

**Arielle Bonneville-Roussy
Fabien Fenouillet
Yannick Morvan**

Introduction aux analyses par équations structurelles

**Applications avec Mplus
en psychologie et sciences sociales**

DUNOD

Cet ouvrage est un retraitage corrigé de l'édition parue en septembre 2021 sous le même titre.

Maquette intérieure :
www.atelier-du-livre.fr
(Caroline Joubert)

<p>Le pictogramme qui figure ci-contre mérite une explication. Son objet est d'alerter le lecteur sur la menace que représente pour l'avenir de l'écrit, particulièrement dans le domaine de l'édition technique et universitaire, le développement massif du photocopillage.</p> <p>Le Code de la propriété intellectuelle du 1^{er} juillet 1992 interdit en effet expressément la photocopie à usage collectif sans autorisation des ayants droit. Or, cette pratique s'est généralisée dans les établissements</p>	<p>d'enseignement supérieur, provoquant une baisse brutale des achats de livres et de revues, au point que la possibilité même pour les auteurs de créer des œuvres nouvelles et de les faire éditer correctement est aujourd'hui menacée.</p> <p>Nous rappelons donc que toute reproduction, partielle ou totale, de la présente publication est interdite sans autorisation de l'auteur, de son éditeur ou du Centre français d'exploitation du droit de copie (CFC, 20, rue des Grands-Augustins, 75006 Paris).</p>
	

© Dunod, 2022

11 rue Paul Bert - 92240 Malakoff
ISBN 978-2-10-083891-2

Le Code de la propriété intellectuelle n'autorisant, aux termes de l'article L. 122-5, 2^o et 3^o a), d'une part, que les « copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective » et, d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, « toute représentation ou reproduction intégrale ou partielle faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause est illicite » (art. L. 122-4).

Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait donc une contrefaçon sanctionnée par les articles L. 335-2 et suivants du Code de la propriété intellectuelle.

Table des matières

<i>Introduction</i>	7
1. Pourquoi un livre sur les modèles d'équations structurelles?.....	9
2. Pourquoi Mplus?.....	10
3. Un livre sans (ou presque sans) équations.....	10
4. La structure du livre.....	11
<i>Remerciements</i>	13

Partie 1 *Concepts de base*

CHAPITRE 1 – DÉFINIR LA MODÉLISATION PAR ÉQUATIONS STRUCTURELLES.....	17
1. Que sont les modèles par équations structurelles?.....	19
2. Concepts liés aux MES.....	21
3. Comment commencer? Les grandes étapes des MES.....	26
CHAPITRE 2 – L'ENVIRONNEMENT MPLUS.....	35
1. Fichier de données.....	37
2. Fichier script (aussi appelé Syntaxe, dans Mplus Input, .inp).....	39
3. Fichier output (Résultats).....	47
4. Autres fonctionnalités utiles.....	49

Partie 2 *Modèles de base*

CHAPITRE 3 – ANALYSES PAR PISTES CAUSALES AVEC VARIABLES OBSERVÉES....	53
1. Comment modéliser une régression simple?.....	56
2. Comment interpréter les résultats?.....	64
3. Comment modéliser une analyse en pistes causales?.....	65
4. Comment interpréter les résultats?.....	73
5. Comment modéliser une analyse de médiation avec variables observées?.....	76
CHAPITRE 4 – LES MODÈLES DE MESURE: ANALYSE FACTORIELLE.....	85
1. Introduction: principe de l'analyse factorielle.....	87
2. Analyse factorielle confirmatoire.....	88

3. Respécifier le modèle (étape 5).....	100
4. Un entre-deux: modèles par équations structurelles exploratoires	103
5. Les modèles bifactoriels	109
CHAPITRE 5 – L’INVARIANCE DES MESURES ET L’ÉQUIVALENCE DES MODÈLES ENTRE DIFFÉRENTS GROUPES.....	
1. Principe de l’invariance	125
2. Les principales étapes de l’invariance.....	126
3. Invariance de la mesure: la méthode simple.....	131
4. Invariance de la mesure: deuxième approche.....	140
5. Comparer les moyennes latentes.....	155
CHAPITRE 6 – MODÉLISATION PAR ÉQUATIONS STRUCTURELLES AVEC VARIABLES LATENTES	
1. Combiner les modèles de régression et les modèles de mesure	163
2. Mettre en œuvre et tester un modèle par équations structurelles.....	163
3. MES entre différents groupes: la mesure d’équivalence.....	172

Partie 3
Modèles longitudinaux

CHAPITRE 7 – ANALYSE LONGITUDINALE EN COURBES LATENTES LGCM.....	191
1. Exemple d’un protocole longitudinal	193
2. Comment modéliser les trajectoires?	196
3. Comment modéliser une trajectoire linéaire?.....	196
4. Adéquation des modèles à nos données.....	207
5. Comment modéliser des trajectoires non linéaires?.....	212
6. Les covariables qui varient et ne varient pas au cours du temps.....	223
7. Comment modéliser l’effet d’une trajectoire sur une variable dépendante distale	234
8. En résumé.....	237

Partie 4
Modèles multi-niveaux (hiérarchiques)

CHAPITRE 8 – ANALYSES FACTORIELLES CONFIRMATOIRES MULTI-NIVEAUX.....	241
1. Introduction aux analyses multi-niveaux.....	243

2. L'importance d'analyser des données hiérarchiques à l'aide d'analyses multi-niveaux.....	244
3. Notation graphique.....	245
4. Spécifier un modèle multi-niveau.....	245
5. Identification et estimation.....	249
6. Les analyses factorielles multi-niveaux.....	250
 CHAPITRE 9 – ANALYSES PAR PISTES CAUSALES MULTI-NIVEAUX.....	 271
1. La régression multi-niveaux.....	273
2. Les analyses par pistes causales multi-niveaux.....	291
 CHAPITRE 10 – LES MODÈLES PAR ÉQUATIONS STRUCTURELLES MULTI-NIVEAUX.....	 311

Partie 5
S'y retrouver dans les MES

CHAPITRE 11 – SUGGESTIONS DE SOLUTIONS À DES PROBLÈMES COURANTS. CE CHAPITRE EST PEUT-ÊTRE CELUI QUE VOUS DÉSIREREZ LIRE EN PREMIER.....	343
1. Glossaire des termes anglais-français utilisés en MES.....	345
2. Glossaire anglais-français des commandes Mplus.....	348
3. Un problème fréquent: la taille d'échantillon.....	349
4. Les vingt sources d'erreurs fréquentes et leurs messages.....	358
 <i>Épilogue</i>	 369
 <i>Bibliographie</i>	 371
 <i>Index</i>	 375



Introduction

1. Pourquoi un livre sur les modèles d'équations structurelles ?

Arielle : L'idée de ce livre est venue en 2014 lorsque j'enseignais les modèles d'équations structurelles (MES) à l'université de Cambridge, en Angleterre. Le cours que je donnais était accessible aux chercheurs et étudiants des cycles supérieurs du monde entier et se voulait une introduction aux MES par le biais d'exemples concrets proches des besoins encourus dans les universités et grandes écoles. C'est en enseignant ce cours en anglais et en rencontrant des chercheurs francophones qui venaient y assister que je me suis rendu compte du besoin criant de démocratiser les MES par des ressources en français. J'y ai fait la rencontre de Yannick Morvan. Puis quelques mois plus tard, Fabien Fenouillet et moi avons présenté dans le même colloque et avons découvert cet intérêt commun aux MES.

Yannick : Après m'être formé aux MES avec Mplus à Londres avec Chris Stride en 2014 puis à Cambridge à deux reprises en 2014 et 2015 auprès d'Arielle Bonneville-Roussy et de Gabriela Roman, j'étais convaincu de l'intérêt de transmettre ces techniques de modélisation auprès d'un public francophone. Avec l'aide de ma collègue Laurence Kern, nous avons organisé une formation à l'université Paris-Nanterre en juillet 2016 et eu le plaisir d'y accueillir Arielle Bonneville-Roussy comme formatrice. Nous nous y sommes retrouvés avec Fabien Fenouillet qui était déjà un participant avancé et également très intéressé pour diffuser ces techniques.

Fabien : Mon intérêt pour les MES a émergé aux alentours des années 2010. Après m'être arraché les cheveux sur nombre d'ouvrages et d'articles en anglais, j'ai eu le plaisir de rencontrer le regretté Pierre Vrignaud qui en était l'un des premiers spécialistes français. Quelques années plus tard, à l'occasion de différents colloques au Canada sur mes thématiques de recherches, j'ai croisé la route d'Arielle Bonneville-Roussy que j'ai eu l'agréable surprise de retrouver la même année mais cette fois comme formatrice sur invitation de Yannick Morvan et Laurence Kern. Il est vite apparu que nous partagions le même goût pour la compréhension de l'immense famille des MES et qu'il aurait été dommage de s'arrêter en si bon chemin.

C'est en discutant tous les trois que nous nous sommes rendu compte du besoin criant d'un ouvrage savant traitant des MES de la manière la plus simple possible. En effet, les MES sont devenues l'une des méthodes les plus utilisées en psychologie et en sciences sociales. Cependant, la majorité des ouvrages traitant des MES ont été écrits en anglais, souvent en utilisant un jargon opaque peu accessible à un public francophone. De plus, les ouvrages existants utilisent fréquemment des équations statistiques comme outil de base de la compréhension, ce qui rend d'autant plus ardue la découverte des MES pour les chercheurs et étudiants en psychologie et sciences sociales. Ce livre se veut donc une réponse à ces deux problématiques.

Le but de cet ouvrage est d'offrir une introduction aux MES qui est à la fois accessible et complète. Son objectif principal est de fournir un guide pratique des MES à l'aide de Mplus. En ce sens, il est écrit pour les chercheurs qui connaissent peu ou qui ne connaissent pas les MES et le logiciel Mplus. En fait, le seul prérequis de ce livre est d'avoir une connaissance de base de la corrélation et de la régression. Il cible tant les étudiants aux cycles supérieurs et les chercheurs postdoctoraux que les chercheurs établis. Ce livre ne cible pas de domaine de recherche spécifique, si ce n'est que les exemples sont tirés majoritairement des domaines de la psychologie, des sciences sociales et de l'éducation. En somme, nous espérons que les chercheurs en psychologie et en sciences sociales débutants puissent acquérir par les premiers chapitres de cet ouvrage une connaissance de base des techniques liées aux MES, et qu'ils puissent les appliquer dans leurs propres recherches. Nous espérons aussi que les chercheurs plus aguerris puissent y découvrir des techniques plus avancées qui sont introduites dans les derniers chapitres du livre.

2. Pourquoi Mplus ?

Plusieurs logiciels sont aptes à faire de la modélisation par équations structurelles. Mentionnons ici les logiciels AMOS, LISREL, R et bien entendu Mplus. Cependant, Mplus est un logiciel relativement facile à utiliser qui peut être considéré comme le « couteau suisse » des MES. Non seulement peut-il analyser des données continues, mais aussi catégorielles de toutes sortes. Il traite aussi les données manquantes et nous permet de combiner des analyses telles que les régressions logistiques et linéaires. Avec Mplus le chercheur peut analyser des données tant corrélationnelles que transversales, longitudinales et multi-niveaux. Enfin, Mplus nous permet de voir graphiquement les résultats des analyses par MES. Il est à noter que tous les scripts ont été testés à l'aide de la version Mplus 8.5 (Mplus 2017-2021).

Bien que ce livre traite des MES avec Mplus, le lecteur intéressé à avoir une introduction plus générale aux MES pourra facilement s'y retrouver puisque chaque chapitre commence par une introduction aux concepts de base des analyses préconisées dans ce chapitre.

3. Un livre sans (ou presque sans) équations

Le présent ouvrage est pensé comme une introduction à l'analyse et l'interprétation de données avec les MES et Mplus. L'accent est mis sur la compréhension

des concepts de base. Il ne s'agit pas d'un livre de statistiques. L'ouvrage se veut comme un guide pratique, l'attention est portée sur l'analyse et l'interprétation à l'aide d'applications pratiques et d'exemples concrets. En ce sens chacun des chapitres commence par l'énoncé d'un problème de recherche concret, tiré bien souvent des recherches antérieures des auteurs, suivi de l'énoncé d'hypothèses qui seront évaluées à l'aide de la technique préconisée dans le chapitre. Chacun des problèmes de recherche sera aussi illustré par des schémas et graphiques propres aux MES. De plus, les scripts et paramètres d'entrée (fichiers scripts, INPUT) Mplus seront précisés et les fichiers de sortie (OUTPUT) interprétés en détail. Enfin, une interprétation globale des résultats sera donnée.

Afin de préserver l'aspect pratique du livre, les équations statistiques sont donc réduites au minimum. Les lecteurs plus friands de formules mathématiques et d'équations statistiques élaborées ne seront pas en reste puisque chaque chapitre inclut des références à des ouvrages plus théoriques où les concepts mathématiques sous-jacents sont expliqués en profondeur.

4. La structure du livre

Le livre comprend cinq parties. La première partie introduit les concepts de base du livre. Le chapitre 1 s'attarde sur les concepts de base des MES, ainsi que leurs notations standardisées et une proposition de marche à suivre pour commencer l'analyse des MES. Le chapitre 2 s'attarde à l'environnement Mplus et introduit aux lecteurs francophones la définition des fichiers, des syntaxes de base et des fonctionnalités de Mplus.

La deuxième partie permettra au chercheur de se familiariser avec les techniques de base des MES ainsi que de leurs applications dans Mplus. Ainsi, nous y introduirons les analyses par piste causale (chapitre 3), les modèles de mesures (chapitres 4 et 5) ainsi que les modèles d'équations structurelles proprement dites (chapitre 6).

La troisième partie s'attarde aux modèles longitudinaux avec des analyses de trajectoires (chapitre 7). La quatrième partie offre un aperçu des analyses MES multi-niveaux (8, 9 et 10). Finalement, la cinquième partie offre des solutions et astuces de dépannages pour vous aider en cas de problèmes avec vos analyses (chapitre 11). À l'heure où nous finalisons la rédaction de cet ouvrage, tous les fichiers Mplus ont été vérifiés à l'aide de la version Mplus 8.6. Ces fichiers sont disponibles sur notre site web compagnon : mesmodelisations.fr.

Remerciements

Le parcours de la conception à la parution de ce livre a été marqué par plusieurs événements de vie, notamment une pandémie et trois naissances chez les auteurs. Nous tenons donc à remercier en premier lieu nos familles pour leur soutien infail-
lible, Florence, Guilhem, Camille, Constance, Cécile, Julie, Jeanne...

Nous voulons aussi remercier le Centre de Psychométrie de l'Université de Cambridge, son directeur émérite le Professeur John Rust ainsi que Dr. Gabriela Roman (co-directrice, avec Arielle Bonneville-Roussy, du cours de modélisation par équations structurelles avec Mplus, donné à Cambridge de 2013 à 2016). Sans le soutien du Centre et la complicité et le soutien de Gabriela, le cours, et par extension l'idée de ce livre, n'auraient jamais vu le jour. Merci aussi à l'Université de Roehampton où la majorité du livre a été écrite.

Nous souhaitons également remercier l'Institut de Psychiatrie ainsi que l'université Paris Nanterre et plus particulièrement le Dr. Laurence Kern pour nous avoir permis d'organiser en 2016 une première formation francophone nationale sur Mplus qui a constitué une des sources de motivation de cet ouvrage, ainsi que les Pr. Krebs et Dr. Chaumette du Centre d'Évaluation des Jeunes Adultes et Adolescents dont l'activité clinique et de recherche a inspiré certaines des modélisations présentées dans cet ouvrage.

Un grand merci à nos relecteurs qui ont fait un travail formidable en nous disant que notre travail en valait la chandelle et surtout en nous aidant à clarifier les passages difficiles : Pr. Jérémie Verner-Filion ainsi que les doctorantes Éliisa Grenon et Éliizabeth Parent-Taillon.

Finalement, un grand merci à l'équipe des éditions Dunod qui, compte-tenu du contexte sanitaire, a fait preuve d'une grande patience et résilience dans l'attente du résultat final.

Partie 1

Concepts de base

Chapitre 1

**Définir la modélisation
par équations structurelles**



Sommaire

1. Que sont les modèles par équations structurelles?.....	19
2. Concepts liés aux MES.....	21
3. Comment commencer ? Les grandes étapes des MES.....	26

1. Que sont les modèles par équations structurelles ?

La modélisation par équations structurelles (MES) est une technique d'analyse qui combine simultanément les modèles de mesure (par exemple, les analyses factorielles) et de prédiction (par exemple, les régressions). Elles servent à confirmer des théories et hypothèses préexistantes. Les théories engendrant des MES tendent à stipuler des liens causaux entre une ou plusieurs variables. Les MES peuvent combiner en une seule analyse l'évaluation des mesures utilisées ainsi que leurs liens causaux. De plus, dans les MES, les erreurs de mesure sont modélisées ce qui permet de réduire les biais psychométriques.

Revenons au terme « modèles par équations structurelles » ou « modélisation par équations structurelles » pour mieux en comprendre sa définition. Le terme « modélisation » se réfère au fait que, dans les MES, le chercheur veuille confirmer, ou infirmer, un modèle dans son entièreté. Un *modèle* est une représentation schématique d'un processus. Ainsi, dans les MES, les schémas graphiques sont très importants. Le terme « modèles » est suivi d'« équations », donc à plusieurs équations estimées simultanément. Les MES sont des analyses multivariées, c'est-à-dire avec plusieurs variables évaluées collectivement. Finalement, le terme « structurelles » se réfère à l'aspect prédictif des modèles. C'est-à-dire le modèle à confirmer, ou à rejeter, sera calculé avec une *structure* précise où il sera postulé a priori qu'une variable en prédira, ou en mesurera, une autre. Ce calcul se fera à l'aide de plusieurs équations évaluées simultanément.

Cinq caractéristiques principales définissent les MES. La première caractéristique est que les MES ont comme but principal de confirmer un modèle théorique existant, qui spécifie des relations structurelles entre les variables. En soi, les MES constituent aussi une approche hiérarchique descendante de l'analyse de données (c'est-à-dire qu'ils partent de théories et d'hypothèses à tester), et sont donc, dans la plupart des cas, confirmatoires. En MES, le modèle doit être guidé par une ou plusieurs théories et des résultats empiriques issus de recherches antérieures. Un des buts principaux des MES est donc d'évaluer la concordance entre le modèle théorique et le modèle obtenu avec les données empiriques.

La deuxième caractéristique est que les MES ont la particularité de modéliser les erreurs de mesure, ce qui veut dire que les résultats des MES contiennent un biais dû à l'erreur qui est minimisé (Wang & Wang, 2019).

La troisième caractéristique est la combinaison des *modèles de mesure* et *modèles de structure* dans la même analyse. Les *modèles de mesure* sont primordiaux en sciences sociales et en psychologie parce que la plupart des phénomènes entourant les personnes ne peuvent pas être mesurés directement. Par exemple, l'intelligence, l'extraversion et le rendement au travail des personnes sont difficiles à mesurer

directement à l'aide d'indicateurs simples. Ces phénomènes ne peuvent être mesurés qu'indirectement. Les chercheurs ont donc recours à des mesures plus ou moins objectives, tels que les tests de quotient intellectuel (QI), de personnalité ou de performance afin d'identifier ces phénomènes. Ces mesures contiennent plusieurs indicateurs qui ont besoin d'être évalués pour leur qualité : leur validité et leur fiabilité, tout en tenant compte du contexte dans lequel le phénomène est observé. Ce sont ces qualités (surtout la validité) que les modèles de mesure évaluent. En ce sens, les modèles de mesure des MES font l'agrégation de plusieurs indicateurs qui sont présumés mesurer un phénomène commun. Ceci est différent de bien des sciences naturelles où les mesures telles que la longueur ou le poids peuvent souvent être mesurées directement à l'aide d'indicateurs simples et objectifs (le centimètre et le gramme, par exemple). Ces mesures objectives ne sont pas présumées changer en fonction du contexte, et donc, dans la plupart des cas, les chercheurs n'ont pas besoin d'établir leur validité ou leur fiabilité une fois que la mesure est établie. Les *modèles de structure*, quant à eux, définissent les liens causaux entre les variables. C'est-à-dire que les MES assument une relation de cause à effet entre une variable et une autre (une variable en prédit une autre). Les termes prédiction et causalité en MES sont utilisés afin d'identifier précisément ces liens prédictifs présumés. Il ne présuppose pas une causalité observée, que seuls les devis expérimentaux stricts peuvent évaluer.

La quatrième caractéristique est que les modèles sont évalués statistiquement dans leur intégralité. Ce n'est que lorsque les modèles de mesure et de structure sont combinés que nous obtenons des *modèles par équations structurelles* proprement dits. Le modèle statistique est donc testé dans une seule analyse qui combine les modèles de mesure et de structure à l'aide de techniques statistiques qui leur sont propres. Dans sa forme la plus simple, un MES combine l'analyse factorielle (comme composante de mesure) et la régression (comme composante de structure). Le modèle statistique combiné est comparé au modèle théorique à l'aide de paramètres que l'on nomme *indices d'adéquation* (ou d'ajustement ; en anglais : *model fit indices*). Si les indices d'adéquation sont convenables, le modèle est accepté ; sinon, il est rejeté.

Finalement, les MES peuvent généralement être représentés par des graphiques. Ces graphiques sont informés par le modèle théorique et les résultats de l'analyse. Les MES sont illustrés schématiquement par des symboles reconnus. Nous reviendrons à ces symboles plus tard dans le chapitre.

2. Concepts liés aux MES

2.1 Les variables observées et latentes

De nombreuses variables dans la recherche sont observées et observables. Par exemple, un chercheur peut être intéressé par des variables observées telles que l'âge et la taille en centimètre. En sciences sociales et en psychologie, d'autres variables peuvent aussi être observées, telles que, en prenant l'exemple d'un test d'intelligence, le temps passé sur une question d'un test et sa réussite ou son échec, chacune des réponses au test, ou une observation spécifique liée au test. Cependant, nous ne pouvons pas observer ou mesurer directement l'intelligence. La variable « intelligence » est un concept abstrait qui, pour être mesuré, doit être défini conceptuellement, puis mesuré à l'aide de plusieurs indicateurs observables, tels que le temps de réponse mesuré en millisecondes, le nombre de réponses correctes aux énoncés d'un questionnaire d'intelligence, ou l'observation de plusieurs tâches cognitives. La variable mesurée (par exemple la réponse à un énoncé d'un questionnaire) est appelée variable observée, ou variable manifeste, et la variable abstraite (l'intelligence elle-même) est appelée variable latente. En d'autres mots, une variable latente est une variable non mesurable qui est la synthèse de plusieurs variables observées. Cette variable latente est fonction de plus d'une mesure observée et tente d'expliquer les corrélations entre ces mesures. Ces variables observées sont, à leur tour, supposées être la manifestation de ce construit latent.

En MES, les *modèles de mesure* visent à identifier, à l'aide de plusieurs variables observées, une ou plusieurs variables latentes (les fameux *facteurs*). Les *modèles de structure*, quant à eux, visent à déterminer les liens causaux (le plus souvent de régression) entre plusieurs variables, qu'elles soient observées ou latentes.

2.2 Variables indépendantes et dépendantes, exogènes et endogènes

Les variables indépendantes (VI) sont les caractéristiques qui prédisent les variables d'intérêt. Les VI sont aussi appelées *variables explicatives* ou *variables exogènes* en MES. Elles sont à l'origine des fluctuations des valeurs des autres variables du modèle. Les variables dépendantes (VD), quant à elles, sont les résultats d'intérêt qui sont directement ou indirectement influencés par les VI. Les VD se nomment aussi *variables à expliquer* ou *variables endogènes* en MES. Les VI et VD peuvent être soit observées ou latentes. Il est à noter qu'en MES, les termes *variables exogènes* et *variables endogènes* sont plus souvent utilisés. Une *variable exogène* ne

dépend pas d'aucune autre variable alors qu'une *variable endogène* est une variable qui dépend d'une autre, mais cette variable peut être soit une variable à expliquer proprement dite, soit une variable à expliquer et explicative à la fois, notamment dans le cas des variables médiatrices (voir les chapitres 3 et 9). Nous utiliserons les termes VI et VD, ou variables exogènes et endogènes, lorsqu'approprié.





2.3 Les paramètres fixes, libres et contraints

En MES, un paramètre est un élément mesurable, ou estimable, tels que les termes de régression, de moyennes, et d'erreurs, qui sont définies par une valeur. Les paramètres fixes correspondent aux valeurs qui ne sont pas estimées dans le modèle et, typiquement, sont fixées à zéro (0) ou un (1). Ces valeurs sont soit fixées par défaut (comme, par exemple, la valeur du premier indicateur qui est fixée à 1 par défaut dans une analyse factorielle confirmatoire) ou par le chercheur. Les paramètres libres ont des valeurs inconnues et sont donc ceux à estimer dans le modèle (tous ceux qui ne sont pas fixés). Finalement, certains paramètres peuvent être contraints à avoir une valeur égale. Par exemple, en analyses en groupes multiples, les paramètres des hommes et des femmes (ou autres) peuvent être contraints à être égaux si le contexte théorique nous indique qu'il n'y a pas de différence entre les sexes. Ces paramètres contraints sont aussi libres, c'est-à-dire estimés (mais *contraints* à l'égalité).

2.4 Les notations symboliques

Les MES sont accompagnés d'une notation graphique universellement reconnue dans le domaine. Cette notation constitue la base des schémas graphiques présentés dans tout article scientifique ou livre savant traitant des MES.

En voici les règles de base :

	Le rectangle ou le carré représente la variable observée.
	L'ovale ou le cercle représente la variable latente.
	La flèche unidirectionnelle représente une relation unidirectionnelle.
	La flèche bidirectionnelle représente une relation bidirectionnelle.

La figure 1.1 illustre un exemple d'un schéma représentant les liens entre l'intelligence et le rendement scolaire. Nous y reviendrons dans la prochaine section.

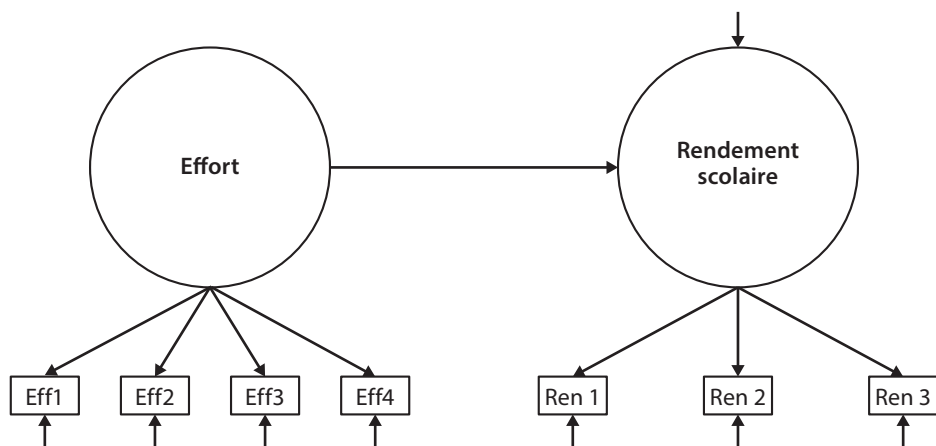


Figure 1.1 – Exemple d'un MES représentant l'effort lors de tâches scolaires qui prédit le rendement scolaire.

2.5 Les équations de base (et parmi les seules du livre)

Cette section est un passage obligé vers les concepts statistiques de base sous-tendant les MES. Les équations statistiques effraient bien des chercheurs, c'est pourquoi nous essaierons de les vulgariser le plus possible. Byrne (2016) résume bien le principe statistique fondamental des MES, qui peut être expliqué par une série d'équations de régression. Par souci de simplicité, les équations présentées ici-bas se veulent des modèles en prose. Le lecteur plus féru de statistiques se référera à Wang et Wang (2019, p. 7-11) pour une explication complète des équations sous-tendant les MES.

Le principe fondamental des MES est d'estimer un modèle comme suit :

$$\boxed{\text{Données} = \text{Modèle} + \text{Erreur}}$$

Équation 1.1 – Concept fondamental.

OÙ

Les DONNÉES sont celles que nous avons récoltées lors de recherches empiriques et constituent des variables observées.

Ces données sont la somme de deux choses :

1. Le MODÈLE hypothétique, c'est-à-dire celui qui a été spécifié lors de l'élaboration du modèle. Ce modèle comprend la plupart du temps à la fois des variables observées et des variables latentes.

2. Un ou plusieurs termes d'ERREUR résiduelle qui correspond à l'écart entre le modèle hypothétique et les données.

En MES, le chercheur définit un modèle hypothétique. Ce modèle est par la suite évalué avec des données empiriques (c'est-à-dire prise de notre collecte de données) qui comprennent toutes les variables observées du modèle hypothétique. Le chercheur compare ensuite l'adéquation (ou ajustement) entre ce modèle hypothétique et les données empiriques. Il arrive rarement que la concordance entre le modèle hypothétique et les données empiriques soit parfaite, d'où l'existence d'une erreur résiduelle.

Dans l'exemple de la figure 1.1, le principe fondamental de l'équation 1.1 cherche à expliquer, à l'aide de données, l'ensemble du modèle explicatif du rendement scolaire et tous les liens inclus dans la figure.

Le **modèle de mesure** des MES peut être simplifié comme ceci :

$$VL = VO + \text{Erreur}$$

Équation 1.2 – Modèle de mesure.

OÙ

Une variable latente (VL) synthétise :

1. Plusieurs variables observées (VO) aussi appelées indicateurs, qui expliquent la variance commune, soit la variance prise en compte par la variable latente.
2. Un terme d'ERREUR de mesure pour chacune des variables observées, c'est-à-dire la variance de chaque VO qui n'est pas prise en compte par la VL. Elle est une combinaison d'erreur systématique spécifique à chaque indicateur, et d'erreur aléatoire.

Une variable latente explique la variation entre un ensemble de mesures observées. Il est important de se rappeler que c'est la variable latente qui explique les variables observées, et non l'inverse.

De manière simplifiée et en prose, les VL Effort et Rendement scolaire de la figure 1.1 sont le résultat des valeurs de leurs VO respectives (quatre pour Effort, nommées Eff1 à Eff4, et trois pour Rendement scolaire, nommées Ren1 à Ren3), et de leurs termes d'erreur (notés à la figure 1.1 par des flèches pointant directement vers les VO). Rappelons, cependant, que le calcul complet est beaucoup plus complexe (voir Kline, 2016, ou Wang et Wang, 2019).

Le concept de base qui sous-tend le **modèle de structure** des MES est celui-ci :

$$VD = VI + \text{Erreur}$$

Équation 1.3 – Modèle de structure.

OÙ

Une variable dépendante (VD) est expliquée par :

1. une ou plusieurs variables indépendantes (VI) ;
2. un terme d'ERREUR résiduelle. Cette erreur résiduelle représente la variance de prédiction de la VD qui n'est pas expliquée par la (ou les) VI.

La partie structurelle des MES peut donc être représentée par une série d'équations. Ces équations représentent l'influence d'une ou plusieurs variables, VI, sur une autre variable, la VD. En tenant compte qu'il est presque impossible qu'une VD soit entièrement expliquée par une ou plusieurs VI, l'équation contient aussi un terme d'erreur qui correspond à la proportion de la VD qui n'est pas expliquée par la, ou les, VI.

Dans le cas de notre exemple de la figure 1.1, le rendement scolaire (VD) est expliqué par l'effort (VI), et une valeur résiduelle qui constitue la part de variance de Rendement scolaire qui n'est pas expliqué par l'effort (ce terme d'erreur est indiqué par une flèche pointant directement vers Rendement scolaire). Ici aussi, le calcul est simplifié (voir Kline, 2016, ou Wang et Wang, 2019 pour les calculs complets).

En somme, le modèle d'équations simplifiées présenté ci-dessus vise à expliquer l'équation 1.1 par une série d'équations structurelles, qui sont une combinaison de modèles de mesure (équation 1.2) et de modèles de structure (équation 1.3). Dans l'exemple de la figure 1.1, nous avons deux variables latentes, soit l'effort scolaire (mesuré par quatre VO) et le rendement scolaire (mesuré par trois VO). Les deux variables latentes et leurs indicateurs constituent la partie « mesure » du modèle, et la flèche du milieu, qui lie Effort et Rendement scolaire, la section « structure » du modèle. Les petites flèches qui aboutissent directement aux variables latentes ou observées constituent les indicateurs des erreurs de mesure (pour les variables observées) et les termes d'erreur (pour la VD, c'est-à-dire la proportion de la variance de Rendement Scolaire qui n'est pas expliquée par le modèle). Les termes d'erreur des indicateurs des facteurs Effort et Rendement scolaire sont expliqués sous l'équation 1.2 plus haut, alors que l'erreur résiduelle du Rendement scolaire est expliquée par l'équation 1.3.

2.6 Les structures de covariance et de moyennes

Les MES s'intéressent tout d'abord aux variances et covariances du modèle à l'étude, c'est-à-dire aux relations entre les variables. Rappelons que les covariances correspondent aux coefficients bêtas non standardisés d'une régression. L'utilisation des covariances a comme but de comprendre les patrons de relations entre les variables du modèle et d'expliquer la plus grande proportion de variance

possible dans le modèle. Les covariances sont analysées dans tous les types de MES. Le tableau 1.1 illustre un exemple de matrice de variance-covariance pour les VO de la figure 1.1.

Tableau 1.1 – Exemple fictif d’une matrice de variance-covariance pour l’exemple d’un MES représentant l’effort lors de tâches scolaires qui prédit le rendement scolaire (figure 1.1).

	Eff1	Eff2	Eff3	Eff4	Ren1	Ren2	Ren3
Eff1	3,414						
Eff2	1,235	0,975					
Eff3	0,984	0,716	1,090				
Eff4	0,564	0,565	0,687	1,438			
Ren1	0,250	0,111	0,188	0,208	1,654		
Ren2	0,232	0,140	0,223	0,181	0,402	1,419	
Ren3	0,323	0,054	0,085	0,039	0,911	0,863	1,046

Les variances sont en caractères gras.

Certains MES se soucient aussi des moyennes, qui prennent le plus souvent la forme de moyennes de variables latentes. Dans la plupart des MES, les moyennes sont estimées (sous forme d’ordonnée à l’origine d’une variable observée dans une analyse factorielle ou de moyenne d’une variable latente) mais ne font pas partie du modèle en tant que tel. Nous verrons dans les chapitres suivants certains types de MES où les moyennes font partie de l’analyse, soit pour en comparer la valeur entre un groupe et un autre, soit pour en comparer l’évolution dans le temps.

3. Comment commencer ? Les grandes étapes des MES

Bollen et Long (1993) et Kline (2016) offrent une marche à suivre afin d’estimer des MES que nous discutons ici. Les étapes sont les suivantes: 1) spécifier le modèle; 2) évaluer l’identification du modèle; 3) estimer le modèle; 4) tester l’adéquation du modèle; et enfin 5) respécifier le modèle si nécessaire.

3.1 Étape 1: spécifier le modèle

La partie la plus importante du processus sous-tendant les MES est la spécification de ces derniers. La spécification du modèle est la représentation théorique des hypothèses contenues dans le modèle hypothétique. Elle comprend tous les liens dictés par le modèle hypothétique. Ce modèle schématique correspond aux relations postulées entre les variables observées et latentes. Le modèle ne présente pas de résultats. La figure 1.1 plus haut est un exemple de la spécification d'un modèle à tester.

Le modèle doit prendre en compte plusieurs aspects :

- Les MES visent à confirmer **un modèle théorique** spécifié *a priori*. Idéalement, la spécification du modèle se fait lors de la préparation d'une étude avant la collecte de données. Elle doit toujours se faire avant de commencer les analyses.
- Il présuppose une **directionnalité**, indiquée avec les flèches présentées plus haut.
- Il suit le principe de **parcimonie**, c'est-à-dire qu'avec les MES nous cherchons à expliquer le plus de variance possible dans le modèle avec le moins de liens et de variables possibles. Donc, en MES, les modèles simples sont à privilégier. D'ailleurs plusieurs indices d'ajustement que nous verrons plus bas pénalisent les modèles complexes. De plus, les modèles plus complexes demandent une plus grande taille d'échantillon. Nous y reviendrons aussi plus bas.

3.2 Étape 2: évaluer l'identification du modèle

Bien qu'une explication approfondie du principe d'identification dépasse la portée de ce livre, elle est essentielle pour comprendre comment les modèles peuvent être spécifiés et les limites de la spécification. Kline (2016) offre une explication exhaustive du concept d'identification.

En termes généraux, l'identification du modèle se concentre sur la question de savoir s'il existe un ensemble unique de paramètres cohérents avec les données. L'application des MES implique l'estimation de paramètres inconnus sur la base des variances et covariances entre les données d'un modèle. Si une solution unique pour les paramètres inconnus est trouvée, le modèle est identifié. Le modèle est donc testable. Une condition nécessaire à l'identification est que l'on ne puisse pas estimer plus d'éléments qu'il n'y a de paramètres possibles dans le modèle.

La règle pour estimer le nombre de *paramètres possibles* (aussi appelés paramètres uniques, équivalents au nombre de points dans une matrice de covariances, incluant les variances) est expliquée à l'équation 2.4 (Bollen, 1989; Kline, 2016) :

$$\text{Nombre de paramètres possibles} = \text{N variables observées} \times (\text{N variables observées} + 1) / 2$$

Équation 1.4. – Paramètres possibles dans les modèles de variance-covariance.

Dans l'exemple de la figure 1.1, nous avons sept variables observées (les quatre indicateurs de l'effort, et les trois du rendement). Selon l'équation 1.4, le nombre de paramètres possibles du modèle est $28 : 7 \times (7 + 1) / 2$. Ce nombre est en tout point égal au nombre de valeurs présentes dans la matrice de variances-covariances du tableau 1.1.

Un principe d'identification est que le nombre de degrés de liberté dans le modèle doit être plus grand ou égal à zéro (0). Les degrés de liberté sont la différence entre le nombre de paramètres possibles et le nombre de paramètres à estimer (Kline, 2016). En d'autres mots, ils sont la différence entre les éléments connus (ceux inclus dans la matrice de variance-covariance, par exemple au tableau 1.1) et les éléments inconnus (ceux que l'on désire estimer, représentés par des flèches dans les modèles graphiques, par exemple à la figure 1.1). Les paramètres à estimer sont, par exemple, les poids factoriels, les pistes causales, les pentes de régression et les covariances.

Il y a trois types d'identification :

- **Sous-identification** : un modèle sous-identifié a *un nombre de degrés de liberté inférieur à zéro (négatif)*, c'est-à-dire que le nombre de paramètres possible est inférieur au nombre de paramètres à estimer. Ces modèles ne contiennent pas assez d'information dans le jeu de données pour être estimables. Ces modèles donnent invariablement un message d'erreur dans Mplus (et autres logiciels).
- **Identification juste** : un modèle à identification juste a *zéro degrés de liberté (0)*, c'est-à-dire que le nombre de paramètres à estimer est égal au nombre de paramètres possibles. Le modèle correspond donc exactement à la matrice de variances-covariances en termes de nombre de paramètres. Ces modèles sont estimables mais sont rarement considérés en MES, car ils ne donnent pas assez d'information quant à l'ajustement du modèle. D'un point de vue MES, les régressions linéaires simples ou multiples ont toujours une identification juste.
- **Sur-identification** : un modèle sur-identifié a *un degré de liberté positif (supérieur à zéro)* et fait partie de ceux à privilégier. Ces modèles ont un nombre de paramètres possibles plus grand que le nombre de paramètres à estimer. Ces modèles sont estimables et accompagnés d'indices d'ajustement suffisants. Les modèles sur-identifiés sont donc les modèles à privilégier.

Le but de l'étape 2 est donc de s'assurer que le modèle soit identifié (préféralement sur-identifié) avant de commencer les analyses. Dans la figure 1.1, nous avons 28 paramètres uniques (voir aussi la matrice de variance-covariance du tableau 1.1) et 15 paramètres à estimer. Ces paramètres à estimer sont : cinq liens entre les

indicateurs et les deux facteurs Effort et Rendement scolaire (le premier indicateur de chaque facteur, *Effort* vers *Eff1* et *Rendement scolaire* vers *Ren1*, est fixé à 1 par défaut dans les analyses factorielles confirmatoires et n'est donc pas à estimer puisqu'il est fixe), les sept variances résiduelles des indicateurs des deux facteurs, le lien causal entre Effort et Rendement scolaire, la variance de Effort, et finalement la variance résiduelle de Rendement scolaire ($5 + 7 + 1 + 1 + 1$). Donc nous avons 13 degrés de liberté (28 paramètres possibles – 15 paramètres à estimer). Selon le principe d'identification, le modèle de la figure 1.1 est sur-identifié, avec 13 degrés de liberté.

3.3 Étape 3 : estimer le modèle

Alors que les étapes de la spécification du modèle et d'identification (étapes 1 et 2) se font préférablement avant la collecte de données ou à tout le moins avant de commencer les analyses, l'étape d'estimation du modèle fait partie de l'analyse en tant que telle. L'estimation d'un modèle se fait en trois parties, soit choisir le bon estimateur, évaluer la qualité d'ajustement, ainsi que les poids estimés pour chaque paramètre.

3.3.1 Choisir le bon estimateur

Le choix d'estimateur en MES est multiple, et est fait en fonction du type de données et du type de MES à estimer. Les estimateurs disponibles selon chaque analyse seront discutés dans leurs chapitres respectifs et nous nous contenterons d'en donner un bref aperçu ici. Wang et Wang (2019) donnent une explication approfondie des estimateurs disponibles dans Mplus, en plus de proposer des estimateurs moins connus.

L'estimateur le plus populaire, et celui par défaut dans Mplus, est ML (*maximum likelihood*). Cet estimateur est adéquat avec un jeu de données continues et normalement distribuées, et une taille d'échantillon assez large. Si toutes ces conditions sont satisfaites, ML donnera des résultats libres de biais d'estimation. Autrement dit, les paramètres ne seront pas ni surestimés ni sous-estimés. De plus, ce mode d'estimation est relativement rapide.

Lorsque les données sont continues mais asymétriques (non normalement distribuées), MLR (ML qui est robuste «R» à la non-normalité des données) est à privilégier. Certains types de MES utilisent MLR par défaut. MLR est aussi à privilégier avec une petite taille d'échantillon. Un autre estimateur, MLM, est aussi disponible pour les données asymétriques mais ne s'applique pas avec un échantillon ayant des données manquantes.

Si le modèle contient des VD, ou des indicateurs d'une analyse factorielle, qui sont catégorielles, l'estimateur WLSMV (*weighted least square with robust means and variances*) sera avantage. Cet estimateur donnera des paramètres ajustés à l'aspect catégoriel des données. WLSMV tient compte des données manquantes.

3.3.2 Le traitement des données manquantes dans le choix d'estimateur

Les données manquantes se traitent par FIML (*full information maximum likelihood*, en français: calcul du maximum de vraisemblance à information complète) par défaut avec presque tous les estimateurs. Cette approche signifie que Mplus traite toutes les informations disponibles dans les données observées et utilise ces informations pour estimer les coefficients, écarts types et erreurs de mesure. Il est aussi à noter que l'on peut utiliser le traitement des données manquantes *listwise*, c'est-à-dire la suppression des sujets ayant des données manquantes. Cependant, le traitement par FIML sera préférablement utilisé si possible.

3.4 Étape 4: tester l'adéquation du modèle

L'objectif principal du processus d'estimation est de fournir des valeurs de paramètres telles que l'écart entre les données observées et le modèle théorique à tester soit minimal. En d'autres termes, le modèle spécifié à l'Étape 1 sera testé à l'aide des données observées et un ajustement sera de qualité si le modèle est similaire aux données. Comme les MES sont des analyses de confirmation, la structure du modèle est imposée aux données et non l'inverse.

Les indices d'ajustement indiquent dans quelle mesure le modèle correspond aux données. Plusieurs indices d'ajustement sont inclus dans les MES. Les plus importants dans Mplus sont le chi carré, le CFI, le TLI, le RMSEA, le SRMR, le BIC et l'AIC. D'autres indices spécifiques à certains types d'analyses précises seront décrits dans les chapitres concernés. Il est important de noter qu'un modèle de qualité est une condition *sine qua none* des MES. Si un modèle est adéquat selon les valeurs d'ajustement présentées ici, le chercheur pourra procéder à l'étape suivante. Sinon, le chercheur passera directement à l'étape 5: *respécifier le modèle si nécessaire*.

Afin d'évaluer la qualité des indices d'ajustement, les chercheurs ont mis au point des règles d'or pour évaluer les seuils pour un ajustement adéquat du modèle. Ces règles n'ont été que récemment évaluées à l'aide de simulations de données et sont à prendre avec un « grain de sel ». Nous référerons le lecteur aux ouvrages savants suivants pour plus de détails sur ces indices, les règles d'or et leur subjectivité