,698	0,809	0,756	0,840	0,794	0,850	0,800
			0,2			0,706	0,630	0,755	0,730	0,847	0,816	0,866	0,860
	7	0,4	0	0,822	0,810								
			0,05			0,788	0,778	0,792	0,770	0,784	0,768	0,810	0,804
			0,1			0,746	0,726	0,806	0,806	0,842	0,830	0,836	0,820
			0,2			0,718	0,716	0,834	0,810	0,852	0,860	0,876	0,872
		0,7	0	0,902	0,864								
			0,05			0,866	0,834	0,868	0,834	0,898	0,878	0,886	0,854
			0,1			0,842	0,814	0,880	0,856	0,902	0,872	0,918	0,882
			0,2			0,768	0,726	0,869	0,852	0,913	0,896	0,942	0,938
		0,9	0	0,957	0,918								
			0,05			0,926	0,890	0,941	0,884	0,960	0,924	0,960	0,920
			0,1			0,923	0,870	0,949	0,918	0,949	0,924	0,962	0,932
			0,2			0,846	0,796	0,929	0,908	0,964	0,946	0,988	0,972

en gris : l'intervalle de confiance à 95% du risque de première espèce ne contient pas la valeur attendue 5%

TABLE D.11 – Estimations de l'effet temps entre le temps 1 et le temps 2 ($\hat{d}_{12}$) pour les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J), de la corrélation de la variable latente (ρ_θ), de la proportion de dropout ($\pi^{(t)}$) et du type de dropout (ρ_{θ_χ}). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour la méthode SM et une structure UN pour la méthode LRM. Valeurs moyennes de $\hat{d}_{12}$ et écarts-types (s.d.).Effet groupe simulé : $\Delta_\theta = 0.5$. Effet temps simulé : $d_\theta = 0$. $d_S = 0$.

				no dropout				MCAR				MNAR											
												$\rho_{\theta\chi} = -0.4$				$\rho_{\theta\chi} = -0.7$				$\rho_{\theta\chi} = -0.9$			
				LRM		SM		LRM		SM		LRM		SM		LRM		SM		LRM		SM	
N	J	ρ_θ	$\pi^{(t)}$	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)
100	4	0,4	0	-0,001	(0,208)	-0,001	(0,154)																
			0,05					0,001	(0,201)	0,000	(0,148)	0,007	(0,194)	0,006	(0,143)	0,013	(0,197)	0,010	(0,147)	0,014	(0,219)	0,010	(0,163)
			0,1					0,011	(0,208)	0,008	(0,155)	0,023*	(0,209)	0,018*	(0,155)	0,032*	(0,198)	0,024*	(0,146)	0,057*	(0,207)	0,041*	(0,154)
			0,2					-0,004	(0,224)	-0,003	(0,167)	0,039*	(0,215)	0,033*	(0,160)	0,059*	(0,214)	0,042*	(0,159)	0,085*	(0,203)	0,063*	(0,150)
		0,7	0	0,002	(0,196)	0,002	(0,145)																
			0,05					-0,009	(0,189)	-0,007	(0,140)	0,007	(0,184)	0,005	(0,135)	0,003	(0,191)	0,002	(0,141)	0,009	(0,181)	0,006	(0,134)
			0,1					0,003	(0,193)	0,002	(0,142)	0,028*	(0,189)	0,021*	(0,140)	0,034*	(0,192)	0,025*	(0,141)	0,033*	(0,197)	0,024*	(0,145)
			0,2					0,002	(0,190)	0,001	(0,141)	0,019*	(0,190)	0,015*	(0,141)	0,056*	(0,204)	0,041*	(0,149)	0,074*	(0,193)	0,054*	(0,142)
		0,9	0	-0,009	(0,173)	-0,007	(0,127)																
			0,05					0,001	(0,166)	0,001	(0,122)	0,006	(0,179)	0,004	(0,130)	0,011	(0,169)	0,009	(0,125)	0,016*	(0,176)	0,012*	(0,129)
			0,1					-0,001	(0,179)	-0,001	(0,131)	0,011	(0,185)	0,008	(0,135)	0,014	(0,176)	0,010	(0,128)	0,030*	(0,171)	0,021*	(0,125)
			0,2					-0,002	(0,198)	-0,002	(0,144)	0,019*	(0,181)	0,013*	(0,131)	0,044*	(0,179)	0,031*	(0,131)	0,053*	(0,193)	0,037*	(0,142)
	7	0,4	0	-0,009	(0,172)	-0,011	(0,213)																
			0,05					0,003	(0,185)	0,003	(0,230)	-0,005	(0,166)	-0,005	(0,205)	0,009	(0,172)	0,011	(0,214)	0,012	(0,171)	0,015	(0,212)
			0,1					-0,001	(0,168)	-0,001	(0,209)	0,017*	(0,179)	0,021*	(0,222)	0,009	(0,169)	0,011	(0,209)	0,030*	(0,173)	0,037*	(0,213)
			0,2					-0,009	(0,184)	-0,012	(0,228)	0,045*	(0,185)	0,054*	(0,227)	0,058*	(0,173)	0,075*	(0,216)	0,069*	(0,180)	0,085*	(0,224)
		0,7	0	0,009	(0,147)	0,011	(0,182)																
			0,05					0,010	(0,151)	0,012	(0,186)	0,014*	(0,159)	0,017	(0,197)	0,012	(0,160)	0,014	(0,199)	0,018*	(0,154)	0,021*	(0,191)
			0,1					0,009	(0,155)	0,009	(0,193)	0,012	(0,152)	0,013	(0,189)	0,026*	(0,161)	0,032*	(0,200)	0,021*	(0,157)	0,026*	(0,195)
			0,2					-0,002	(0,172)	-0,004	(0,211)	0,029*	(0,165)	0,035*	(0,203)	0,061*	(0,159)	0,070*	(0,194)	0,074*	(0,160)	0,092*	(0,198)
		0,9	0	0,004	(0,133)	0,005	(0,164)																
			0,05					-0,007	(0,141)	-0,009	(0,174)	0,010	(0,137)	0,013	(0,169)	-0,005	(0,147)	-0,006	(0,181)	0,003	(0,145)	0,003	(0,179)
			0,1					-0,001	(0,140)	-0,002	(0,173)	0,016*	(0,144)	0,020*	(0,176)	0,021*	(0,147)	0,026*	(0,182)	0,011	(0,138)	0,013	(0,169)

Suite page suivante...

Table D.11 – Suite

				no dropout				MCAR				MNAR											
												$\rho_{\theta\chi} = -0.4$				$\rho_{\theta\chi} = -0.7$				$\rho_{\theta\chi} = -0.9$			
				LRM		SM		LRM		SM		LRM		SM		LRM		SM		LRM		SM	
N	J	ρ_θ	$\pi^{(t)}$	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)
			0,2					-0,004	(0,149)	-0,005	(0,183)	0,028*	(0,143)	0,036*	(0,175)	0,036*	(0,149)	0,047*	(0,184)	0,050*	(0,150)	0,061*	(0,186)
200	4	0,4	0	0,012	(0,146)	0,009	(0,109)																
			0,05					-0,009	(0,144)	-0,007	(0,108)	0,002	(0,141)	0,001	(0,105)	0,017*	(0,145)	0,013*	(0,107)	0,011	(0,143)	0,008	(0,107)
			0,1					0,004	(0,152)	0,002	(0,112)	0,022*	(0,145)	0,017*	(0,108)	0,032*	(0,141)	0,024*	(0,104)	0,050*	(0,146)	0,037*	(0,108)
			0,2					0,004	(0,147)	0,003	(0,110)	0,026*	(0,144)	0,019*	(0,107)	0,058*	(0,149)	0,044*	(0,110)	0,086*	(0,145)	0,064*	(0,108)
		0,7	0	-0,002	(0,128)	-0,002	(0,095)																
			0,05					0,000	(0,127)	0,000	(0,095)	0,003	(0,124)	0,002	(0,092)	0,001	(0,127)	0,000	(0,095)	0,007	(0,125)	0,005	(0,093)
			0,1					-0,004	(0,128)	-0,003	(0,095)	0,020*	(0,138)	0,015*	(0,102)	0,020*	(0,138)	0,015*	(0,102)	0,038*	(0,128)	0,028*	(0,095)
			0,2					-0,003	(0,142)	-0,002	(0,106)	0,034*	(0,134)	0,024*	(0,100)	0,060*	(0,141)	0,045*	(0,106)	0,079*	(0,144)	0,059*	(0,107)
		0,9	0	0,004	(0,128)	0,003	(0,095)																
			0,05					-0,003	(0,128)	-0,003	(0,096)	0,002	(0,120)	0,002	(0,089)	0,012*	(0,117)	0,009*	(0,086)	0,013*	(0,126)	0,010*	(0,094)
			0,1					0,001	(0,126)	0,001	(0,093)	0,016*	(0,115)	0,012*	(0,085)	0,023*	(0,122)	0,017*	(0,090)	0,029*	(0,122)	0,021*	(0,090)
			0,2					0,015*	(0,129)	0,011*	(0,095)	0,027*	(0,132)	0,021*	(0,096)	0,040*	(0,136)	0,030*	(0,101)	0,061*	(0,134)	0,044*	(0,099)
	7	0,4	0	-0,002	(0,119)	-0,002	(0,148)																
			0,05					0,004	(0,120)	0,005	(0,149)	0,006	(0,124)	0,008	(0,154)	0,016*	(0,125)	0,021*	(0,156)	0,013*	(0,115)	0,016*	(0,144)
			0,1					-0,003	(0,122)	-0,004	(0,151)	0,014*	(0,118)	0,017*	(0,147)	0,034*	(0,120)	0,043*	(0,149)	0,032*	(0,127)	0,040*	(0,159)
			0,2					0,006	(0,131)	0,007	(0,163)	0,030*	(0,123)	0,034*	(0,155)	0,062*	(0,132)	0,078*	(0,162)	0,078*	(0,130)	0,098*	(0,161)
		0,7	0	-0,001	(0,102)	0,000	(0,128)																
			0,05					-0,002	(0,105)	-0,003	(0,130)	0,002	(0,099)	0,003	(0,123)	0,008	(0,111)	0,009	(0,138)	0,016*	(0,106)	0,020*	(0,132)
			0,1					0,005	(0,116)	0,006	(0,145)	0,014*	(0,114)	0,017*	(0,142)	0,021*	(0,111)	0,026*	(0,139)	0,029*	(0,111)	0,036*	(0,139)
			0,2					-0,002	(0,114)	-0,003	(0,142)	0,024*	(0,116)	0,028*	(0,145)	0,042*	(0,115)	0,052*	(0,144)	0,071*	(0,111)	0,088*	(0,138)
		0,9	0	0,000	(0,098)	0,000	(0,122)																
			0,05					0,006	(0,099)	0,007	(0,124)	0,001	(0,097)	0,001	(0,120)	0,001	(0,096)	0,002	(0,120)	0,014*	(0,097)	0,018*	(0,121)
			0,1					0,002	(0,100)	0,002	(0,125)	0,011*	(0,097)	0,014*	(0,120)	0,025*	(0,096)	0,030*	(0,120)	0,017*	(0,100)	0,021*	(0,124)
			0,2					0,001	(0,105)	0,001	(0,130)	0,022*	(0,098)	0,027*	(0,121)	0,046*	(0,103)	0,055*	(0,128)	0,049*	(0,100)	0,060*	(0,124)

* indique que le test de Student est significatif à 5% ($H_0 : \mu_{\hat{d}_{12}} = 0$)

TABLE D.12 – Estimations de l'effet temps entre le temps 1 et le temps 2 ($\hat{d}_{12}$) pour les méthodes Score and Mixed models (SM) et Longitudinal Rasch Mixed model (LRM) pour différentes valeurs de la taille d'échantillon (N), du nombre d'items (J), de la corrélation de la variable latente (ρ_θ), de la proportion de dropout ($\pi^{(t)}$) et du type de dropout (ρ_{θ_χ}). Résultats d'analyses avec une structure de matrice de variance covariance AR(1) pour la méthode SM et une structure UN pour la méthode LRM. Valeurs moyennes de $\hat{d}_{12}$ et écarts-types (s.d.). Effet groupe simulé : $\Delta_\theta = 0.5$. Effet temps simulé : $d_\theta = 0, 2$. $d_S = 0.15$ pour J=4, $d_S = 0.25$ pour J=7.

				no dropout				MCAR				MNAR											
												$\rho_{\theta\chi} = -0.4$				$\rho_{\theta\chi} = -0.7$				$\rho_{\theta\chi} = -0.9$			
				LRM		SM		LRM		SM		LRM		SM		LRM		SM		LRM		SM	
N	J	ρ_θ	$\pi^{(t)}$	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)
100	4	0,4	0	0,203	(0,192)	0,152	(0,141)																
			0,05					0,193	(0,207)	0,143	(0,153)	0,199	(0,199)	0,147	(0,148)	0,214	(0,207)	0,159	(0,153)	0,214	(0,200)	0,159	(0,148)
			0,1					0,204	(0,212)	0,152	(0,157)	0,212	(0,210)	0,157	(0,155)	0,231*	(0,220)	0,171*	(0,162)	0,242*	(0,212)	0,180*	(0,158)
			0,2					0,189	(0,229)	0,141	(0,169)	0,242*	(0,214)	0,179*	(0,157)	0,278*	(0,205)	0,205*	(0,151)	0,277*	(0,210)	0,205*	(0,155)
		0,7	0	0,185	(0,190)	0,137	(0,140)																
			0,05					0,197	(0,188)	0,146	(0,138)	0,208	(0,188)	0,153	(0,139)	0,215	(0,183)	0,159	(0,135)	0,212	(0,189)	0,157	(0,140)
			0,1					0,204	(0,181)	0,151	(0,132)	0,229*	(0,187)	0,169*	(0,139)	0,231*	(0,202)	0,171*	(0,147)	0,239*	(0,191)	0,176*	(0,141)
			0,2					0,213	(0,195)	0,157	(0,143)	0,244*	(0,196)	0,179*	(0,144)	0,270*	(0,197)	0,199*	(0,145)	0,279*	(0,196)	0,207*	(0,145)
		0,9	0	0,196	(0,177)	0,146	(0,129)																
			0,05					0,203	(0,173)	0,150	(0,127)	0,210	(0,174)	0,154	(0,127)	0,210	(0,181)	0,154	(0,132)	0,208	(0,165)	0,152	(0,119)
			0,1					0,200	(0,188)	0,147	(0,136)	0,226*	(0,183)	0,164*	(0,133)	0,226*	(0,188)	0,166*	(0,137)	0,231*	(0,186)	0,170*	(0,135)
			0,2					0,179*	(0,189)	0,132*	(0,138)	0,234*	(0,192)	0,171*	(0,141)	0,249*	(0,180)	0,183*	(0,131)	0,272*	(0,193)	0,199*	(0,139)
	7	0,4	0	0,209	(0,168)	0,259	(0,207)																
			0,05					0,214	(0,170)	0,264	(0,212)	0,216*	(0,172)	0,269*	(0,214)	0,226*	(0,163)	0,281*	(0,202)	0,223*	(0,167)	0,277*	(0,206)
			0,1					0,215	(0,178)	0,265	(0,219)	0,218*	(0,183)	0,271*	(0,225)	0,225*	(0,173)	0,279*	(0,214)	0,221*	(0,188)	0,274*	(0,234)
			0,2					0,205	(0,177)	0,253	(0,217)	0,255*	(0,191)	0,314*	(0,235)	0,259*	(0,183)	0,324*	(0,226)	0,269*	(0,182)	0,333*	(0,226)
		0,7	0	0,199	(0,152)	0,247	(0,187)																
			0,05					0,206	(0,162)	0,256	(0,201)	0,197	(0,148)	0,243	(0,182)	0,206	(0,149)	0,255	(0,184)	0,211	(0,155)	0,262	(0,192)
			0,1					0,209	(0,156)	0,259	(0,193)	0,214	(0,163)	0,266	(0,201)	0,224*	(0,153)	0,277*	(0,188)	0,234*	(0,152)	0,290*	(0,187)
			0,2					0,199	(0,172)	0,246	(0,211)	0,228*	(0,160)	0,283*	(0,196)	0,240*	(0,169)	0,301*	(0,208)	0,266*	(0,157)	0,330*	(0,196)
		0,9	0	0,210	(0,137)	0,260	(0,167)																
			0,05					0,211	(0,135)	0,26	(0,165)	0,218*	(0,136)	0,269*	(0,166)	0,215*	(0,141)	0,265*	(0,173)	0,218*	(0,136)	0,270*	(0,168)
			0,1					0,197	(0,142)	0,242	(0,176)	0,220*	(0,138)	0,272*	(0,170)	0,206	(0,142)	0,254	(0,175)	0,230*	(0,145)	0,284*	(0,177)

Suite page suivante...

Table D.12 – Suite

				no dropout				MCAR				MNAR									
												$\rho_{\theta\chi} = -0.4$				$\rho_{\theta\chi} = -0.7$				$\rho_{\theta\chi} = -0.9$			
				LRM		SM		LRM		SM		LRM		SM		LRM		SM		LRM		SM	
N	J	ρ_θ	$\pi^{(t)}$	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)	$\hat{d}_{12}$	(s.d.)
			0,2					0,205	(0,155)	0,252	(0,191)	0,223*	(0,149)	0,273*	(0,183)	0,244*	(0,149)	0,299*	(0,183)	0,254*	(0,147)	0,313*	(0,181)
200	4	0,4	0	0,204	(0,149)	0,152	(0,111)																
			0,05					0,202	(0,143)	0,150	(0,107)	0,212	(0,146)	0,158	(0,108)	0,210	(0,144)	0,157	(0,107)	0,215*	(0,140)	0,160*	(0,105)
			0,1					0,207	(0,150)	0,154	(0,110)	0,217*	(0,149)	0,162*	(0,112)	0,215*	(0,144)	0,160*	(0,107)	0,234*	(0,141)	0,175*	(0,104)
			0,2					0,207	(0,154)	0,153	(0,114)	0,225*	(0,154)	0,167*	(0,113)	0,259*	(0,142)	0,193*	(0,106)	0,280*	(0,148)	0,209*	(0,109)
		0,7	0	0,206	(0,130)	0,154	(0,097)																
			0,05					0,205	(0,128)	0,152	(0,095)	0,203	(0,127)	0,151	(0,094)	0,211	(0,133)	0,157	(0,099)	0,213*	(0,126)	0,158*	(0,093)
			0,1					0,201	(0,137)	0,148	(0,101)	0,226*	(0,123)	0,168*	(0,091)	0,223*	(0,133)	0,167*	(0,099)	0,241*	(0,135)	0,180*	(0,101)
			0,2					0,206	(0,143)	0,153	(0,105)	0,230*	(0,147)	0,170*	(0,109)	0,262*	(0,133)	0,193*	(0,099)	0,273*	(0,148)	0,203*	(0,110)
		0,9	0	0,191	(0,122)	0,142	(0,090)																
			0,05					0,200	(0,121)	0,149	(0,091)	0,210	(0,129)	0,156	(0,095)	0,215*	(0,123)	0,159*	(0,091)	0,204	(0,126)	0,151	(0,094)
			0,1					0,201	(0,125)	0,149	(0,093)	0,212*	(0,126)	0,156	(0,093)	0,228*	(0,129)	0,168*	(0,095)	0,224*	(0,117)	0,166*	(0,086)
			0,2					0,212	(0,137)	0,156	(0,101)	0,226*	(0,131)	0,168*	(0,096)	0,247*	(0,129)	0,183*	(0,096)	0,264*	(0,133)	0,195*	(0,098)
	7	0,4	0	0,201	(0,123)	0,250	(0,153)																
			0,05					0,191	(0,121)	0,238	(0,151)	0,211*	(0,118)	0,263	(0,147)	0,207	(0,124)	0,258	(0,154)	0,219*	(0,124)	0,272*	(0,153)
			0,1					0,206	(0,119)	0,256	(0,148)	0,213*	(0,121)	0,265*	(0,151)	0,222*	(0,120)	0,277*	(0,149)	0,228*	(0,115)	0,283*	(0,143)
			0,2					0,194	(0,130)	0,242	(0,162)	0,237*	(0,127)	0,294*	(0,158)	0,257*	(0,134)	0,322*	(0,168)	0,283*	(0,131)	0,353*	(0,162)
		0,7	0	0,209	(0,104)	0,260	(0,129)																
			0,05					0,206	(0,104)	0,257	(0,129)	0,208	(0,104)	0,258	(0,128)	0,213*	(0,111)	0,265*	(0,138)	0,203	(0,105)	0,252	(0,131)
			0,1					0,201	(0,114)	0,250	(0,141)	0,207	(0,104)	0,258	(0,129)	0,223*	(0,102)	0,277*	(0,126)	0,239*	(0,103)	0,297*	(0,129)
			0,2					0,205	(0,114)	0,256	(0,142)	0,236*	(0,119)	0,292*	(0,148)	0,255*	(0,110)	0,317*	(0,136)	0,258*	(0,115)	0,320*	(0,143)
		0,9	0	0,200	(0,099)	0,249	(0,122)																
			0,05					0,199	(0,096)	0,248	(0,119)	0,204	(0,099)	0,253	(0,122)	0,203	(0,093)	0,252	(0,115)	0,216*	(0,100)	0,268*	(0,123)
			0,1					0,205	(0,103)	0,255	(0,128)	0,212*	(0,101)	0,264*	(0,126)	0,217*	(0,101)	0,270*	(0,126)	0,224*	(0,099)	0,278*	(0,123)
			0,2					0,198	(0,108)	0,246	(0,133)	0,233*	(0,102)	0,288*	(0,127)	0,242*	(0,102)	0,301*	(0,124)	0,253*	(0,102)	0,313*	(0,125)

* indique que le test de Student est significatif à 5% - LRM : $H_0 : \mu_{\hat{d}_{12}} = d_\theta$ - SM : $H_0 : \mu_{\hat{d}_{12}} = d_S$ ($d_S = 0.15$ pour J=4, $d_S = 0.25$ pour J=7)

Bibliographie

[1] N. K. Aaronson, S. Ahmedzai, B. Bergman, M. Bullinger, A. Cull, N. J. Duez, A. Filiberti, H. Flechtner, S. B. Fleishman, and J. C. de Haes. The european organization for research and treatment of cancer QLQ-C30 : a quality-of-life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute*, 85(5) :365–376, 1993.

[2] C. Acquadro, R. Berzon, D. Dubois, N. K. Leidy, P. Marquis, D. Revicki, M. Rothman, and P. H. Group. Incorporating the patient's perspective into drug development and communication : An ad hoc task force report of the Patient-Reported outcomes (PRO) harmonization group meeting at the food and drug administration, february 16, 2001. *Value in Health*, 6(5) :522–531, 2003.

[3] H. Akaike. A new look at the statistical model identification. system identification and time-series analysis. *IEEE Transactions on Automatic Control*, 19(6) :716–723, 1974.

[4] E. B. Andersen. Asymptotic properties of conditional Maximum-Likelihood estimators. *Journal of the Royal Statistical Society. Series B (Methodological)*, 32(2) :283–301, 1970.

[5] E. B. Andersen. The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(1) :42–54, 1972.

[6] E. B. Andersen. A goodness of fit test for the rasch model. *Psychometrika*, 38(1) :123–140, 1973.

[7] E. B. Andersen. Sufficient statistics and latent trait models. *Psychometrika*, 42 :69–81, 1977.

[8] E. B. Andersen. Estimating latent correlations between repeated testings. *Psychometrika*, 50(1) :3–16, 1985.

[9] D. Andrich. A rating formulation for ordered response categories. *Psychometrika*, 43(4) :561–573, Dec. 1978.

[10] D. Andrich and G. Luo. Conditional pairwise estimation in the rasch model for ordered response categories using principal components. *Journal of Applied Measurement*, 4(3) :205–21, 2003.

[11] R. Barclay-Goddard, J. D. Epstein, and N. E. Mayo. Response shift : a brief overview and proposed research priorities. *Quality of Life Research*, 18(3) :335–346, 2009.

[12] A. Birnbaum. Some latent trait models and their use in inferring an examinee's ability. In *Statistical theories of mental test scores*. F. M. Lord & M. R. Novick, New York, Addison-Wesley edition, 1968.

[13] M. Blanchin, J. Hardouin, T. L. Neel, G. Kubis, C. Blanchard, E. Mirallié, and V. Sébille. Comparison of CTT and Rasch-based approaches for the analysis of longitudinal Patient Reported Outcomes. *Statistics in Medicine*, 30(8) :825–838, 2011.

[14] M. Blanchin, J. Hardouin, T. L. Neel, G. Kubis, and V. Sébille. Analysis of longitudinal Patient-Reported Outcomes with informative and non-informative dropout : Comparison of CTT and Rasch-based methods. *International Journal of Applied Mathematics & Statistics*, 24(SI-11A) :107–124, 2011.

[15] R. D. Bock and M. Aitkin. Marginal maximum likelihood estimation of item parameters : Application of an EM algorithm. *Psychometrika*, 46(4) :443–459, 1981.

[16] B. C. Brown, S. P. McKenna, M. Solomon, J. Wilburn, D. A. McGrouther, and A. Bayat. The patient-reported impact of scars measure : development and validation. *Plastic and Reconstructive Surgery*, 125(5) :1439–1449, May 2010.

[17] W. Brown. Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3(3) :296–322, 1910.

[18] C. Caillard, F. Sebag, M. Mathonnet, H. Gibelin, L. Brunaud, C. Loudot, J. Kraimps, A. Hamy, L. Bresler, B. Charbonnel, J. Leborgne, J. Henry, J. Nguyen, and E. Mirallié. Prospective evaluation of quality of life (SF-36v2) and nonspecific symptoms before and after cure of primary hyperparathyroidism (1-year follow-up). *Surgery*, 141(2) :153–160, Feb. 2007.

[19] C. C. Chen and R. K. Bode. Psychometric validation of the manual ability measure-36 (MAM-36) in patients with neurologic and musculoskeletal disorders. *Archives of Physical Medicine and Rehabilitation*, 91(3) :414–420, Mar. 2010.

[20] Committee for medicinal products for human use. Reflection paper on the regulatory guidance for the use of healthrelated quality of life (HRQL) measures in the evaluation of medicinal products. Technical Report EMEA/CHMP/EWP/139391/2004, European Medicines Agency, Londres, July 2005.

[21] F. Cousson, M. Bruchon-Schweitzer, B. Quintard, J. Nuissier, and N. Rascle. Analyse multidimensionnelle d'une échelle de coping : validation française de la W.C.C. (ways of coping checklist). *Psychologie française*, 41(2) :155–164, 1996.

[22] L. J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3) :297–334, 1951.

[23] D. Curran, M. Bacchi, S. F. Schmitz, G. Molenberghs, and R. J. Sylvester. Identifying the types of missingness in quality of life data from clinical trials. *Statistics in Medicine*, 17(5-7) :739–56, 1998.

[24] D. Curran, G. Molenberghs, P. M. Fayers, and D. Machin. Incomplete quality of life data in randomized trials : missing forms. *Statistics in Medicine*, 17(5-7) :697–709, 1998.

[25] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1) :1–38, 1977.

[26] P. Diggle and M. G. Kenward. Informative Drop-Out in longitudinal data analysis. *Applied Statistics*, 43(1) :49–93, 1994.

[27] E. M. Durna, S. M. Crowe, L. R. Leader, and J. A. Eden. Quality of life of breast cancer survivors : the impact of hormonal replacement therapy. *Climacteric*, 5(3) :266–276, 2002.

[28] S. E. Embretson. The new rules of measurement. *Psychological Assessment*, 8(4) :341–349, 1996.

[29] S. E. Embretson and S. P. Reise. The new rules of measurement. In *Item response theory for psychologists*, Multivariate Applications Series. Lawrence Erlbaum Associates Inc, 2000.

[30] B. Falissard. *Mesurer la subjectivité en santé : Perspective méthodologique et statistique*. Masson, 2001.

[31] X. Fan. Item response theory and classical test theory : An empirical comparison of their Item/Person statistics. *Educational and Psychological Measurement*, 58(3) :357 –381, June 1998.

[32] P. Fayers, N. K. Aaronson, K. Bjordal, D. Curran, and M. Groenvold on behalf of the EORTC Quality of Life Study Group. *EORTC QLQ-C30 Scoring Manual (Third edition)*. EORTC Quality of Life Group, Brussels, 2001.

[33] P. M. Fayers, D. Curran, and D. Machin. Incomplete quality of life data in randomized trials : missing items. *Statistics in Medicine*, 17(5-7) :679–96, 1998.

[34] L. S. Feldt and R. L. Brennan. Reliability. In R. L. Linn, editor, *Educational Measurement*, pages 105–146. Macmillan USA, New York, 3rd edition, 1989.

[35] G. Fischer. Derivations of the rasch model. In G. H. Fischer and I. W. Molenaar, editors, *Rasch models : foundations, recent developments, and applications*. Springer, 1995.

[36] G. Fischer and P. Parzer. An extension of the rating scale model with an application to the measurement of change. *Psychometrika*, 56 :637–651, 1991.

[37] G. Fischer and I. Ponocny. An extension of the partial credit model with an application to the measurement of change. *Psychometrika*, 59(2) :177–192, 1994.

[38] G. H. Fischer. Logistic latent trait models with linear constraints. *Psychometrika*, 48(1) :3–26, 1983.

[39] G. H. Fischer. Linear logistic models for change. In I. W. Molenaar and G. Fischer, editors, *Rasch models : foundations, recent developments, and applications*. Springer, New York, Apr. 1995.

[40] R. A. Fisher. *Statistical methods for research workers*. Oliver and Boyd, Edinburgh, 1925.

[41] G. M. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs. *Longitudinal data analysis*. Chapman and Hall/CRC, 2009.

[42] G. M. Fitzmaurice, N. M. Laird, and J. H. Ware. Estimation and statistical inference. In *Applied longitudinal analysis*, Wiley Series in Probability and Statistics. Wiley-IEEE, Hoboken, 2004.

[43] C. A. Glas, H. Geerlings, M. A. van de Laar, and E. Taal. Analysis of longitudinal randomized clinical trials using item response models. *Contemporary Clinical Trials*, 30(2) :158–170, 2009.

[44] C. A. W. Glas and I. Hendrawan. Testing linear models for ability parameters in item response models. *Multivariate Behavioral Research*, 40(1) :25, 2005.

[45] R. Glynn, N. Laird, and D. Rubin. Selection modelling versus mixture modelling with nonignorable nonresponse. In H. Wainer, editor, *Drawing Inferences from Self-selected Samples*, pages 115–142. New York : Springer-Verlag, 1986.

[46] C. C. Gotay, C. T. Kawamoto, A. Bottomley, and F. Efficace. The prognostic significance of Patient-Reported outcomes in cancer clinical trials. *J Clin Oncol*, 26(8) :1355–1363, 2008.

[47] H. Gulliksen. *Theory of Mental Tests*. John Wiley & Sons Inc, New York, Dec. 1950.

[48] R. K. Hambleton and R. W. Jones. Comparison of classical test theory and item response theory and their application to test development. *Educational Measurement : Issues and Practice*, 12(3) :38–47, 1993.

[49] R. K. Hambleton and W. J. van der Linden. Advances in item response theory and applications : An introduction. *Applied Psychological Measurement*, 6(4) :373–378, 1982.

[50] J.-B. Hardouin. Rasch analysis : Estimation and tests with raschtest. *Stata Journal*, 7(1) :22–44(23), 2007.

[51] D. A. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358) :320–338, 1977.

[52] R. Holman and C. A. W. Glas. Modelling non-ignorable missing-data mechanisms with item response theory models. *The British Journal of Mathematical and Statistical Psychology*, 58(Pt 1) :1–17, May 2005. PMID : 15969835.

[53] M. G. Kenward and J. H. Roger. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53(3) :983–997, 1997.

[54] N. M. Laird and J. H. Ware. Random-Effects models for longitudinal data. *Biometrics*, 38(4) :963–974, 1982.

[55] S. Lawson. One parameter latent trait measurement : Do the results justify the effort ? In B. Thompson, editor, *Advances in Educational Research : Substantive findings, methodological developments*, volume 1, pages 159–168. Greenwich, CT : JAI, 1991.

[56] A. Leplège, E. Ecosse, A. Verdier, and T. V. Perneger. The french SF-36 health survey : translation, cultural adaptation and preliminary psychometric evaluation. *Journal of Clinical Epidemiology*, 51(11) :1013–23, Nov. 1998.

[57] K.-Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1) :13–22, 1986.

[58] M. J. Lindstrom and D. M. Bates. Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404) :1014–1022, 1988.

[59] R. C. Littell, G. A. Milliken, W. W. Stroup, and R. D. Wolfinger. *SAS system for mixed models.* SAS Institute, Inc., Cary, NC, 1996.

[60] R. J. A. Little. Pattern-Mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421) :125–134, 1993.

[61] R. J. A. Little. A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81(3) :471–483, 1994.

[62] R. J. A. Little. Modeling the Drop-Out mechanism in Repeated-Measures studies. *Journal of the American Statistical Association*, 90(431) :1112–1121, 1995.

[63] R. J. A. Little and D. B. Rubin. *Statistical analysis with missing data.* Wiley, 2002.

[64] F. M. Lord and M. R. Novick. *Statistical Theories of Mental Test Scores*. Addison-Wesley Publishing Company, Inc., June 1968.

[65] P. Macdonald and S. V. Paunonen. A monte carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*, 62(6) :921 –943, Dec. 2002.

[66] G. N. Masters. A rasch model for partial credit scoring. *Psychometrika*, 47(2) :149–174, 1982.

[67] C. A. McHorney, J. E. Ware, and A. E. Raczek. The MOS 36-Item Short-Form health survey (SF-36) : II. psychometric and clinical tests of validity in measuring physical and mental health constructs. *Medical Care*, 31(3) :247–263, 1993.

[68] T. Meiser. Rasch models for longitudinal data. In M. von Davier and C. H. Carstensen, editors, *Multivariate and Mixture Distribution Rasch Models*, Statistics for Social and Behavioral Sciences, pages 191–199. New York, springer edition, 2007.

[69] B. Michiels, G. Molenberghs, L. Bijnens, T. Vangeneugden, and H. Thijs. Selection models and pattern-mixture models to analyse longitudinal quality of life data subject to drop-out. *Statistics in Medicine*, 21(8) :1023–41, 2002.

[70] P. Minini and M. Chavance. Sensitivity analysis of longitudinal normal data with drop-outs. *Statistics in Medicine*, 23(7) :1039–1054, 2004.

[71] R. J. Mislevy, A. E. Beaton, B. Kaplan, and K. M. Sheehan. Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29(2) :133–161, 1992.

[72] R. J. Mislevy, E. G. Johnson, and E. Muraki. Scaling procedures in NAEP. *Journal of Educational Statistics*, 17(2) :131–154, 1992.

[73] I. W. Molenaar. Estimation of item parameter. In G. H. Fischer and I. W. Molenaar, editors, *Rasch models : foundations, recent developments, and applications*. Springer, Apr. 1995.

[74] G. Molenberghs, C. Beunckens, C. Sotto, and M. G. Kenward. Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 70(2) :371–388, 2008.

[75] G. Molenberghs and M. G. Kenward. A perspective on simple methods. In *Missing data in clinical studies*, pages 41–54. John Wiley and Sons, 2007.

[76] G. Molenberghs, M. G. Kenward, and E. Lesaffre. The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika*, 84(1) :33–44, 1997.

[77] G. Molenberghs, B. Michiels, M. G. Kenward, and P. J. Diggle. Monotone missing data and pattern-mixture models. *Statistica Neerlandica*, 52(2) :153–161, 1998.

[78] C. Monseur and R. Adams. Plausible values : how to deal with their limitations. *Journal of Applied Measurement*, 10(3) :320–334, 2009.

[79] E. Muraki. A generalized partial credit model : Application of an EM algorithm. *Applied Psychological Measurement*, 16(2) :159 –176, June 1992.

[80] J. M. Norquist, R. Fitzpatrick, J. Dawson, and C. Jenkinson. Comparing alternative rasch-based methods vs raw scores in measuring change in health. *Medical Care*, 42(1 Suppl) :I25–36, Jan. 2004.

[81] J. C. Nunnally. *Psychometric Theory*. Mcgraw-Hill College, 2nd edition, 1978.

[82] W. H. O. Q. of Life Assessment Group. What quality of life ? the WHOQOL group. *World Health Forum*, 17(4) :354–356, 1996.

[83] H. G. Osburn. Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*, 5(3) :343–355, 2000.

[84] G. R. Parkerson, W. E. Broadhead, and C. K. Tse. The duke health profile. a 17-item measure of health and dysfunction. *Medical Care*, 28(11) :1056–1072, 1990.

[85] H. D. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3) :545–554, 1971.

[86] J. Pfanzagl. On item parameter estimation in certain latent trait models. In G. H. Fischer and D. Laming, editors, *Contributions to mathematical psychology, psychometrics, and methodology*, pages 249–263. Springer, New York, 1994.

[87] L. Prieto, J. Alonso, and R. Lamarca. Classical test theory versus rasch analysis for quality of life questionnaire reduction. 1 :27, 2003.

[88] S. Rabe-Hesketh, A. Skrondal, and A. Pickles. GLLAMM manual. *U.C. Berkeley Division of Biostatistics Working Paper Series*, (Working Paper 160), 2004.

[89] G. Rasch. On specific objectivity : An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy*, 14 :58–94, 1977.

[90] G. Rasch. *Probabilistic models for some intelligence and attainment tests.* University of Chicago Press, 1980.

[91] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3) :581–592, 1976.

[92] D. B. Rubin. *Multiple imputation for nonresponse in surveys.* Wiley-IEEE, 2004.

[93] F. Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, (17), 1969.

[94] G. Saporta. *Probabilités, analyses des données et statistiques.* Editions Technip, Paris, 1990.

[95] F. Satterthwaite. Synthesis of variance. *Psychometrika*, 6 :309–316, 1941.

[96] C. E. Schwartz and M. A. Sprangers. Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Social Science & Medicine (1982)*, 48(11) :1531–1548, June 1999. PMID : 10400255.

[97] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2) :461–464, 1978.

[98] K. Sijtsma and B. T. Hemker. A taxonomy of IRT models for ordering persons and items using simple sum scores. *Journal of Educational and Behavioral Statistics*, 25(4) :391–415, 2000.

[99] J. A. Sloan, M. Y. Halyard, M. H. Frost, A. C. Dueck, B. Teschendorf, M. L. Rothman, and the Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group. The mayo clinic manuscript series relative to the discussion, dissemination, and operationalization of the food and drug administration guidance on Patient-Reported outcomes. *Value in Health*, 10 :S59–S63, 2007.

[100] A. C. Smidt, J. Lai, D. Cella, S. Patel, A. J. Mancini, and S. L. Chamlin. Development and validation of Skindex-Teen, a quality-of-life instrument for adolescents with skin disease. *Archives of Dermatology*, 146(8) :865–869, Aug. 2010.

[101] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1) :72–101, 1904.

[102] C. Spearman. Demonstration of formulae for true measurement of correlation. *The American Journal of Psychology*, 18(2) :161–169, Apr. 1907.

[103] M. A. Sprangers and C. E. Schwartz. Integrating response shift into health-related quality of life research : a theoretical model. *Social Science & Medicine (1982)*, 48(11) :1507–1515, June 1999. PMID : 10400253.

[104] M. A. G. Sprangers. Response-shift bias : a challenge to the assessment of patients' quality of life in cancer clinical trials. *Cancer Treatment Reviews*, 22(Supplement 1) :55–62, Jan. 1996.

[105] S. Stevens. Mathematics, measurement and psychophysics. In *Handbook of Experimental Psycholgy*. New York, 1951.

[106] S. S. Stevens. On the theory of scales of measurement. *Science*, 103(2684) :677–680, 1946.

[107] The WHOQOL Group. Development of the world health organization WHOQOL-BREF quality of life assessment. *Psychological Medicine*, 28(3) :551–558, 1998.

[108] N. Thomas. Assessing model sensitivity of the imputation methods used in the national assessment of educational progress. *Journal of Educational and Behavioral Statistics*, 25(4) :351–371, 2000.

[109] L. Thurstone. *The Reliability and Validity of Tests*. Edwards Brothers, Ann Arbor, MI, 1931.

[110] A. B. Troxel, D. L. Fairclough, D. Curran, and E. A. Hahn. Statistical analysis of quality of life with missing data in cancer clinical trials. *Statistics in Medicine*, 17(5-7) :653–666, 1998.

[111] R. Turner and R. J. Adams. The programme for international student assessment : an overview. *Journal of Applied Measurement*, 8(3) :237–248, 2007.

[112] U.S. Department of Health and Human Services FDA Center for Drug Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Biologics

Evaluation and Research, and U.S. Department of Health and Human Services FDA Center for Devices and Radiological Health. Guidance for industry : patient-reported outcome measures : use in medical product development to support labeling claims : draft guidance. *Health and Quality of Life Outcomes*, 4 :79–79, 2006.

[113] G. Verbeke and E. Lesaffre. The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis*, 23(4) :541–556, 1997.

[114] G. Verbeke and G. Molenberghs. Estimation of the marginal model. In *Linear Mixed Models for Longitudinal Data*, Springer Series in Statistics, pages 41–54. Springer, New York, 2000.

[115] G. Verbeke and G. Molenberghs. Inference for the marginal model. In *Linear Mixed Models for Longitudinal Data*, Springer Series in Statistics, pages 55–76. Springer, New York, 2000.

[116] G. Verbeke and G. Molenberghs. Joint modeling of measurements and missingness. In *Linear Mixed Models for Longitudinal Data*, Springer Series in Statistics, pages 209–219. Springer, New York, 2000.

[117] G. Verbeke and G. Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer, Berlin, 2000.

[118] G. Verbeke and G. Molenberghs. Simple missing data methods. In *Linear Mixed Models for Longitudinal Data*, Springer Series in Statistics, pages 221–229. Springer New York, 2000.

[119] G. Verbeke, G. Molenberghs, H. Thijs, E. Lesaffre, and M. G. Kenward. Sensitivity analysis for nonrandom dropout : A local influence approach. *Biometrics*, 57(1) :7–14, 2001.

[120] H. Wainer and D. Thissen. Estimating ability with the wrong model. *Journal of Educational Statistics*, 12(4) :339–368, 1987.

[121] T. Warm. Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54 :427–450, 1989.

[122] B. D. Wright and G. N. Masters. *Rating Scale Analysis*. MESA Press, Chicago, 1 edition, Jan. 1982.

[123] M. Wu. The role of plausible values in Large-Scale surveys. *Studies in Educational Evaluation*, 31(2) :114–128, 2005.

[124] X. Zheng and S. Rabe-Hesketh. Estimating parameters of dichotomous and ordinal item response models with gllamm. *Stata Journal*, 7(3) :313–333, 2007.

[125] A. H. Zwinderman. Pairwise parameter estimation in rasch models. *Applied Psychological Measurement*, 19(4) :369–375, 1995.

Comparaison des approches CTT et IRT pour l'analyse des effets temps et groupe de données longitudinales de type Patient-Reported Outcomes et impact du dropout

L'évaluation des traitements des maladies chroniques a longtemps été basée uniquement sur la progression de la maladie et la survie. Avec l'amélioration des traitements et l'allongement de la durée de vie, la question de l'impact de la maladie et des traitements sur la qualité de vie s'est posée. Les Patient-Reported Outcomes (PRO), évaluant des concepts tels que la qualité de vie liée à la santé, sont de plus en plus utilisés. Deux approches existent pour l'analyse de PRO : la théorie classique des test (CTT) et la théorie de réponse aux items (IRT). La validation de questionnaires s'appuie aussi bien sur la CTT que l'IRT mais la CTT semble plus souvent utilisée en phase d'analyse. L'intérêt se porte souvent sur l'étude de l'évolution d'un PRO au cours du temps. La méthode d'analyse doit alors être adaptée à l'analyse de données corrélées dans le temps. Les données longitudinales sont fréquemment sujettes à du dropout, informatif ou non, qui peut avoir un impact sur les résultats de l'analyse.
Ce travail vise à déterminer la méthode la plus adéquate pour analyser des PRO recueillis de manière longitudinale et issus d'un questionnaire validé à la fois en CTT et en IRT. Différentes méthodes d'analyse, basées sur la CTT ou l'IRT, ont été comparées à travers des études de simulation. L'impact du dropout, informatif ou non, a également été étudié. La comparaison était basée sur le risque α, la puissance et le biais des estimations des effets temps et groupe.
Les deux approches présentent des résultats comparables pour des données complètes ou sujettes à du dropout ignorable et sont valides dans ce contexte. Elles ne sont pas valides en cas de dropout informatif.
Mots-clés : Patient-Reported Outcomes, Classical Test Theory, Item Response Theory, modèle de Rasch, données longitudinales, dropout

Comparison of CTT and IRT approaches for joint analysis of group and time effects of longitudinal Patient-Reported Outcomes and impact of dropout

For a long time, treatments for chronic diseases were evaluated solely on disease progression and survival. As treatments improved and survival increased, the influence of the disease and the treatments on patients quality of life became of interest. Patient-Reported Outcomes (PRO), assessed through questionnaires to evaluate concepts, such as health-related quality of life, are now widely used. Two approaches exist to handle such data, the Classical Test Theory (CTT) and the Item Response Theory (IRT). The evaluation of the evolution of PRO is often the major concern of the study and this raises the point of the choice of a method of analysis adapted to correlated data. Longitudinal studies are frequently faced with potentially informative dropout that can have an impact on the results of the analysis. CTT and IRT are often used for development and validation of questionnaires but CTT remains the most frequently used method at the analysis stage.
This work aims at identifying the most adequate approach between CTT and IRT to analyze the data when the questionnaire used to collect PRO data has been validated with a CTT and an IRT model. Different methods of analysis, based either on CTT or IRT, were compared in the context of longitudinal PRO data, potentially subject to dropout, through simulation studies. The comparison was made in terms of type I error, power and bias for the estimation of time and group effects.
Both approaches presented comparable results on complete data and data subject to ignorable dropout and methods are valid in such cases. However, both approaches are not adequate for the analysis of data subject to informative dropout.
Keywords: Patient-Reported Outcomes, Classical Test Theory, Item Response Theory, Rasch model, longitudinal data, dropout

BLANCHIN Myriam
6 Cours Sully
44000 NANTES