

**Faculté de psychologie et des sciences
de l'éducation**

Développement d'un module d'analyse Rasch pour l'interprétation des mesures en psychologie

Auteur : Catherine Absil
Promoteurs : Massimo Penta
Lecteur : Jacques Grégoire
Année académique 2019-2020
Master en sciences psychologiques à finalité spécialisée en clinique de
l'adulte

**Développement d'un module d'analyse Rasch
pour l'interprétation des mesures en psychologie**

Promoteur : Monsieur Massimo Penta

Co-promoteur : Monsieur Martin Edwards

**Mémoire présenté en vue de l'obtention du grade de Master psychologie
clinique par Catherine ABSIL**

NOMA 7525-18-00

catherineabsil@gmail.com

Louvain-la-Neuve

2019-2020

Mes premiers mots vont à toutes les personnes qui de près ou de loin ont contribué à la réalisation de ce mémoire.

En premier, Monsieur Penta, mon promoteur, que je remercie pour nos échanges, sa disponibilité et son extrême patience face à ma situation d'adulte qui cumulait travail et reprise d'études et un petit garçon de moins de 2 ans. Merci aussi pour la confiance qui m'a été accordée. Son expertise sur le modèle de Rasch a été très inspirante et motivante.

Mes remerciements vont aussi à Monsieur Edwards, mon co-promoteur, pour ses conseils avisés par rapport à la structure de ce mémoire un peu particulier.

Ensuite, je remercie également Philippe, mon mari, et James, mon fils, pour leur soutien, leur compréhension et leur patience face aux sacrifices imposés lors de ces quatre années. Merci de m'avoir permis de réaliser ce projet fou de reprendre des études. Merci d'avoir cru en moi et j'espère rattraper les longues absences.

Merci aussi à mes amis et proches dont mes condisciples d'études depuis quatre ans et en particulier Barbara, Lieselot, Sandra, Marie. Vous rencontrer a été non seulement un plaisir mais a aussi contribué avec certitude à ma réussite. Vos encouragements, votre enthousiasme ou regards positifs sur différents obstacles et dans les moments plus difficiles ont été d'un grand réconfort et mon permis de tenir jusqu'au bout de la réalisation de cette aventure.

Mes remerciements vont aussi aux membres de la communauté R qui partagent des réflexions ou des pistes de solutions ce qui m'a permis de progresser tout au long de l'élaboration de cet outil d'analyse Rasch en R avec eRm et qui je l'espère aidera d'autres étudiants, psychologues chercheurs intéressés par le modèle de Rasch.

Table des matières

Introduction.....	1
1 Contributions du modèle Rasch dans la recherche en psychologie.....	2
1.1 Psychométrie	2
1.2 Concept de mesure.....	2
1.3 Modèle de la théorie classique des tests (TCT).....	6
1.3.1 Le modèle	6
1.3.2 Indicateurs de qualité psychométrique.....	8
1.3.3 Quelques critiques du modèle.....	9
1.4 Le modèle de Rasch	11
1.4.1 Aspects théoriques	11
1.4.2 Plusieurs modèles.....	15
1.4.3 « Rating scale » (RSM) et « Partial crédit » (PCM).....	17
1.4.4 Indicateurs utilisés par le Modèle de Rasch	17
1.4.4.1 Fidélité et nombre de classes d'aptitude.....	17
1.4.4.2 Adéquation globale avec le modèle de Rasch et unidimensionnalité.....	18
1.4.4.3 Critère indépendance locale.....	20
1.4.4.4 Ciblage, courbe d'information et redondance	20
1.5 Recherches illustrant l'utilisation du modèle de Rasch	21
1.5.1 Réduction d'un questionnaire de qualité de vie (NHP38).....	21
1.5.1.1 Introduction	21
1.5.1.2 Approches de réduction	22
1.5.1.3 Réduction avec le modèle TCT	23
1.5.1.4 Réduction avec le modèle de Rasch	23
1.5.1.5 Comparaison des versions	24
1.5.1.6 Résultats.....	24
1.5.1.7 Conclusion.....	25
1.5.2 Analyse Rasch des échelles de dépression, anxiété et de stress.....	26
1.5.2.1 Introduction	26
1.5.2.2 Analyses Rasch.....	26
1.5.2.3 Résultats.....	27
1.5.2.4 Conclusion.....	28

1.5.3	Limites psychométriques de l'échelle réduite du sentiment de cohérence	28
1.5.3.1	Introduction	28
1.5.3.2	Analyse Rasch et résultats	29
1.5.3.3	Conclusion	29
1.5.4	Haut potentiel et blocage cognitif, émotionnel et motivationnel.....	30
1.5.5	Identification du haut potentiel	31
1.5.6	Unidimensionnalité de l'échelle d'engagement au travail d'Utrecht.....	33
1.5.6.1	Introduction	33
1.5.6.2	Analyse Rasch et résultats	33
1.5.6.3	Conclusion.....	34
1.6	Utilité d'un outil d'analyse R.....	34
1.6.1	Le langage R.....	34
1.6.2	Outil open source d'analyse R	35
1.6.3	RMarkdown	36
2	Développement d'un module d'analyse Rasch en R pour l'interprétation des mesures en psychologie	37
2.1	Introduction.....	37
2.2	Architecture de la solution et explication des choix	37
2.2.1	Notebooks d'Analyse	38
2.2.1.1	Notebook Analyse Rasch	38
2.2.1.2	Notebook Analyse TCT.....	38
2.2.2	Comparateur des résultats RUMM / R	39
2.2.3	Simulateur de jeu de données polytomiques.....	40
2.2.4	RStudio.....	41
2.2.5	Package eRm.....	41
2.2.6	Performance	42
2.2.7	Données.....	42
2.2.8	Librairies	42
2.2.9	Structure d'un Notebook.....	42
2.2.10	Procédure de production d'un rapport d'analyse	43
2.2.11	Visualisations graphiques	45
2.3	Conception et défis rencontrés.....	45
2.3.1	Notebook Rasch.....	45
2.4	Versions.....	50
2.5	Sorties pour différents jeux de données et comparaisons avec RUMM2030.....	50
2.5.1	Correspondance.....	51

2.5.2	Non correspondance	52
2.6	Comparaison entre les modules proposés et RUMM2030.....	55
3	Comparaison entre l'analyse TCT et le modèle de Rasch	59
3.1	Présentation du subtest Vocabulaire de la WAIS-III.....	59
3.2	Analyse TCT	60
3.2.1	Fidélité du subtest	60
3.2.2	Difficulté et discrimination des items.....	60
3.2.3	Ecart-type des items	62
3.2.4	Fiabilité et validité des items	62
3.2.5	Fonctionnement différentiel en fonction de l'âge et/ou du genre	63
3.3	Analyse Rasch.....	63
3.3.1	Les statistiques générales.....	63
3.3.2	Ciblage et difficulté des items	64
3.3.3	Critères d'unidimensionnalité	65
3.3.3.1	Fonctionnement différentiel des items+	65
3.3.4	Corrélation des résidus.....	66
3.3.5	Redondance des items.....	66
4	Conclusion.....	67
5	Bibliographie.....	1
6	Annexe 1	1
6.1	La psychométrie (Roulin, 2018)	1
7	Annexe 2.....	1
7.1	Réduction de l'échelle NHP38.....	1
7.1.1	Réduction de l'échelle NHP38 – Modèle TCT	1
7.1.2	Réduction de l'échelle NHP38 – Modèle Rasch.....	1
7.1.3	Contenu comparé des deux versions NHP réduites	2
7.1.4	Rasch analysis	2
8	Annexe 3.....	3
8.1	Echelle d'engagement au travail d'Utrecht	3
9	Annexe 4.....	4
9.1	Cahier de charge.....	4
10	Annexe 5 - Détails techniques d'implémentation de la solution	5
10.1	Défi : Utilisation de RMarkdown, sortie de tableaux.....	5
10.2	Défi: Configuration du notebook pour les paramètres.....	5
10.3	Défi : Extension de packages.....	7
10.4	Défi : Performance	8
10.5	Défi : Gestion des valeurs manquantes et conversion en NA	8
10.6	Défi : Selection de la plage de données souhaitée	9

10.7	Défi : Elimination des personnes et items non souhaités.....	9
10.8	Défi : Rescoring.....	10
10.9	Défi : Comptage des éléments.....	11
10.10	Défi : Calcul du fit et des thresholds sur base de l'algorithme choisi.....	12
10.11	Défi : Réalisation du LR test de Andersen	13
10.11.1	Idée de base.....	13
10.11.2	Mise en oeuvre	14
10.11.3	Intepétation de la pvalue	15
10.11.4	Evaluation graphique du Goodness of Fit.....	15
10.12	Défi : Person Parameters.....	16
10.13	Défi : Evaluation du niveau du trait latent des personnes.....	18
10.14	Défi : Calcul de l'itemfit et de personfit.....	18
10.15	Défi : Confection du tableau des items	19
10.16	Défi : Matrice de corrélation des résidus.....	19
10.17	Défi : Targeting	21
10.17.1	Targeting des personnes.....	21
10.17.2	Targeting des items	22
10.17.3	Person Item Threshold Distribution	23
10.17.4	Person-Item Location Distribution	24
10.18	Défi : Fonction d'information.....	26
10.19	Défi : Person et Item map (développement spécifique).....	26
10.20	Défi : Plot ICC	28
10.21	Défi : Visualisation du TCC Global.....	29
10.22	Défi : Visualisation Plot Bond-and-Fox pathway map (items, personnes).....	30
10.23	Défi : Calcul des métriques PSI, SSD, MSE, G/Separation Ratio,.....	31
10.24	Défi: Clustering en categories	32
10.25	Défi : Visualisations DIF.....	33
10.26	Défi : Résumé des données.....	37
10.27	Défi : Préparation et présentation des données sous forme de tableaux triés par critères	37
10.27.1	Préparation des données.....	37
10.28	Présentation sous forme de tableaux	38
10.29	Défi : ACP des résidus.....	38
10.30	Versions nécessaires.....	40
1.1.1.	Packages	40
1.1.2.	RStudio.....	42
1.1.3.	R.....	43
11	Annexe 6 – Analyses du subtest Vocabulaire de la WAIS-III.....	43
11.1.1	Test Lof	43

11.1.2	Exemple d'un graphe pour analyser le DIF	44
--------	--	----

Table des figures¹

Figure 1 - Probabilité de réussite à un item en fonction de l'âge pour le test d'intelligence de Binet et Simon (1916) – (Bertrand et al., 2004).....	11
Figure 2 - Score attendu à l'item i de difficulté δ_i – courbe caractéristique de l'item (CCI). (Penta et al., 2005).....	12
Figure 3 - Illustration du fait qu'une même augmentation de score observé ne donne pas une même progression sur l'échelle d'aptitude. (Wright & Masters, 1982).....	13
Figure 4 - Courbes CPC pour les catégories de réponses de l'item 9 du subtest vocabulaire de la WAIS-III avec le module d'analyse développé	15
Figure 5- (CCI) Courbe caractéristique de l'item 13 du subtest vocabulaire de la WAIS-III avec le module d'analyse développé	16
Figure 6 - Targetting et courbe d'information réalisé avec le module R.....	20
Figure 7 - Comparaison des apports ou limitations pour chacun des deux modèles - Penta & Lories (2018)	21
Figure 8- Corrélations entre les différentes formes d'échelles – Prieto et al., (2003)	24
Figure 9- calcul des corrélations - Prieto et al., (2003).....	25
Figure 10 - Capture d'écran de la vidéo mentionnée - RMarkdown. (s.d.). [Vidéo]	37
Figure 11 - Modules formant la solution	37
Figure 12 - Visualisation TCC : rapport entre difficulté et discrimination des items du subtest Vocabulaire de la WAIS-III.....	38
Figure 13 - Appel de la méthode de comparaison RUMM2030 / R.....	39
Figure 14 - Exemple de comparaison RUMM / Outil R.....	40
Figure 15 - Fit PCM sur données simulées	40
Figure 16 - Aperçu de la solution permettant le développement du module d'analyse Rasch en R	41
Figure 17 – Niveaux de titre (en gras) et chunks de code.....	42
Figure 18- bouton knit to HTML	43
Figure 19 - Données conservées dans l'environnement.....	43
Figure 20 - exécution isolée d'un chunk	43
Figure 21 - Notebook (.Rmd) et rapport (.html)	44
Figure 22- Accès au rapport.....	44
Figure 23 - Rapport au format HTML	45

¹ De nombreuses figures sont ou des créations originales ou des captures d'écran de codes et n'ont donc pas de sources.

Figure 24 - Comparaison RUMM2030 / Outil R.....	51
Figure 25- Vvaleurs RUMM2030.....	51
Figure 26 - Vvaleurs eRm.....	51
Figure 27 - RUMM vs Outil pour jeu de données depress	52
Figure 28- Valeurs RUMM2030 et R – Stats items.....	52
Figure 29 - GOF Outil R sur depress	53
Figure 30 - GOF outil R - confidence bands.....	54
Figure 31 - distribution personnes et items dans depress.....	55
Figure 32 - Comparatif de référence pour eRm - Robinson et al. (2019)	58
Figure 33 - Eléments à comparer entre les outils - Robinson et al. (2019).....	59
Figure 34 - occurrences des réponses à chaque item du subtest Vocabulaire de la WAIS-III	60
Figure 35 - rapports entre les indices de « difficulté » et discrimination des 33 items du subtest Vocabulaire de la WAIS-III.....	61
Figure 36- Ecart-type des items du subtes Vocabulaire de la WAIS-III.....	62
Figure 37- Fidélité des items du subtest Vocabulaire de la WAIS-III.....	63
Figure 38 - Targettingtreshold et persons avec courbe d'information	64
Figure 39 - TargettingItems et persons avec courbe d'information	64
Figure 40 - Matrice des corrélations des résidus.....	66
Figure 41 - Angles Clés pour le module d’analyse Rasch en R développé dans ce mémoire	67
Figure 42 - Mind map psychométrie - Roulin (2018).....	1
Figure 43- tableau des résultats de l'analyse TCT relative à la réduction du questionnaire de qualité de vie NHP38 – Prieto et al., (2003)	1
Figure 44 - tableau des résultats de l'analyse Rasch relative à la réduction du questionnaire de qualité de vie NHP38 – Prieto et al., (2003)	1
Figure 45 - liste des items du NHP38 et des versions courtes NHP20 (TCT) et NHP22 (Rasch)(Prieto et al., 2003)	2
Figure 46-(Prieto et al., 2003).....	2
Figure 47- Courbes des probabilités des différentes catégories des items polytomiques de l’UWES-17 avant et après recodage (De Bruin et al., 2013).....	3
Figure 48- cahier de charge du projet de développement du module d'anlyse Rasch en R	4
Figure 49- Entête et configuration du répertoire de travail.....	6
Figure 50- Localisation du fichier paramètres	6
Figure 51- Fichier paramètres relatif aux Données du subtest Vocabulaire de la WAIS-III qui doit être complété par l'utilisateur du module d'analyse Rasch en R.....	7
Figure 52 - Référencement des extensions	8

Figure 53- Mise en place du package par doParallel	8
Figure 54 -caFunctions.R - transformation des * en NA	9
Figure 55 - Selection des données pertinentes	10
Figure 56- Déclaration du rescoring	10
Figure 57 - Rescoring avec plyr et mapvalues	10
Figure 58 - Fonctions de support au comptage (caUtils.R)	11
Figure 59 - Comptage des éléments et somme totale pour verification.....	11
Figure 60 - Print screen des totaux et compte de vérification.....	11
Figure 61 - Estimation des paramètre	12
Figure 62- Tableau des locations et seuils des items	13
Figure 63 - Items faciles ou difficiles repérés dans le tableau des occurences.....	14
Figure 64 - Analyse d'un subset par rapport au fit	15
Figure 65 - Exemple des erreurs standard de fit avec 2 groupes (exemple technique)	15
Figure 66 - Exemple de fit correct	16
Figure 67 - Exemple de « fit » à problème avec bandes de confiance (item 1 et 3).....	16
Figure 68 - Calcul des paramètres des personnes et de l'intervalle de confiance de ceux-ci.....	16
Figure 69 - Plot de person.parameters	17
Figure 70 - Structure du résultat de person.parameters	17
Figure 71 - Intervalle de confiance des parametres des personnes	17
Figure 72 - Extraction des résidus pour les paramètres des personnes	17
Figure 73 - Evaluation des valeurs du trait latent des personnes	18
Figure 74 - Exemple de valeurs calculées pour le trait latent des personnes.....	18
Figure 75- Exemple d'item fit	19
Figure 76- Exemple de personfit.....	19
Figure 77 - Tableau synoptique des items	19
Figure 78 – Chunk relatif à la matrice de corrélation des résidus	20
<i>Figure 79 - Tracé de la matrice avec la palette spécifique</i>	20
<i>Figure 80 - Matrice de corrélation avec palette spécifique</i>	21
<i>Figure 81 - Table de corrélation des résidus</i>	21
Figure 82 -	22
Figure 83 -	22
Figure 84 -	22
Figure 85 -	23
Figure 86 -	23

Figure 87 -	24
Figure 88 -	24
Figure 89 -	25
Figure 90-	26
Figure 91-	26
Figure 92-	26
Figure 932-	26
Figure 94-	27
Figure 95-	27
Figure 96-	28
Figure 97 -	29
Figure 98-	29
Figure 99-	30
Figure 100-	30
Figure 101-	31
Figure 102-	31
<i>Figure 103 - Clustering</i>	32
<i>Figure 104 – Partitionnement</i>	33
<i>Figure 105 - Centres des groupes et vecteur associant un groupe à chaque personne</i>	33
Figure 106-	34
Figure 107-	34
Figure 108-	35
Figure 109-	35
Figure 110	35
Figure 111-	36
Figure 112-	36
Figure 113-	37
Figure 114 -	38
Figure 115-	38
Figure 116 -	38
<i>Figure 117 - Composantes Principales</i>	39
Figure 118 - Version de Rstudio utilisée pour l'élaboration du module d'analyse Rasch.....	42
Figure 119 - Version de R utilisée pour l'élaboration du module d'analyse Rasch.....	43
Figure 120- Résultat au test Lof.....	43

Table des tableaux

Tableau 1 : Aptitude du sujet par rapport à la difficulté de l’item.....	12
Tableau 2 - Tableau de synthèses des indices d’analyse des résidus selon RUMM2030 et R (Pental et al., 2005).	19
Tableau 3- Défis de conception et solutions apportées (outil Rasch)	46
Tableau 4 - Relevé items "à problème" selon les indicateurs	65

Table des abréviations et symboles

<u>Abréviations ou symboles</u>	<u>Significations</u>
TCT	Théorie classique des tests
MRI/TRI	Modèle de réponse à l'item / théorie de réponse à l'item
X	Le score observé
V	Le score vrai
E	L'erreur de mesure
$\epsilon(X)$	L'espérance mathématique du score observé
ρ_{EV}	Coefficient de corrélation entre le score vrai et l'erreur de mesure
$\rho_{E_1E_2}$	Coefficient de corrélation entre deux erreurs de mesures de deux passations du test à des temps différents. Les numérotations sont utilisées pour différencier les deux passations considérées.
$\rho_{E_1V_2}$	Coefficient de corrélation entre l'erreur de mesures obtenu à un test et le score vrai d'un autre test. Les numérotations sont utilisées pour différencier les deux passations ou tests considéré(e)s.
$\sigma^2_{E_1}$ et $\sigma^2_{E_2}$	Les variances de deux erreurs de mesure de deux passations différentes du test.
V_1 et V_2	Les deux scores vrais de deux passations différentes du test.
$V_2 + k$	Le score vrai d'une deuxième passation du test additionné à un nombre constant k.
p_i	L'indice de « difficulté » de l'item en TCT.
p	La p-value pour la probabilité « seuil » d'un test statistique.
D_i ou r_i	L'indice de discrimination de l'item en TCT.
Courbe CPC	Courbe « category probability curve » de l'item. Représente la probabilité de réponse pour chacune des catégories de réponse d'un item.
Courbe ICC	Courbe « Item characteristic curve » de l'item. Représente le score attendu à un item.

DASS	Dépression – Anxiety – Stress Scale
FDI/DIF	Fonctionnement différentiel de l’item/ Differential item functioning
ANOVA	Analyse of variance – analyse de variance
PSI	Fidélité de séparation des personnes
G	de séparation des personnes
ACP	Analyse en composantes principales
(β_n)	l'aptitude du sujet
(δ_i) .	la difficulté de l’item
CPC	Category probability curves/ courbes de probabilités des catégories de réponses
ICC/CCI	Item characteristic curve/ Courbe caractéristique de l’item
P_{ni1}	est la probabilité du n ^{ième} sujet de répondre correctement (score X =1) à l’item dichotomique i de difficulté (δ_i) en fonction de son aptitude (β_n) ,
$\epsilon(X_{ni})$	L’espérance mathématique du score observé X_{ni}

Introduction

Ce mémoire retrace les étapes de création d'un outil d'analyse Rasch à l'aide du langage R et présente diverses contributions du modèle de Rasch en psychométrie dans la recherche en psychologie. Différents chapitres le composent articulant présentation de concepts, comparaison de modèles et étapes de développement de cet outil qui permettra une utilisation gratuite du modèle de Rasch aux psychologues qui le souhaiteraient.

Proposer un tel outil en « open source » a du sens. En effet, les différents logiciels permettant des analyses Rasch et bien documentés sont payants, ce qui peut freiner certains utilisateurs potentiels. De plus, bien que le modèle de Rasch existe depuis 1960, selon Embretson et Reise (2000), la majeure partie des tests en psychologie (à leur époque) utilisent toujours des méthodes issues de la théorie classique des tests (TCT). Une des raisons explicatives est le manque de logiciels « accessibles » pour actionner ce modèle de Rasch. Borsboom (2006) fait ce constat également pour la recherche expérimentale sur la personnalité ou les attitudes ou encore le développement cognitif et l'intelligence. Shea, Tennant, et Pallant (2009) rapportaient que depuis une dizaine d'années, le modèle de mesure de Rasch, permettant d'étudier avec précision différents aspects d'une échelle, commence à être utilisé en psychologie. En 2020, le modèle de Rasch semble donc avoir fait son entrée en psychologie (haut potentiel, réduction d'échelle de qualité de vie) et en sciences de l'éducation (testing adaptatif). Cependant, les logiciels permettant son utilisation restent onéreux et nécessitent souvent une formation pointue pour les utiliser.

Le chapitre 1 de ce mémoire investigate le concept de mesure en psychologie et deux modèles d'évaluation d'échelles de mesure. Il s'agit du modèle de la théorie classique des tests et du modèle de Rasch. Ensuite, des études publiées utilisant le modèle de Rasch sont présentées dont certaines avec une comparaison TCT.

Le chapitre 2 explique certains des défis rencontrés lors des développements du module réalisé en langage R. Ensuite, les résultats obtenus pour différents jeux de données entrés dans le nouveau module R sont comparés avec ceux obtenus via le logiciel RUMM2030². Dans le chapitre 3, une analyse d'un subtest de la WAIS-III est réalisée avec le modèle de Rasch grâce au nouveau module R comparativement à l'analyse TCT de ce subtest. L'objectif est de présenter les informations pertinentes que l'analyse Rasch révèle. Enfin, le chapitre 4 explore les apports, limites et prolongements éventuels de ce travail.

² <http://www.rummlab.com/>

1 Contributions du modèle Rasch dans la recherche en psychologie

1.1 Psychométrie

Avant d'entrer dans le vif du sujet, il est important d'en dresser le cadre. Le modèle de Rasch s'inscrit dans ce qui s'appelle en psychologie, la psychométrie. Il en existe plusieurs définitions. Par exemple, « *La psychométrie concerne l'ensemble des principes à la base de la mesure en psychologie.* » (Roulin, 2018, p.8)

Ou

La psychométrie est un domaine d'étude qui concerne la théorie et la méthodologie de construction et d'utilisation des échelles de mesure des caractéristiques mentales. Celles-ci peuvent être des connaissances, des traits de personnalité ou encore des composantes de la cognition, comme le langage, l'intelligence ou la mémoire. (Selon Jacques Grégoire pour l'encyclopédie Universalis en ligne³)

Par conséquent, la psychométrie demande un éventail de connaissances et compétences dans les domaines des statistiques, de la construction des tests, de la déontologie et de l'échantillonnage ; sous-tendus par la notion de mesure (cf. [Annexe 1](#)) (Le Corff & Yergeau, 2017 ; Roulin, 2018). Dès lors, afin de pouvoir aborder dans ce mémoire les différents apports du modèle de Rasch, il est nécessaire de clarifier le concept de mesure.

1.2 Concept de mesure

Lorsque certaines caractéristiques d'un sujet (sa taille, son intelligence, ...) sont mesurées, un processus d'association avec un type d'étiquetage vérifiant certaines règles mathématiques s'établit, avec le sous-entendu que ces propriétés mathématiques «correspondent» aux propriétés des « objets » mesurés.

Cependant, force est de constater que ce concept de mesure n'est pas si évident. Les avis sont nombreux et certains d'entre eux s'opposent complètement.

Plusieurs auteurs (Bond & Fox, 2015 ; Vautier, 2015 ; Wright & Linacre, 1989) soulèvent par exemple des distinctions à faire entre une mesure et une évaluation ou le fait qu'il arrive que des scores bruts soient utilisés comme estimation de la capacité d'une personne

³ <https://www.universalis.fr/encyclopedie/psychometrie/>

lors d'examens alors que les items utilisés sont implicitement considérés comme « bons » et rarement analysés. Bond et Fox (2015) rapportent aussi des difficultés de compréhension du concept de mesure de la part d'étudiants dans leurs analyses ou une confusion répandue entre les mesures ordinales traitées directement erronément comme des mesures d'intervalles. Ils soulignent qu'au-delà de l'importance de l'analyse statistique nécessaire dans le processus scientifique, il est important que les chercheurs se penchent aussi sur la nature de leurs données, données auxquelles ils appliquent ces méthodes statistiques.

Un point de départ pour définir ce concept de mesure pourrait être que lorsqu'il y a mesure d'une manifestation d'une « grandeur », il y a attribution d'une « valeur » prise dans un ensemble d'observables utilisés comme mesure de cette grandeur (Vautier, 2016). Cependant, cela se complique assez vite selon ce qui est mesuré. En effet, la mesure peut concerner des phénomènes directement observables ou des phénomènes dérivés. Même parmi les grandeurs physiques, il y a des phénomènes non directement observables comme par exemple, la mesure de la température. En effet, il s'agit plutôt d'une mesure de l'impact de la température sur certains produits (par ex. le mercure), impact mesuré en mm (variation de longueur). En psychologie, de nombreux questionnements concernent des situations de mesure de caractéristiques des individus non directement observables comme par exemple, l'intelligence, l'anxiété, etc. C'est ce que les psychologues appellent des traits (ou variables) latent(e)s ou attributs ou facteurs. Les psychologues partent du principe que, bien que ces traits ne soient pas directement observables, il est possible de les inférer s'ils sont à l'origine de certaines manifestations externes (Le Corff & Yergeau, 2017 ; Penta, Arnould, & Decruynaere, 2005). Ils sont aussi supposés suffisamment stables pour que les chercheurs trouvent utile de développer une méthodologie permettant de les mesurer (Laveault & Grégoire, 2014). S'ajoute aussi que, pour de nombreuses théories psychologiques, le concept de variable latente a une place centrale au sein des modèles explicatifs (Borsboom, Mellenbergh, & van Heerden, 2003).

Pour aller plus loin, lorsque l'on effectue des recherches sur le concept de mesure en psychologie, on aboutit nécessairement à l'article de Stevens (1946). Intitulé « On the theory of scales of measurement », il est considéré par beaucoup d'auteurs comme étant une contribution majeure du 20e siècle en ce qui concerne le domaine de la mesure en psychologie. Dans cet article, il présente sa catégorisation de différentes échelles

numériques de mesure qui sont les échelles nominale, ordinale, d'intervalles et de ratio. Il utilise une définition dite « au sens large » avec laquelle tous les auteurs ne s'accordent pas. Pour Stevens (1946), mesurer consiste dans le fait « *d'attribuer des nombres à des objets ou à des événements selon des règles.* ». Derrière ce mot « règles », il y a la notion d'isomorphisme⁴ entre ce que l'on peut réaliser avec les aspects des objets et les propriétés des nombres utilisés. Sa catégorisation des échelles est donc fonction de ce que l'on peut réaliser ou pas avec les aspects de ces objets bien qu'il soit la plupart du temps tous représentés par des nombres utilisables dans les quatre types d'échelles. Par exemple, dans le cadre d'une échelle nominale associée aux genres homme et femme, les nombres 0 et 1 sont souvent utilisés et ils ne pourront dans ce cas pas être ordonnés alors que l'utilisation de ces mêmes nombres dans le cas d'une échelle ordinale de type Likert (éventuellement complétée d'autres niveaux) permet un ordre. Lorsque l'on passe de l'échelle nominale à l'échelle ordinale, une propriété est ajoutée à l'isomorphisme : la conservation de l'ordre. D'un point de vue mathématique, il s'agit d'une partition de l'ensemble des observations dans le cas d'une échelle nominale à laquelle s'ajoute une relation d'ordre dans le cas d'une échelle ordinale. Ce type d'échelle est très présent en psychologie.

Avec l'échelle d'intervalles, l'isomorphisme a une propriété supplémentaire : les écarts entre les observables ont une signification comme c'est le cas entre les nombres associés. Cette différence ou distance entre deux graduations de l'échelle correspond à l'unité de mesure connue et stable en tout point. La graduation 0 est fixée par une convention. L'échelle de ratio ou proportionnelle est une échelle d'intervalle munie d'un zéro absolu. Plus l'isomorphisme a de propriétés et plus l'échelle permet d'opérations sur les nombres associés. D'un point de vue statistique descriptive par exemple, on peut seulement s'intéresser aux fréquences d'observation et calculer le mode avec les échelles nominales. S'ajoutent à cela la médiane et les quartiles avec l'échelle ordinale. L'échelle d'intervalle nous permet de calculer en plus la moyenne.

Cette définition au « sens large » est l'objet de critiques. Certains auteurs considèrent que le score d'un test ne peut être plus qu'une échelle ordinale (Merbitz, Morris & Grip, 1989 ; Wright & Linacre, 1989). Or, le fait d'associer un même nombre de points à chaque item d'un test peut conduire à le considérer comme une échelle

⁴ En mathématiques, un isomorphisme est une application linéaire bijective entre deux espaces vectoriels de même dimension.

d'intervalles (Laveault & Grégoire, 2014). Et, en ce qui concerne les items, permettent-ils tous le même accroissement du trait mesuré ?

Penta et al. (2005, p.) signalent à juste titre que « *toutes les mesures sont des nombres, mais tous les nombres ne sont pas des mesures* ».

Une autre difficulté surgit lorsque l'on veut mesurer des traits latents en psychologie. En effet, on ne dispose pas d'unité de référence visible.

De plus, le concept de mesure prend aussi racine au sein de ce qu'on nomme l'épistémologie à savoir « *la discipline qui prend la connaissance scientifique pour objet* » selon le Larousse en ligne⁵ (2020). Or, selon les écoles de pensée, il existe des visions différentes du concept de mesure (Guyon, 2016). Les débats au sujet du concept de mesure ont été nombreux, prenant source soit en référence au paradigme utilisé pour définir la mesure soit critiquant le modèle utilisé pour établir une « mesure ». Avec les développements des modèles de la réponse à l'item (MRI) et le modèle de Rasch, les débats ont repris en force car ces modèles fournissent des échelles en accord avec le principe d'additivité (Bertrand & Blais, 2004). Par exemple, pour Borsboom (2006), la psychométrie n'a pas encore réellement percé en psychologie alors que de nombreux problèmes de mesure peuvent être relevés notamment en lien avec la validité. Pour lui, il y a des interprétations de résultats aux tests psychologiques discutables à mettre en relation avec le manque d'outils psychométriques récents. Parmi ces outils récents, il y a le modèle de Rasch qui, pour Andrich (2004), introduit un nouveau paradigme dans la mesure en psychologie.

Bien que l'objectif de ce mémoire ne soit pas d'explorer ces divergences, il est important de les mentionner afin de situer le cadre de ce travail et, de préciser, que les développements présentés s'appuieront sur le concept d'une mesure objective au sens donné par l'Institute for Objective Measurement⁶ :

Mesurer de manière objective consiste à répéter une quantité unitaire de grandeur constante, dans des limites acceptables d'erreur, quel que soit l'instrument utilisé pour mesurer la variable d'intérêt et quelle que soit la personne ou la chose mesurée.» (Traduit par Penta et al, 2005, p25).

⁵ <https://www.larousse.fr/dictionnaires/francais/%C3%A9pist%C3%A9mologie/30520>

⁶ Institute for Objective Measurement : IOM consulté sur www.rasch.org

Par conséquent, le terme de « mesure » conviendra pour les valeurs localisées sur une échelle d'intervalles que fournit le modèle de Rasch. Les autres nombres présents dans les jeux de données feront référence à ce qui est appelé, en psychologie, un score.

1.3 Modèle de la théorie classique des tests (TCT)

« Le test est une épreuve strictement définie dans ses conditions d'application et son mode de notation, qui permet de situer un sujet par rapport à une population elle-même bien définie (biologiquement et socialement). » (Zazzo, 1960 cité par Huet, 2018, p.26).

Un test est constitué d'items. Un item est un dispositif de recueil d'information par rapport à laquelle l'administrateur du test a un objectif d'exploitation (Huet, 2018).

La TCT aussi appelée théorie du « score vrai » s'appuie sur plusieurs postulats et différents indicateurs afin d'élaborer de nombreux outils de mesure de traits latents en psychologie. La standardisation, l'objectivité de la mesure et les propriétés psychométriques sont les aspects indispensables que tout instrument psychométrique se doit de vérifier (Le Corff & Yergeau, 2017). La TCT débute avec les travaux de Spearman entre 1907 et 1913 en parallèle avec les développements psychométriques de l'intelligence (Juhel, 1991). Ensuite, le modèle a été affiné par Gulliksen en 1950, Magnusson en 1967 et Lord et Novick en 1968 (Laveault & Grégoire, 2014).

1.3.1 Le modèle

Cette théorie repose sur plusieurs postulats.

Postulat 1 : $X = V + E$

L'hypothèse de départ posée par Spearman est que l'on ne peut mesurer sans qu'il y ait une erreur. Dès lors, le score observé (X) d'un sujet à un test est la somme du score vrai (V) inobservable de ce sujet (fixe d'une répétition à l'autre du test) et d'une erreur de mesure (E) aléatoire et constante pour tous les niveaux du trait latent. C'est ce qu'on appelle le postulat des sources de variance ou l'équation fondamentale de la TCT (Bertrand & Blais, 2004 ; DeVellis, 2006).

Postulat 2 : $V = \epsilon(X)$

La TCT suppose que les scores observés suivent une distribution normale. Par conséquent, le score vrai est la moyenne des scores observés dans le cas où il serait possible de réaliser une infinité d'administrations du test indépendantes. Comme les

erreurs successives sont aléatoires, leur moyenne est nulle (DeVellis, 2006). Bien que ce fait soit souvent énoncé comme un postulat, il s'agit plutôt d'une propriété découlant du postulat 1 (Bertrand & Blais, 2004):

$$\epsilon(X) = \epsilon(V + E) = \epsilon(V) + \epsilon(E) = V + 0 = V$$

Postulat 3 : $\rho_{EV} = 0$

Il n'y a pas de corrélation entre l'erreur de mesure et le score vrai. Par exemple, un score élevé n'entraîne pas une erreur élevée (Laveault & Grégoire, 2014).

Postulat 4 : $\rho_{E_1E_2} = 0$

Il n'y a pas de corrélation entre deux erreurs de mesures relatives à deux passations de l'instrument pour une population de sujets à qui le test a été administré. Aucune prédiction sur une erreur future n'est possible en connaissant une erreur passée. Il y a infraction à ce postulat quand il y des sources de variation communes comme par exemple des conditions environnementales impactant la passation, de la fatigue pour tout le groupe sujet, etc. (Bertrand & Blais, 2004).

Postulat 5 : $\rho_{E_1V_2} = 0$

Il n'y a pas de corrélation entre l'erreur d'un test et le score vrai à un autre test.

Laveault et Grégoire (2014) illustrent ce fait avec un exemple de test mesurant la créativité chez des sujets dont la propension à répondre au hasard augmente avec le niveau de créativité. Cela implique que l'erreur de mesure n'est pas aléatoire. Ce postulat n'est pas vérifié à partir du moment où le test mesure une aptitude qui a une influence (in)directe sur comment le sujet répondra au test.

Postulat 6 :

Deux tests sont dits parallèles si $V_1 = V_2$ **et** $\sigma^2_{E_1} = \sigma^2_{E_2}$

Leurs scores vrais sont égaux et leurs variances d'erreurs de mesure également. Il s'agit là de concepts théoriques permettant de développer le modèle TCT. En effet, cela va permettre, à l'aide du coefficient de corrélation entre deux tests parallèles, de définir la fidélité d'un test (Bertrand & Blais, 2004).

Postulat 7 :

Deux tests sont dits tau-équivalents si $V_1 = V_2 + k$ **avec** $k = \text{constante}$

Il est nécessaire d'avoir des tests tau-équivalents pour que le coefficient α qui donne une mesure de la cohérence interne d'un test, puisse prendre sa valeur maximale possible (Laveault & Grégoire, 2014).

1.3.2 Indicateurs de qualité psychométrique

Lors de la construction d'un test, le chercheur souhaite obtenir des scores observés au plus proches du score vrai et donc minimiser l'erreur de mesure. Afin de réaliser des mesures les plus précises possibles du trait latent, la TCT va parfois exclure des items de tests en fonction de leurs paramètres psychométriques et d'indices de corrélation. L'objectif est d'obtenir les items qui représentent le mieux le score vrai (Juhel, 1991). A cette fin, la TCT repose sur plusieurs indices de qualité psychométriques (le score total dépendant du nombre de réponses au test, la difficulté (p_i) et la discrimination (D_i) de l'item) ainsi que sur les notions de fidélité et validité du test (Laveault & Grégoire, 2014).

La validité du test porte sur ce qu'un test mesure en vérifiant qu'il mesure bien le trait qu'il est supposé mesurer alors que la fidélité porte sur la reproductibilité de la mesure. Une mesure non valide pourrait être fidèle mais le contraire n'est pas possible (Hogan, Parent & Stephenson, 2017 ; Laveault & Grégoire, 2014). Il y a différents types de validité (apparente, de contenu, critérielle et prédictive) (Hogan et al., 2017; Huet, 2018). Il s'agit d'établir en lien avec des objectifs ou populations ou contextes spécifiques, un certain degré de validité (Le Corff & Yergeau, 2017).

Avec la fidélité du test, il y a la notion de constance des résultats lors de différentes passations d'un même test avec une certaine marge d'erreur. La fidélité permet de rendre compte de la capacité du test à bien représenter le score vrai et donc quantifie d'une certaine façon l'impact de l'erreur de mesure (Le Corff & Yergeau, 2017). Le coefficient de fidélité (r) se définit comme la part de variance observée due à la variance vraie (DeVellis, 2006 ; Hogan, et al., 2017). Il y a trois types de méthodes permettant d'estimer la fidélité à savoir la méthode de stabilité (test-retest - à l'aide de coefficients de corrélation), la méthode de cohérence interne (homogénéité - à l'aide du coefficient alpha de Cronbach ou les formules de Kuder Richardson KR_{20} et KR_{21}), la méthode d'équivalence (le split-half - tests parallèles - à l'aide de coefficients de corrélation) (Hogan, et al., 2017 ; Huet, 2018 ; Le Corff & Yergeau, 2017). Hogan, et al., 2017 citent plusieurs auteurs qui trouvent une valeur minimum de 0,70 acceptable pour la fidélité. Cependant, la fidélité ne prend pas en compte les erreurs systématiques. Celles-ci réduisent ou augmentent le score attendu du sujet et doivent être repérées et distinguées de l'erreur aléatoire (Hogan, et al., 2017 ; Laveault & Grégoire, 2014).

Lors de la conception d'un test, le chercheur analyse les items grâce aux données collectées sur une population de référence. Il calcule l'indice de difficulté de chaque item du test ainsi que leur indice de discrimination.

La « difficulté » d'un item correspond à la moyenne des cotes attribuées à cet item pour tout le groupe des répondants. Attention, un item avec une « difficulté » $p = 0,9$ est en fait un item facile vu que 90% des répondants l'ont réussi (Hogan, et al., 2017; Laveault & Grégoire, 2014). De plus ce calcul ne tient pas compte des non-réponses (par ex. pour des items en fin de test) ni de la réussite de l'item qui serait due au hasard lorsque les items n'offrent pas tous le même nombre de choix de réponse. Dans ce cas, il existe une difficulté corrigée. Ensuite, lors de l'analyse d'une échelle, les items dont la difficulté varie entre 0,3 et 0,7 sont conservés (Laveault & Grégoire, 2014 ; Huet, 2018). Pour les items situés entre 0,1 et 0,3 ou entre 0,7 et 0,9, une analyse de leur niveau de discrimination permettra de décider de leur maintien éventuel. Les sujets d'aptitude plus faible sur le trait sont départagés par les items faciles du test alors que les items difficiles ne discriminent que les sujets d'aptitude plus haute sur le trait. Les items de difficulté moyenne discriminent parmi tous les sujets (Laveault & Grégoire, 2014).

La variance de l'item nous donne également une indication sur la dispersion des résultats à cet item et donc nous donne une indication des différences entre les répondants. Elle est maximale quand l'indice de difficulté vaut 0,5.

La discrimination d'un item est la capacité de l'item à discriminer les sujets de forte aptitude sur le trait (que le test mesure) de ceux de moindre aptitude. Elle peut se calculer grâce la différence entre la difficulté dans un groupe de sujets « forts » et la difficulté dans une groupe de sujets « faibles » (indice de Finley). Mais en général, elle est calculée en corrélant chaque item avec le score total au test ou le score total moins le score à cet item. Plus la corrélation est élevée et plus l'item est discriminant et cohérent avec le trait mesuré (DeVellis, 2006). Un minimum de 0,2 est nécessaire pour que l'item soit discriminant (Laveault & Grégoire, 2014). Le concepteur du test doit s'interroger sur les items qui discriminent peu ou négativement.

La corrélation de l'item avec le score total pondéré par l'écart-type de l'item nous donne la fidélité de l'item. En procédant de même avec le score à un autre test de référence, on obtient la validité de l'item.

1.3.3 Quelques critiques du modèle

1. Kubinger (2005) précise que beaucoup de tests sont conçus avec comme présupposé sous-jacent que le nombre de réponses correctes aux items serait une estimation juste de la « réussite » ou de l'aptitude. Il pose la question de savoir si le nombre de questions résolues à un test est réellement une statistique suffisante pour déterminer le paramètre d'aptitude attendu d'un sujet. Par exemple si des réponses à plusieurs items sont fortement

corrélées et que le sujet les réussit, son score élevé est une surestimation du score vrai. Ou lorsque dans un test, deux étudiants ont le même score alors qu'un seul d'entre eux a réussi tous les items très difficiles, que penser de leurs scores égaux ? Autrement dit, le score total au test ne rend pas bien compte du niveau réel d'aptitude (Bertrand & Blais, 2004 ; Hambleton & Jones, 1993).

2. Ce modèle suppose une erreur standard identique pour chacun des sujets quel que soit sa position sur la dimension examinée. Cette hypothèse est peu plausible (DeVellis, 2006 ; Roulin, 2018) car elle ne peut être identique à tous les niveaux du trait latent.
3. Le postulat 2 suppose une infinité d'administrations du test indépendantes mais, dans la réalité, un biais d'apprentissage peut rendre les mesures dépendantes des passations antérieures. De plus, l'opération de mesure en elle-même, peut aussi amener une certaine conscientisation chez le sujet appelée « effet opération de mesure » (Laveault & Grégoire, 2014).
4. La TCT se concentre sur le résultat au test et ne permet pas de faire des prédictions sur la réussite à un item du test. (DeVellis, 2006 ; Roulin, 2018).
5. Avoir deux versions parallèles a quelques inconvénients. D'un côté, cela représente un coût en argent et en temps et de plus, est impossible d'avoir des formes totalement équivalentes. Il y a donc une surestimation de l'erreur de mesure issue de cette « équivalence » imparfaite (Le Corff & Yergeau, 2017).
6. Les indices de difficulté et discrimination de l'item ainsi que la fidélité et la validité du test dépendent de l'échantillon utilisé. L'interprétation des indices se fait en fonction des sujets sélectionnés. En plus d'une certaine relativité des propriétés métriques, il sera difficile de comparer les items (Roulin, 2018). On peut illustrer ceci par la situation d'un échantillon faible par rapport au trait latent mesuré, ce qui amènera à considérer certains des items ratés comme difficiles alors qu'avec un échantillon mieux adapté, certains de ces items ne seraient pas classés comme difficiles et donc toute classification antérieure serait invalidée. Laveault et Grégoire (2014) pointent cette difficulté qui devient aigüe pour la création de banque d'items.
7. Les différences individuelles au sein de l'échantillon, la difficulté ou un nombre élevé d'items et le temps limité sont des éléments qui vont impacter la fidélité du test. Cortina (1993, cité par Laveault & Grégoire, 2014) a montré que pour 40 items ou plus, une valeur du coefficient alpha de 0,70 est facilement atteinte.

1.4 Le modèle de Rasch

1.4.1 Aspects théoriques

D'autres modèles dit « modèles de la réponse à l'item » (MRI) se sont développés avec comme objectif à la fois d'obtenir les caractéristiques psychométriques des items d'un test indépendantes de l'échantillon sélectionné ainsi qu'une mesure pertinente de l'aptitude des sujets pour un trait latent mesuré. On trouve à la base de ces modèles des fonctions dites logistiques. Ce sont des fonctions qui font partie de l'éventail de fonctions de régression. La méthode d'estimation associée est appelée méthode « du maximum de vraisemblance » (Bertrand & Blais, 2004). Ces fonctions ont un graphique dit de courbe en « S » qui avait déjà été observée début du siècle passé lors de travaux sur l'intelligence (Hambleton et al., 1985 cité par Bertrand & Blais, 2004).

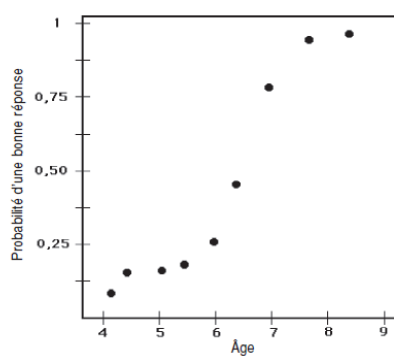


Figure 1 - Probabilité de réussite à un item en fonction de l'âge pour le test d'intelligence de Binet et Simon (1916) – (Bertrand et al., 2004)

Par la suite, dans les modélisations du lien entre l'aptitude d'un sujet et la probabilité de bien répondre à un item, l'adéquation de ces fonctions logistiques a été mise en évidence pour représenter ce lien. Les différents modèles MRI sont associées à différentes fonctions logistiques et l'un d'entre eux est « le modèle de Rasch » classé comme MRI unidimensionnel et à un paramètre bien qu'il n'ait pas été au départ conçu à partir de la même approche (Andrich, 1978, 2004).

Georg Rasch est un mathématicien Danois qui, à la fin des années 50, mène une étude sur les difficultés de lecture de sujets de 20 ans. En paramétrant l'aptitude des sujets et les difficultés des items, il arriva à un modèle de Poisson (Andrich, 2004). Il élaborait alors un modèle pour items dichotomiques permettant une mise en relation entre le score d'un sujet à un item donné avec son niveau d'aptitude sur le trait. Dans ce modèle de Rasch dichotomique, la probabilité de réussir cet item pour le sujet va uniquement dépendre de l'écart entre son niveau d'aptitude (β_n) et la difficulté de l'item (δ_i). Si P_{ni1} est la probabilité du $n^{\text{ième}}$ sujet de répondre correctement (score $X_{ni} = 1$) à l'item

i de difficulté δ_i en fonction de son aptitude β_n , trois situations sont possibles (Penta et al., 2005) :

Tableau 1 : Aptitude du sujet par rapport à la difficulté de l'item

L'aptitude du sujet est supérieure à la difficulté de l'item.	L'aptitude du sujet est égale à la difficulté de l'item.	L'aptitude du sujet est inférieure à la difficulté de l'item.
$\beta_n - \delta_i > 0$	$\beta_n - \delta_i = 0$	$\beta_n - \delta_i < 0$
$P_{ni1} > 0,5$	$P_{ni1} = 0,5$	$P_{ni1} < 0,5$

- ♦ Avec $P_{ni1}(X_{ni} = 1) = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}}$ (trait plein sur la figure 2)

Et la probabilité d'échec associée : $P_{ni0}(X_{ni} = 0) = \frac{1}{1 + e^{\beta_n - \delta_i}}$ (pointillés)

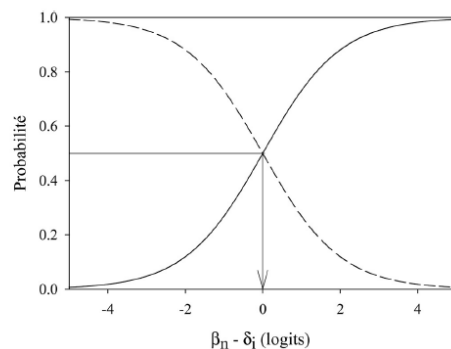


Figure 2 - Score attendu à l'item i de difficulté δ_i – courbe caractéristique de l'item (CCI). (Penta et al., 2005)

- ♦ P_{ni1} donne le score attendu à l'item i.
- ♦ Le quotient $\frac{P_{ni1}}{P_{ni0}}$ représente le rapport de vraisemblance pour la réussite à l'item i du sujet n d'aptitude β_n . Son logarithme donne, la différence $\beta_n - \delta_i$ en logits.
- ♦ Le logit (log-odds unit) est l'unité utilisée à la fois pour l'aptitude des sujets et la difficulté des items. Un logit représente une situation où l'écart entre l'aptitude et la difficulté est de 1 et par conséquent le rapport de vraisemblance qui correspond à l'exponentielle de cette différence vaudra e (2,718 ... ; Nombre de Neper) et cela tout au long de l'échelle.

Ce modèle postule que la probabilité de réussite à un item en fonction de l'aptitude de la personne croît de façon continue relativement au trait latent mesuré ; les items ayant tous une discrimination égale. Ces constats, permettent de dire que le modèle probabiliste de Rasch permet une mesure objective en transformant les scores observés sur une échelle ordinale afin de les situer sur une échelle linéaire d'unité constante (Andrich, 2004 ; Penta et al., 2005).

Le modèle de Rasch se distingue des autres modèles des MRI du fait qu'en plus de la modélisation par cette fonction logistique, il précise quels sont les critères nécessaires pour obtenir une mesure objective (Flieller, 1994). Dès lors, il est possible de vérifier si les données observées rencontrent ces critères.

De quels critères s'agit-il ?

♦ Le critère d'objectivité spécifique

La mesure produite par le modèle de Rasch respecte le principe **d'objectivité spécifique**. Cela signifie que la difficulté (localisation) des items est indépendante de l'aptitude des sujets, elle-même indépendante de la difficulté des items. « *L'objectivité au sens de Rasch concerne l'indépendance de la mesure par rapport à l'instrument de mesure.* » (Flieller, 1994). Cela différencie fortement le modèle de Rasch de la TCT où les mesures et observations obtenues sont dépendantes de l'échantillon. Pour Rasch il s'agit d'un principe épistémologique (Juhel, 1991).

♦ Le critère d'ordre

Le modèle travaille avec des données que l'on peut ordonner et le fait qu'un score plus élevé au test soit associé à une quantité plus grande du trait latent mesuré par le test.

♦ Le critère de linéarité

La transformation logit permet de linéariser les scores bruts. Avoir une échelle linéaire garantit pour un même accroissement de mesure, une même acquisition du trait mesuré. Cette propriété ne se vérifie pas sur une échelle ordinale. Sur l'échelle en logits la progression de l'aptitude est mesurable en tout point et non par incréments comme sur une échelle ordinale. Il s'agit d'un continuum de mesures du trait latent.

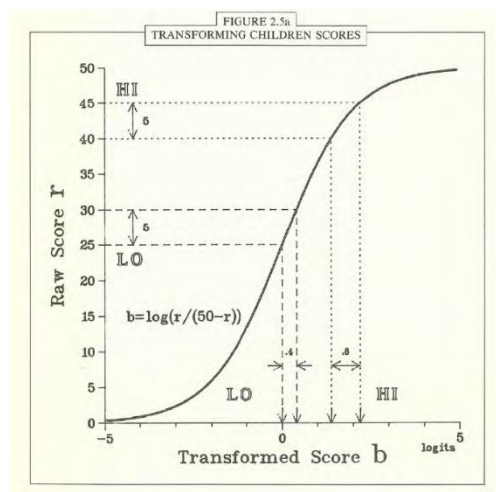


Figure 3 - Illustration du fait qu'une même augmentation de score observé ne donne pas une même progression sur l'échelle d'aptitude. (Wright & Masters, 1982)

C'est en prenant le logarithme népérien du rapport entre les probabilités de réussite et d'échec que l'on obtient cette échelle linéaire.

♦ Le critère d'unidimensionnalité

Parmi les critiques faites sur les scores totaux de différentes échelles, il y a le fait de mesurer plusieurs traits latents d'un sujet. Cela implique qu'il est alors difficile d'en inférer quelque chose de pertinent pour ce sujet. Le nombre de traits latents mesurés

correspond à la dimension de l'échelle utilisée à partir du moment où l'on considère qu'ils sont les seuls éléments explicatifs des choix des répondants. Avec le modèle de Rasch, la mesure s'effectue pour un seul trait latent et l'on tente de déterminer un jeu d'items cohérents avec une seule dimension (cf. [1.5.3.](#)) tout en sachant que l'unidimensionnalité des données n'est jamais parfaite (Flieller, 1994).

Le modèle de Rasch énonce différentes conditions à vérifier afin de se rapprocher du concept d'unidimensionnalité de la mesure. Par exemple, des statistiques d'ajustement relatives aux résidus permettent de repérer des items « problématiques ». Le terme « problématique » signifiant qu'ils manifestent un comportement autre que celui prévu par le modèle de Rasch. Il existe aussi le test sur items dichotomiques de Martin-Löf (avec extension aux polytomiques) qui utilise le rapport de vraisemblance et donne une indication pour le test entier (Boadé, 2013). Ces conditions à vérifier sont détaillées (cf. [1.4.4.](#)).

Wright & Linacre (1989) précisent que ces statistiques sont là pour aider aux compromis entre l'idéal d'unidimensionnalité du modèle de Rasch et les exigences de la pratique ; ceux-ci étant de la responsabilité du concepteur de l'échelle à la fois au niveau de choix posés mais aussi de la minimisation de leurs effets.

Il est aussi important de comprendre que l'unidimensionnalité peut cependant dépendre du niveau de concept relatif au trait. Kubinger (2005) souligne que le non ajustement des données au modèle démontre plutôt l'invalidité de la règle de notation utilisée. Bryce (1981) et Laveault & Grégoire (2014) attirent aussi notre attention sur le fait que ne pas satisfaire aux conditions du modèle de Rasch d'unidimensionnalité ne signifie pas toujours que l'échelle n'est pas unidimensionnelle.

♦ **Le critère d'indépendance locale**

L'indépendance locale est reliée à la notion d'événements indépendants en probabilités. En termes de mesure Rasch, cela signifie qu'il n'y a pas de relation entre les réponses aux items d'un sujet ayant passé le test. Dès lors, la probabilité de répondre à tous les items du test pour le sujet sera égale au produit des probabilités individuelles de réponse à chaque item du test (Flieller, 1994). Cela implique que le score attendu au test est calculé comme la somme du score attendu à chacun des items du test. Kubinger (2005) ajoute :

Il faut donc souligner que chaque fois qu'un auteur de test suggère l'utilisation du nombre d'items correctement répondus, le modèle de Rasch doit nécessairement s'appliquer. Dans le cas contraire, le score du test ne permettrait pas du tout de comparer équitablement les résultats des personnes testées.

1.4.2 Plusieurs modèles

L'usage est de parler du modèle de Rasch mais en fait cette appellation recouvre plusieurs modèles. Au départ Rasch a élaboré le **modèle dichotomique** (Bertrand & Blais, 2004). Ensuite ce modèle a été adapté aux items polytomiques sous 2 formes. D'une part, Andrich (1978) a développé le **modèle « Rating scale »** et d'autre part, Wright & Masters (1982) ont développé le **modèle « Partial Crédit »**. Ces modèles polytomiques concernent les échelles dont les items ont au moins 3 catégories de réponses chacune se voyant attribuer un score distinct. Les scores commencent à 0 pour la catégorie correspondant au niveau le plus bas du trait latent mesuré par l'item en augmentant d'une unité à chaque catégorie. Pour chacun des items de l'échelle, nous pouvons représenter :

- ◆ Les courbes donnant la probabilité de répondre aux différentes catégories (CPC)

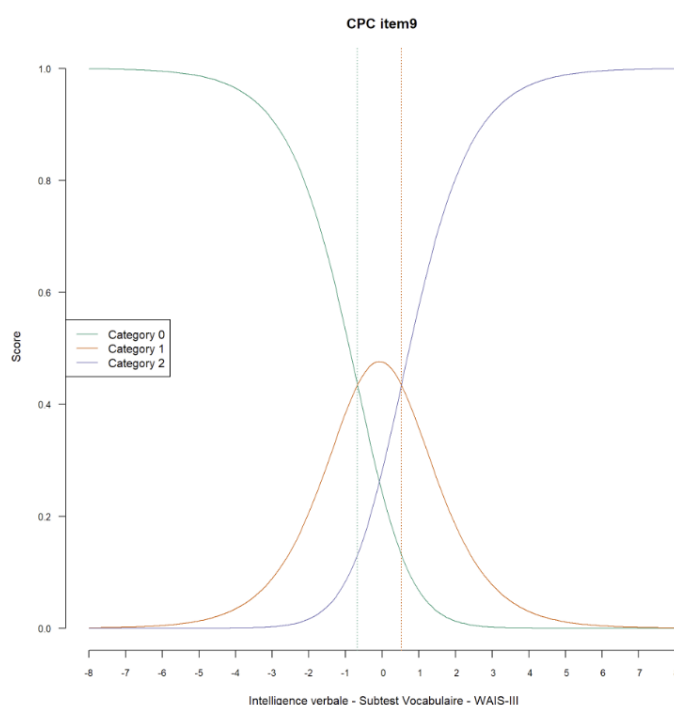


Figure 4 - Courbes CPC pour les catégories de réponses de l'item 9 du subtest vocabulaire de la WAIS-III avec le module d'analyse développé

En fonction de l'aptitude du sujet, on peut lire sur le graphe CPC sa réponse la plus probable à l'item. Les pointillés indiquent les endroits où les probabilités de répondre à deux catégories adjacentes sont équiprobables. Ce sont les « seuils » de l'item dont on mesure la position au niveau de l'échelle. On voit sur ce CPC deux seuils aux intersections des courbes de probabilité des différentes catégories. Ils sont dits « ordonnés » parce qu'ils respectent la progression de croissance du trait (monotonie).

La localisation (difficulté) d'un item est obtenue en calculant la moyenne des seuils de cet item. Lorsque la référence des seuils se fait par rapport à leur localisation exacte

sur l'échelle, les seuils sont dits « décentralisés ». Si cette référence se fait par rapport à leur écart à leur moyenne alors ils sont dits « centralisés ».

- ◆ Ensuite, la courbe caractéristique de l'item (CCI) donnant le score attendu à l'item.

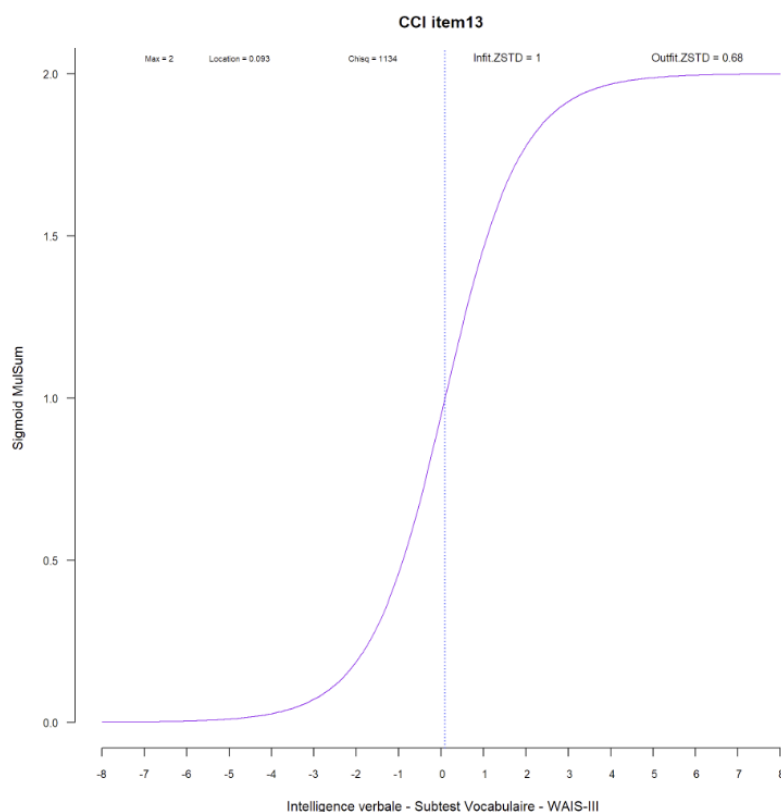


Figure 5- (CCI) Courbe caractéristique de l'item 13 du subtest vocabulaire de la WAIS-III avec module d'analyse développé

On voit en pointillé bleu la localisation de la difficulté de l'item (0,093) correspondant au point d'inflexion de la courbe. Le score attendu s'obtient en sommant les résultats des produits obtenus en multipliant chaque score de chaque catégorie par sa probabilité de réalisation pour chaque valeur β_n de l'échelle. Ce calcul manifeste la nature probabiliste du modèle de Rasch qui détermine la réponse la plus probable d'un sujet à un item en fonction de son niveau d'aptitude.

Dans ce modèle, la probabilité de réussir cet item pour le sujet va dépendre de son niveau d'aptitude, de la difficulté de l'item et faire intervenir la position des seuils de l'item sur l'échelle (cf. Penta et al. (2005) pour la formule détaillée, p.42). Il est possible grâce aux réponses des sujets aux items, d'estimer la valeur de ces différents paramètres inconnus au départ via différents algorithmes qui fournissent aussi une erreur de mesure.

Certains logiciels dont RUMM2030 utilisent la « procédure pairée » qui permet de déterminer les localisations relatives de deux items grâce à la probabilité conditionnelle d'en réussir un seul sur les deux. Cette estimation est généralisée à tous les tels couples

d'items du test Penta et al. (2005). Ensuite RUMM2030 utilise la procédure « UCON » pour estimer la localisation des sujets en se basant sur tous les paramètres (items et sujets) du modèle. Cette procédure utilise un processus itératif afin de déterminer les valeurs des paramètres qui maximisent la probabilité d'obtenir la matrice des scores observés. Utilisée seule pour les items et les personnes, cette seconde procédure comporte un biais dans la localisation des items du fait qu'elle fait intervenir la localisation des personnes pour les déterminer. Certains auteurs ont montré une correction de ce biais liée au nombre d'items.

Le package eRm de R utilise la méthode du maximum de vraisemblance conditionnel (CML) qui vise à déterminer la localisation des items sans faire intervenir celle de sujets (Andersen, 1973 ; Boadé, 2013). Il n'y a que dans le modèle de Rasch que cette méthode est possible (Flieller, 1994). Cependant, pour les paramètres des personnes, eRm utilise la méthode joint maximum likelihood (JML) (Mair, & Hatzinger, 2007a ; 2007b ; Mair, Hatzinger, & Maier, s.d.).

1.4.3 « Rating scale » (RSM) et « Partial crédit » (PCM)

Le choix du modèle sera fonction des besoins du chercheur qui crée une échelle de mesure ou de l'échelle dont on cherche à améliorer les propriétés psychométriques. Pour une échelle où tous les items sont cotés avec la même échelle ordinale et où tous les items ont les mêmes seuils centralisés, le modèle RSM sera appliqué.

Par conséquent, les écarts de progression nécessaires pour augmenter en aptitude sur le trait sont identiques. Par contre pour une échelle où tous les items sont cotés avec la même échelle ordinale avec des écarts entre les seuils variables, il convient d'appliquer le modèle du PCM. Bien qu'il apporte plus de variabilité permettant un meilleur ajustement, le modèle PCM peut conduire à une estimation de paramètres trop faibles (Linacre, s.d.).

1.4.4 Indicateurs utilisés par le Modèle de Rasch

Les scores attendus calculés grâce aux modèles de Rasch et les scores vrais peuvent être comparés. Voici différentes analyses à mener afin de vérifier que les données s'ajustent aux modèles de Rasch.

1.4.4.1 Fidélité et nombre de classes d'aptitude

Un coefficient de fidélité de séparation des personnes et un index de séparation des personnes sont calculés (Wright & Masters, 1982). La fidélité est satisfaisante à partir de 0,70, correcte de 0,70 à 0,85 ; excellente de 0,85 à 0,95 et parfaite au-delà (Huet, 2018).

1.4.4.2 Adéquation globale avec le modèle de Rasch et unidimensionnalité

Pour un même score total, plusieurs sujets peuvent avoir répondu différemment aux items. Au vu du postulat de croissance de l'aptitude ou de la difficulté le long de l'échelle, il est important d'analyser si les schémas de réponses y correspondent ou révèlent quelque chose d'inattendu. Pour y arriver, pour chaque observation, l'écart entre le score observé (X_{ni}) et le score attendu $\epsilon(X_{ni})$ est calculé. Il s'agit des résidus (R_{ni}) (Flieller, 1994). Ils représentent une mesure de l'écart entre ce que le modèle prescrit et les données relevées. Au plus ils sont proches de 0, meilleur est l'ajustement.

Résidus bruts

$$R_{ni} = X_{ni} - \epsilon(X_{ni})$$

Résidus standardisés

$$Z_{ni} = \frac{X_{ni} - \epsilon(X_{ni})}{\sqrt{V(X_{ni})}} \sim N(0, 1)$$

Les résidus calculés pour chaque observation peuvent être cumulés de différentes façon et donner ainsi des indices d'analyse.

♦ L'item – Fit :

Cela donne un indice pour chaque item permettant de vérifier l'ajustement entre le modèle et les résultats de chaque individu.

♦ Le person-Fit

Cela donne un indice pour chaque sujet permettant de vérifier l'ajustement entre le modèle et les résultats de chaque individu

♦ L'overall – Fit : concerne le test en entier. Un test non significatif indique une adéquation de scores vrais avec une échelle unidimensionnelle. Différents tests peuvent être utilisés.

Rumm2030 utilise une approche par intervalles. Afin de déterminer leur amplitude, les résidus sont cumulés dans chaque intervalle ce qui donne l'indice FitRes. La valeur de cet indice doit se situer dans l'intervalle $[-2,5 ; 2,5]$. Ensuite, un χ^2 est calculé afin de déterminer la probabilité d'observer des résidus avec une telle amplitude (Penta et al., 2005).

Le package eRm de R travaille avec les réponses individuelles. Dès lors, à partir des résidus standardisés, la somme de leurs carrés moyens est calculée. Ensuite, ils sont standardisés. Cela donne les indices :

$$OUTFIT.MNSQ \rightarrow OUTFIT.ZSTD \sim N(0,1)$$

Ensuite, la somme des carrés moyens pondérée par la variance est calculée. Ensuite, ils sont standardisés. Cela donne les indices :

$$INFIT.MNSQ \rightarrow INFIT.ZSTD \sim N(0,1)$$

Dans les deux méthodes, trois situations se présentent :

Tableau 2 - Tableau de synthèses des indices d'analyse des résidus selon RUMM2030 et R (Pental et al., 2005).

RUMM2030		
<i>FitRes</i> < -2,5	<i>FitRes</i> = 0	<i>FitRes</i> > 2,5
χ^2 significatif	χ^2 non significatif	χ^2 significatif
Dans ce cas, l'item montre une discrimination plus forte (pente plus forte) qu'attendu, on se rapproche d'un modèle plus déterministe, les réponses des sujets sont trop systématiques. Néanmoins, le plus souvent ces items sont non problématiques. Ex : item ch3	Dans ce cas, le modèle et les données s'ajustent. C'est la situation recherchée. Ex : item ch3	Dans ce cas, l'item montre une discrimination plus faible qu'attendue. Il est difficile de distinguer les sujets par rapport à leur aptitude avec ce type d'item (pente plus basse ou désordre). De plus, il semble que les réponses des sujets puissent dépendre d'une autre dimension. Ex : item ch3.
<u>OVERFIT</u>	<u>FIT</u>	<u>UNDERFIT</u>
$OUTFIT.MNSQ < 1 - \frac{6}{\sqrt{N}}$	1	$OUTFIT.MNSQ > 1 + \frac{6}{\sqrt{N}}$
$INFIT.MNSQ < 1 - \frac{2}{\sqrt{N}}$	1	$INFIT.MNSQ > 1 + \frac{2}{\sqrt{N}}$
$OUTFIT.ZSTD < -2$ Et $INFIT.ZSTD < -2$	0	$OUTFIT.ZSTD > -2$ Et $INFIT.ZSTD > -2$
R		

Toujours pour vérifier le postulat de monotonie, l'ordre des seuils doit être vérifié. Si ceux-ci sont désordonnés alors il convient d'analyser les items concernés afin d'en comprendre la cause et envisager la suppression de l'item.

De plus, dans le cas où le choix se porte sur le modèle RSM (Andrich, 1978), une vérification des distances inter seuils est à faire. Linacre (2002) indique que cette distance

ne doit pas être inférieure à 1,4 logits ou supérieure à 5 logits sans quoi les interprétations cliniques deviennent difficiles à réaliser. Les items ne respectant pas cette distance adéquate entre les seuils sont généralement supprimés (en restant vigilant à l'impact de cette suppression sur la fidélité notamment).

Ensuite, une analyse du fonctionnement différentiel des items est menée à l'aide d'une ANOVA sur ces résidus pour RUMM2030. L'échelle ne doit pas présenter de fonctionnement différentiel en misfit (Facteur 1 : localisation des sujets) ou uniforme (Facteur 2 : sexe, âge, etc.) ou encore non-uniforme (interaction F1 et F2). L'analyse se fait par rapport à la correction de Bonferroni pour les différentes caractéristiques connues des sujets. On peut également utiliser un test t sur les difficultés estimées des items dans deux groupes formé relativement à un critère (Penta et al., 2005)

1.4.4.3 Critère indépendance locale

Une analyse de la matrice des corrélations entre les résidus des différents items est aussi réalisée afin de vérifier qu'elles sont toutes dans l'intervalle $[-0,3 ; 0,3]$. En dehors de cet intervalle, cela signifierait qu'il y a une certaine redondance entre les items aux résidus corrélés. Par conséquent de tels items seraient candidats à la suppression.

1.4.4.4 Ciblage, courbe d'information et redondance

Il s'agit d'examiner la distribution des items et d'analyser la zone de mesures en logits couverte par les items et les seuils des items afin de repérer les effets plafond ou plancher éventuels et les zones d'aptitude éventuellement non couvertes par les items.

Parmi les candidats à la suppression suite aux différents critères, il est important de s'assurer qu'ils ne sont pas seuls à discriminer les sujets dans une zone de l'échelle.

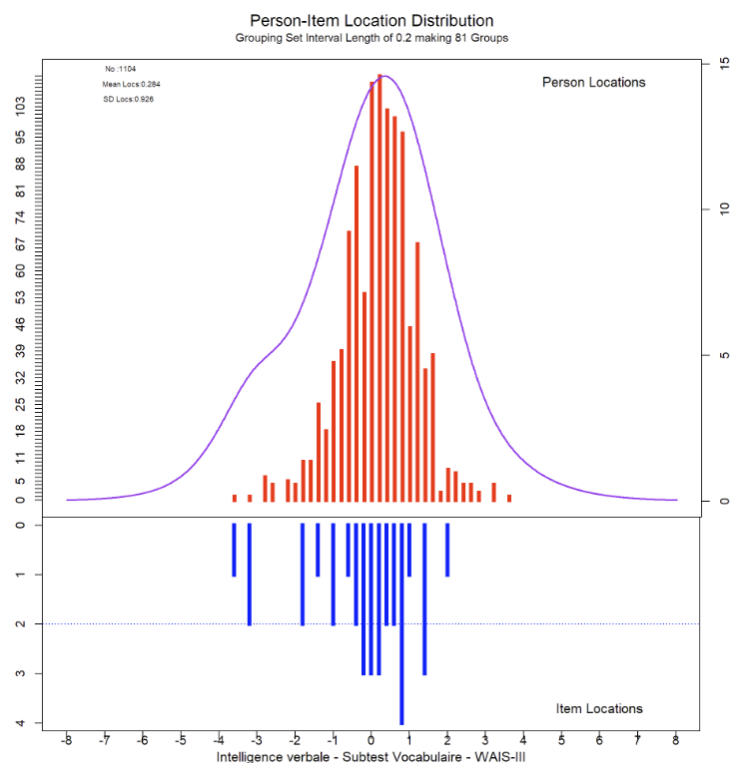


Figure 6 -
Targetting et
courbe
d'information
réalisée avec le
module R

De plus, un des objectifs du concepteur d'échelle est d'avoir une échelle la plus courte possible fournissant la mesure la plus précise possible du trait. Par conséquent si plusieurs items discriminent une même zone, il sera envisagé d'en éliminer certains.

La courbe d'information nous informe du pouvoir informatif soit d'un item soit du test tout entier et donc est à mettre en lien avec le concept de fidélité.

1.5 Recherches illustrant l'utilisation du modèle de Rasch

Cette section investigate et résume différentes études utilisant le modèle de Rasch. Certaines d'entre elles le font en comparaison avec la TCT. Afin de faciliter la compréhension des différentes recherches, voici un tableau de synthèse mettant en évidence les apports du modèle de Rasch sur la mesure de la variable latente par rapport au modèle TCT.

Paramètre psychométrique	Théorie classique	Modèle de Rasch
Échelle	Score brut	Mesure linéaire
Données	Doivent être complètes	Valeurs manquantes possibles
Interprétation	Classement ordinal sur variable latente espérée	Localisation linéaire sur variable latente explicitement définie par les items
Continuité	Échelle discrète	Échelle continue
Fidélité	Erreur moyenne via l'alpha	Erreur type pour chaque personne / item
Analyse linéaire	Impropre	Applicable
Validité	Coefficient de fidélité	Validité de construit, déterminée par analyse de l'unidimensionnalité
Diagnostic	Coefficient de corrélation item/total	Indice d'ajustement pour chaque item / sujet
Conceptualisation	Concrète	Abstraite
Contexte	Dépendant des distributions	Invariant

Modifié d'après B.D. Wright. *Rasch Measurement Transactions*, 1992, 6:1, 208

Figure 7 - Comparaison des apports ou limitations pour chacun des deux modèles - Penta & Lories (2018)

1.5.1 Réduction d'un questionnaire de qualité de vie (NHP38)

1.5.1.1 Introduction

Prieto, Alonso, et Lamarca (2003) comparent les modèles TCT et de Rasch dans leur utilisation en vue de réduire un questionnaire sur la qualité de vie. En effet, les instruments relatifs à la qualité de la vie se trouvent limités dans leur utilisation clinique car ils sont trop longs. Leur étude présente les différences dans cette démarche de réduction d'un questionnaire de bien être qui garde son efficacité. Ils ont constaté que jusqu'en 2003, les méthodologies utilisées pour effectuer des réductions de questionnaires ont le plus souvent été hétérogènes et non standardisées. La majorité

d'entre elles se base le plus souvent sur des procédures statistiques de la TCT. Les auteurs pointent pourtant l'existence de limitations conceptuelles à l'approche TCT, principalement la non-existence d'un continuum d'items ordonnés qui constitueraient une représentation unidimensionnelle du trait latent mesuré et l'absence d'une propriété d'additivité des données de l'échelle de mesure. Le modèle de Rasch s'annonce donc prometteur dans le fait qu'il fournit une méthodologie pour construire une échelle de mesure permettant un examen de la structure hiérarchique, respectant des critères d'unidimensionnalité et vérifiant l'additivité des mesures.

Afin de réaliser cette comparaison, les auteurs ont choisi le questionnaire Nottingham Health Profile (NHP) constitué de 38 items (NHP38) dichotomiques. Le questionnaire NHP38 contient des questions décrivant des problèmes selon six dimensions de la santé. Les réponses de 9419 sujets ont été analysées au travers de 45 études avec données parmi 119 études identifiées sur la NHP38, de 1987 à 1995. A la fin du processus de réduction, le modèle TCT aboutira à un questionnaire de 20 items (NHP20) et le modèle de Rasch à un questionnaire de 22 items (NHP22). Dans le cas de la version réduite Rasch, le questionnaire fournit aussi une mesure physique et une mesure psychologique.

1.5.1.2 Approches de réduction

En ce qui concerne le modèle TCT, l'approche habituelle est d'utiliser une procédure qui inclut l'analyse de facteurs, la corrélation entre la forme courte et la forme longue, les corrélations entre les scores à l'item et le score composites, l'alpha de Cronbach (cf. [1.3.2.](#)). La sélection finale est réalisée en fonction la consistance interne. Alternativement, le modèle de Rasch fournit une autre approche pour établir une échelle où les items sont placés hiérarchiquement et dispose d'autres indicateurs pour mener une procédure de réduction (cf. [1.4.4.](#)). Etant donné qu'un instrument réduit peut présenter des faiblesses dans ses propriétés de mesure, il est important de les vérifier.

La comparaison des résultats des deux approches menées par Prieto et al. (2003) s'effectue sur la réduction de la version espagnole de l'échelle du profil de santé de Nottingham (NHP38). Les propriétés de mesure des deux questionnaires finaux (NHP20 et NHP22) ont été testées et comparées.

L'étude de la qualité de vie sur différentes populations amène à réaliser des évaluations globales au sein de de groupes sociodémographiques différents. La nécessité de travailler avec un seul indicateur décrivant l'état de santé de la population étudiée s'impose facilement. Par conséquent, les auteurs ont travaillé dans cette optique afin d'obtenir, via

des questionnaires réduits, un indicateur unique réalisant la synthèse de l'état de santé du répondant. Cependant, les auteurs ont prévu également un profil en plusieurs nombres pour les chercheurs et utilisateurs ne se satisfaisant pas d'un indicateur unique (plus difficile à interpréter).

1.5.1.3 Réduction avec le modèle TCT

L'approche utilise des procédures standard de statistiques appliquées aux 38 items.

- ◆ Calcul des indices de discrimination (r_i) et de difficulté (p_i) pour chaque item.
- ◆ 10 items avec un $r_i < 0.4$ et un p_i bas (< 0.20) ou haut (> 0.80) ont été exclus.
- ◆ Une analyse exploratoire des facteurs (AEF) est effectuée sur les items restants (l'AEF supprimant tous les cas avec des valeurs manquantes).
- ◆ Une réduction supplémentaire est effectuée éliminant les items qui n'ont pas une contribution importante à la variance des facteurs communs (communality < 0.3) ainsi que d'autres critères.
- ◆ L'alpha de Cronbach est calculé pour estimer la consistance interne de chaque nouveau score composite. Le score résumé du questionnaire réduit a été obtenu en sommant et moyennant les scores des dimensions de ses composants.
- ◆ Une AEF additionnelle indique si les nouvelles dimensions déterminées peuvent donner un score unique en sortie.

Le tableau reprenant certains résultats est en [annexe 7.1.1](#).

1.5.1.4 Réduction avec le modèle de Rasch

Comme présenté plus avant dans le chapitre, cette approche donne lieu à l'estimation de la difficulté de l'item et de l'aptitude des sujets mutuellement indépendantes. De plus, la calibration des items définit la hiérarchie relative des items (en termes de difficulté) tout au long du continuum de la qualité de vie.

Il est possible d'évaluer l'unidimensionnalité de l'échelle à l'aide de statistiques d'ajustement (cf. [1.4.4](#)). Un indicateur important dans l'approche est l'index de séparation qui permet d'obtenir une valeur de la fidélité et le nombre de niveaux d'aptitude que le test mesure dans un échantillon donné.

Plusieurs analyses successives ont été réalisées jusqu'à ce qu'un jeu d'items final satisfasse aux exigences d'ajustement du modèle. Ensuite les auteurs se sont également interrogés sur le fait de savoir si un nombre substantiel de personnes répondent comme le prédit la nouvelle échelle NHS22. De plus, la réduction du questionnaire amenant une perte de sensibilité, les auteurs ont introduit deux options de cotation et ont regroupé, en tenant compte de leur expérience préalable du NHP38, les 38 items en deux échelles différentes. La première est relative aux aspects physiques (énergie, douleur, mobilité

physique) et contient 19 items. La seconde, également de 19 items, mesure la dimension psychologique (réactions émotionnelles, sommeil et isolation sociale). Une analyse Rasch est effectuée sur chaque groupe d'items et donne alors deux scores, qui sont alors sommés pour donner un score final. L'analyse a montré via les statistiques des résidus, 9 items mal ajustés. Ceux-ci ont été supprimés de même que les items mal ajustés dans les analyses suivantes jusqu'au moment où il n'y avait plus aucune amélioration de l'ajustement. Cela a conduit finalement à l'élimination de 16 items parmi les 38 et 18 des 22 items de la version courte satisfont au critère d'unidimensionnalité. Les 4 items non ajustés ayant un impact important sur le coefficient de fidélité ainsi que sur des localisations extrêmes sur l'échelle conceptuellement valides, ont finalement été retenus. L'analyse finale a fourni 3 strates de niveaux de qualité de vie.

Le tableau reprenant certains résultats est en [annexe 7.1.2](#).

1.5.1.5 Comparaison des versions

Pour valider la stabilité des résultats obtenus par les deux modèles, les sujets de la base de données sont découpés aléatoirement en deux sous-jeux de données indépendants en s'assurant qu'ils sont représentatifs de la population dans son ensemble :

- ♦ Un jeu A de 85% de la population, soit 8015 sujets. Ce jeu sert à la calibration.
- ♦ Un jeu B de 15% de la population, soit 1404 sujets. Ce jeu sert à la vérification.

Divers instruments de comparaison ont été utilisés.

1.5.1.6 Résultats

Une vue synoptique (cf. Figure 8) des deux approches montre les indicateurs pertinents permettant la validation ou non des versions réduites :

Table 5: Association* of the original NHP38 and the two alternative short-forms: the NHP20 and NHP22

	NHP20					NHP22		
	Total score	Emotional	Physical	Sleep	Pain	Total score	Physical	Psychological
Total NHP38	.97	.83	.82	.73	.76	.97	.88	.88
Energy	.77	.72	.69	.48	.60	.77	.76	.65
Pain	.84	.59	.80	.51	.91	.82	.88	.61
Emotional Reactions	.79	.92	.52	.58	.53	.78	.59	.85
Sleep	.78	.56	.50	.98	.53	.73	.55	.79
Social Isolation	.54	.59	.39	.37	.39	.61	.47	.66
Physical Mobility	.82	.58	.94	.47	.65	.84	.93	.60
Total NHP20	-	-	-	-	-	-	-	-
Emotional	.83	-	-	-	-	-	-	-
Physical	.84	.56	-	-	-	-	-	-
Sleep	.76	.56	.48	-	-	-	-	-
Pain	.79	.56	.66	.51	-	-	-	-
Total NHP22	.95	.82	.82	.71	.76	-	-	-
Physical	.88	.63	.91	.52	.80	.91	-	-
Psychological	.87	.87	.58	.78	.59	.88	.65	-

* Spearman's Correlation Coefficients

Figure 8- Corrélations entre les différentes formes d'échelles – Prieto et al., (2003)

Le contenu comparé des deux versions réduites est le suivant est en [annexe 7.1.3](#).

Les deux approches ne retiennent pas les mêmes items pour leur version réduite de NHP38. En termes de corrélation des résultats des tests réduits par rapport au test NHP38, on voit que les deux tests réduits montrent des corrélations fortes (cf. Figure 9).

Table 5: Association* of the original NHP38 and the two alternative short-forms: the NHP20 and NHP22

	NHP20					NHP22		
	Total score	Emotional	Physical	Sleep	Pain	Total score	Physical	Psychological
Total NHP38	.97	.83	.82	.73	.76	.97	.88	.88
Energy	.77	.72	.69	.48	.60	.77	.76	.65
Pain	.84	.59	.80	.51	.91	.82	.88	.61
Emotional Reactions	.79	.92	.52	.58	.53	.78	.59	.85
Sleep	.78	.56	.50	.98	.53	.73	.55	.79
Social Isolation	.54	.59	.39	.37	.39	.61	.47	.66
Physical Mobility	.82	.58	.94	.47	.65	.84	.93	.60
Total NHP20	-	-	-	-	-	-	-	-
Emotional	.83	-	-	-	-	-	-	-
Physical	.84	.56	-	-	-	-	-	-
Sleep	.76	.56	.48	-	-	-	-	-
Pain	.79	.56	.66	.51	-	-	-	-
Total NHP22	.95	.82	.82	.71	.76	-	-	-
Physical	.88	.63	.91	.52	.80	.91	-	-
Psychological	.87	.87	.58	.78	.59	.88	.65	-

* Spearman's Correlation Coefficients

Figure 9- calcul des corrélations - Prieto et al., (2003)

Ces résultats indiquent que les formes réduites NHP22 et NHP20 semblent être de bonnes alternatives à NHP38. Cependant, l'analyse d'adéquation Rasch appliquée à NHP20 (le test réduit issu de l'approche TCT) donne des résultats non conformes aux critères de validité et de fiabilité. Les scores des diverses dimensions de NHP20 donnent des mauvais résultats en analyse Rasch.

1.5.1.7 Conclusion

Les auteurs se sont aussi questionnés sur le fait que l'approche Rasch avait éliminé une des catégories de l'indice de qualité de vie NHP38 (isolement social) alors que le NHP20 garde des items dans toutes les catégories initiales. La forte corrélation entre les scores totaux de la forme NHP22 et NHP38 suggère que cette suppression n'a pas affecté structurellement le concept de qualité de vie mesuré par le NHP38.

L'étude ne donne pas une réponse définitive à la question de quelle approche est la plus pertinente pour la réduction mais la mauvaise performance de NHP20 (cf. [Annexe 7.1.4.](#)) aux critères Rasch, a amené les auteurs à retenir NHP22 comme version réduite de NHP38.

Un désavantage supplémentaire de la forme NHP20 est l'absence d'une hiérarchie des items, ce qu'offre NHP22. En conclusion, NHP22 est une forme réduite prometteuse de NHP38.

1.5.2 Analyse Rasch des échelles de dépression, anxiété et de stress

1.5.2.1 Introduction

Les échelles DASS sont un instrument composé de 42 items et qui mesure les niveaux des états émotionnels de dépression, anxiété et stress. La DASS-42 mesure les spécificités de la dépression et de l'anxiété de façon distinctive. A cela s'ajoute une sous-échelle relative au stress psychologique. Pour les auteurs, arriver à diagnostiquer la dépression est très important du fait des nombreuses conséquences de cette maladie sur la santé du sujet (à moyen terme et à long terme), son travail, sa vie sociale et sa vie familiale. Un outil de dépistage fiable, valide et assez court serait très utile pour le diagnostic de la dépression par les médecins traitants. L'étude menée concerne donc la forme réduite de la DASS constitué de 21 items. Celle-ci contient aussi 3 sous-échelles qui sont composées de 7 items de 4 catégories de réponses (0 à 3). Une analyse TCT précédente donne à la DASS-21⁷ des propriétés psychométriques meilleures que la DASS initiale. Cependant, les auteurs rapportent différentes études mettant en évidence des problèmes d'ajustements ou qu'une DASS-21 à 2 facteurs serait plus adaptée.

Shea, Tennant, & Pallant (2009) ont souhaité réalisé une analyse Rasch pour évaluer les propriétés psychométriques de cette version courte, via deux formes d'administration (papier crayon : 210 sujets et en ligne : 535 sujets). Les auteurs ont néanmoins évoqué un biais possible du fait que des études suggèrent une désinhibition possible des sujets lors d'une enquête en ligne. Ensuite, il y a eu un tirage au sort de 210 répondants en ligne parmi les 535 afin de respecter les exigences de l'analyse ANOVA et avoir deux groupes de même effectif. Les objectifs de l'étude portaient sur le fait d'en découvrir davantage sur la dimensionnalité de l'échelle ou le fonctionnement différentiel des items en particulier par rapport aux deux formes d'administration.

1.5.2.2 Analyses Rasch

L'analyse Rasch s'est effectuée en deux temps. En premier une analyse Rasch des 3 sous-échelles à 7 items séparées. Ensuite, afin d'étudier la validité du score totale de la DASS-21, une analyse Rasch sur les 21 items en tant qu'échelle globale de la mesure de la détresse psychologique générale. S'agissant d'items polytomiques, la question s'est posée de choisir entre le modèle rating scale ou partial crédit. Une étude détaillée de l'ordre des seuils et des distances inter seuils ont été réalisées. L'ajustement des items a été vérifié ainsi que leur FDI selon 3 groupes d'âge, 3 groupes de niveau d'éducation, 2 formes d'administration du test. Les valeurs des index de fidélité de séparation de chaque

⁷ La DAS-21 peut être téléchargée sur <http://www2.psy.unsw.edu.au/groups/dass/>

sous-échelle ou l'échelle globale sont supérieures ou égales à 0,7 (fidélité satisfaisante à bonne).

1.5.2.3 Résultats

En ce qui concerne la sous-échelle « dépression », l'analyse Rasch a au départ indiqué une p-value de l'ajustement significative ($< 0,001$). Deux items (it. 5 et 10) avaient une p-value d'ajustement significative. L'item 5 a été supprimé (résidu élevé) ce qui a amélioré la p-value précédente (0,03) et plus aucun item avec un problème d'ajustement. A cette étape, la fidélité de séparation (0,89) indique une très bonne fidélité. Aucun FDI pour l'âge ou le niveau d'éducation ou la forme d'administration n'a été mesuré. Par contre 2 items présentaient un FDI selon le genre (it. 13 et 16). L'ACP n'a pas révélé plus qu'une dimension.

Par contre, les sujets de l'échantillon ne présentent pour beaucoup d'entre eux que très peu de symptômes de dépression. En effet, 35% d'entre eux obtiennent une localisation sur le bas de l'échelle. Le ciblage n'était donc pas adéquat.

Le procédé a été répété avec la sous-échelle « anxiété ». Une première analyse a révélé un mauvais ajustement du modèle ($p < 0,001$) avec 2 items dont les seuils étaient désordonnés (it. 4 et 7). Un recodage n'a pas amélioré l'ajustement du modèle et les auteurs ont alors choisi de garder la situation initiale.

En analysant les résidus des items, 3 items mal ajustés (it. 2, 15, 20) ont été repérés. La suppression de l'item 2 a amélioré l'ajustement global (0,02).

Aucun FDI pour l'âge, le genre, le niveau d'éducation ou la méthode d'administration n'a été relevé. La fidélité de séparation finale était de 0,77 ce qui est une fidélité moindre mais satisfaisante pour les auteurs. L'ACP n'a pas révélé plus d'une dimension.

A nouveau, un problème de ciblage était présent du fait d'un échantillon dont 40 % des sujets étaient situés sur le bas de l'échelle avec un niveau moyen d'anxiété de -1,78 logits.

En ce qui concerne la sous-échelle « stress ». Une première analyse a révélé un mauvais ajustement du modèle ($p < 0,001$). Aucun item n'avait de seuils désordonnés. En analysant les résidus des items, 2 avaient des résidus fortement négatifs (it. 11 et 12). Seule la suppression de l'item 11 a amélioré l'ajustement global (0,01).

Aucun FDI pour les différentes variables n'a été relevé.

La fidélité de séparation finale était de 0,80 ce qui est une fidélité acceptable. L'ACP n'a pas révélé plus d'une dimension.

Le niveau moyen de stress des sujets était de -1,03 logits sans concentration de sujets au bas de l'échelle.

Pour l'échelle Globale, une première analyse a révélé un mauvais ajustement du modèle ($p < 0,001$) avec 52 items dont les seuils étaient désordonnés (it. 4, 7, 17, 20 et 21). Un recodage n'a pas amélioré l'ajustement du modèle et les auteurs ont alors choisi de garder la situation initiale.

En analysant les résidus des items, 5 items mal ajustés (it. 2, 1, 5, 19, 14) ont été supprimés et cela conduit à une fidélité de séparation finale de 0,90 ce qui est excellent.

L'ACP n'a pas révélé un facteur lié aux items de dépression et un second aux items de stress et d'anxiété. Les auteurs ont alors décidé de tester la possibilité de former une nouvelle échelle unidimensionnelle avec les items « anxiété » et « stress ». Après la suppression de 3 items, un ajustement global au modèle de Rasch a été réalisé avec une fidélité correcte (0,84) et aucun FDI. L'ACP n'a pas révélé de dimension supplémentaire. En terme de localisation, les items « stress » sont situés au bas de l'échelle et les items « anxiété » en haut de l'échelle.

1.5.2.4 Conclusion

En conclusion les auteurs préconisent que l'analyse Rasch de cette échelle soit continuée sur d'autres échantillons afin de confirmer les résultats obtenus avec suppression d'un item dans chacune des sous-échelles (pour l'ajustement au modèle de Rasch). Il est important de signaler que Linacre (1994) a montré que la taille de l'échantillon utilisée dans cette étude est suffisante et permet une précision d'un demi-logits au seuil de 99% même en étant mal ciblé.

1.5.3 Limites psychométriques de l'échelle réduite du sentiment de cohérence

1.5.3.1 Introduction

Lerdal, Opheim, Gay, Moum, Fagermoen, & Kottorp (2017) ont mené une étude sur les limites psychométriques de l'échelle SOC-13, échelle du sentiment de cohérence à 13 items. C'est une réduction de la version complète SOC-29. Cette échelle a 3 sous-échelles : une cognitive (compréhension), une comportementale (gérabilité), et une motivationnelle (sens). Jusque-là, les propriétés psychométriques de la SOC-13 ont été étudiées via la TCT. Ces études ont conclu que la SOC-13 est un instrument valable et fiable. Cependant, les auteurs ont constaté lors d'une précédente recherche, certains problèmes psychométriques lors de son utilisation avec des sujets en obésité morbide. La suppression de l'item 1 avait donné une échelle SOC-12 avec de meilleures propriétés

psychométriques. Ils se sont demandés si ce constat était spécifique à ce groupe cible ou s'il pouvait se généraliser à d'autres populations.

1.5.3.2 Analyse Rasch et résultats

La démarche utilisée dans leur précédente étude a donc été reproduite pour un groupe de sujets atteint de maladies inflammatoires de l'intestin (MII). Les 428 adultes norvégiens interrogés ont rempli la version norvégienne de la SOC-13 et à l'aide de l'analyse Rasch les auteurs ont évalué leurs réponses au questionnaire SOC-13 qui est composé de 3 sous-échelles. Chaque item permet 7 niveaux de réponse (Likert). L'analyse Rasch dans cette étude a révélé des problèmes de seuils désordonnés pour les items 7 et 12 (sens) et 5 (gérabilité). L'ajustement du modèle est acceptable et aucun effet plancher ou plafond n'a été détecté mais l'indice de séparation des personnes par contre a donné des valeurs trop basses aux sous-échelles sens et compréhension et une valeur limite pour la gérabilité ($1,54 \rightarrow 2$ strates ; fidélité de séparation $\cong 0,70$).

Ensuite, seul un item (it. 6) a montré un FDI avec la caractéristiques « état civil ». En effet, il était plus facile d'être en accord pour les sujets non mariés avec cet item. Du fait des faibles résultats en validité et fidélité aux 3 sous-échelles, les auteurs ont poursuivi leur analyse sur la SOC-13 entière. Les résultats pour l'unidimensionnalité sont limites. L'indice de séparation est de 2,19. Les items 1 et 5 présentant des problèmes d'ajustement ont été enlevés de l'échelle et les analyses ont continué sur la SOC-11 ainsi obtenue et dont l'indice de séparation avait à peine varié (2,10). L'item 1 a été éliminé à nouveau car il est inadapté pour le modèle de Rasch.

1.5.3.3 Conclusion

Les auteurs pensent donc au terme de l'étude que cette échelle réduite SOC-13 ne peut correspondre à un concept d'échelle unidimensionnelle à moins que certains items soient éliminés. De plus, les 3 sous-échelles ont une fiabilité soit non acceptable soit limite ; alors que, lorsque n'on ne distingue plus les 3 sous-échelles, les indices sont meilleurs. Les auteurs concluent en préconisant de reprendre l'échelle longue SOC-29 et d'entamer une analyse Rasch des items pour créer une nouvelle échelle réduite éventuellement munie de sous-échelle mais toutes avec de meilleurs indices d'ajustement et de fidélité. L'analyse a aussi révélé que les 7 niveaux de réponses ne semblent pas pertinents et qu'une réduction du nombre de catégories (recodage) devrait être envisagée.

1.5.4 Haut potentiel et blocage cognitif, émotionnel et motivationnel

Schiltz & Schiltz (2007) rapportent les résultats d'une étude expérimentale sur l'utilisation d'un testing adaptatif informatisé (en mathématiques) ciblant des élèves à haut potentiel rencontrant des difficultés scolaires et vivant des blocages cognitif, émotionnel et motivationnel à la puberté. Cette étude concerne 112 élèves de 6^e secondaire⁸ parmi lesquels se trouvaient 35 élèves à haut potentiel.

Au début d'un test adaptatif, le sujet reçoit un item dont la difficulté est moyenne. A chaque réussite, le test lui présente un item plus difficile. A partir du moment où il échoue à un item, le test lui présente un item plus facile. Il s'agit donc d'un test « sur mesure ».

Les auteurs mettent en évidence les résultats de la recherche sur ce public en blocage cognitif et/ou émotionnel et/ou motivationnel qui peut les conduire à la préadolescence à vivre une baisse d'estime de soi, des difficultés relationnelles, une morosité prolongée ou des comportements agressifs. Pour eux, selon leur expérience clinique, associer une prise en charge psychothérapeutique et une prise en charge pédagogique est indiquée. Ils ont testé l'efficacité et l'efficience de cette combinaison d'accompagnements dans des études longitudinales. Ils préconisent donc le développement de moyens pédagogiques adaptés à la situation motivationnelle des élèves surdoués. A travers cette étude expérimentale, ils souhaitaient vérifier l'impact d'une situation d'examen adaptée au niveau de l'étudiant. L'hypothèse étant qu'il serait possible de lever le blocage cognitif se manifestant par un échec aux examens de mathématiques. Si cette hypothèse se vérifie, ce type de testing pourra servir à dépister des ressources latentes du sujet à haut potentiel dans une telle situation de blocage lors de bilan psychologiques. Les modèles d'analyses des données qui ont permis dans cette étude expérimentale la construction de ce type de testing adaptatif sont les modèles de réponses à l'item à 2 paramètres.

Cependant, il aurait été possible de construire du testing adaptatif à l'aide du modèle de Rasch. C'est ce que Wenders & Visschers (2016) ont réalisé dans le cadre de leur mémoire à l'UCL en sciences informatiques sous la direction du professeur Olivier Bonaventure. Elles ont réalisé le premier test adaptatif de l'UCL pour un cours d'informatique permettant l'auto-évaluation des étudiants sur tout le contenu du cours en vue de présenter leur examen. Il était disponible sur la plate-forme INGINious afin d'être facilement accessible. Parmi les avantages mentionnés, les auteurs ont mentionné, le gain de temps car ce type d'évaluation adaptée au niveau de l'étudiant se fait avec moins

⁸ 6^e secondaire au Luxembourg, ce qui correspond au passage du primaire au secondaire.

d'items et une évaluation plus ciblée et précise de l'aptitude de l'étudiant. Le langage R a été utilisé également parmi d'autres solutions techniques.

1.5.5 Identification du haut potentiel

Toujours dans la recherche relative aux enfants à haut potentiel, une réflexion est menée par Caroff (2004) sur les limites de l'approche psychométrique « classique » permettant leur identification et les possibilités offertes par les MRI dans ce contexte d'identification. Il pointe le peu de références bibliographique (5%) de l'époque consacrées à la problématique de l'identification alors que la fréquence d'utilisation des procédures d'identification est nettement plus élevée (par le fait notamment que toute recherche relative au haut potentiel amène à réaliser des identifications). De plus, les concepts psychométriques nécessaires à cette identification sont complexes notamment du fait des nombreuses définitions du haut potentiel mais aussi du fait de certaines limites de l'approche de la TCT. Il illustre son propos à l'aide de la WISC.

L'auteur retrace l'évolution du concept de précocité passant d'une vision unitaire (mesure globale du QI) à des conceptions multifactorielles ou multidimensionnelles (intelligences multiples). Bien que de plus en plus d'auteurs soient d'accord sur le principe d'une multi-évaluation pour identifier le haut potentiel, la majorité des recherches continuent à n'utiliser qu'un seul critère d'identification qui est soit une mesure d'intelligence soit de performance (Ziegler et Raul, 2000, cités par Caroff, 2004). L'auteur signale qu'opérationnaliser des critères qui reposent sur des conceptions du haut potentiel différentes (même parfois divergentes) est complexe et que cette situation empêche notamment les comparaisons entre études.

Parmi les différentes techniques d'évaluation, l'auteur rapporte différentes études qui listent les outils les plus fréquemment utilisés (Les échelles de Wechsler et l'échelle Stanford-Binet aux États-Unis, la troisième version de l'échelle d'intelligence de Wechsler pour enfant WISC-III en France). De plus, au moment de l'article, il n'y a pas de consensus dans les recherches menées sur le seuil de QI pertinent pour le diagnostic du haut potentiel. On trouve des variations entre 120 et 140 et parfois une grande tolérance par rapport au seuil du QI si l'enfant manifeste soit un QI verbal soit un QI de performance supérieur à ce seuil. L'auteur précise aussi que l'étalonnage de la WISC-III (Wechsler, 1996) a été réalisé sur 1120 enfants entre 6 et 16 ans mais dont peu d'entre eux avaient montré des performances extrêmes. Selon l'auteur, la WISC-III bien adaptée à la population générale est peu sensible aux valeurs extrêmes or la qualité du diagnostic dépend de la sensibilité du test. Un autre problème est également d'arriver à différencier

les hauts potentiels entre eux. En effet, la WISC-III présente certains effets plafonds pour certains subtests « Performance » ce qui entraîne une étendue non suffisante des scores possibles dans les valeurs extrême pour arriver à différencier les hauts potentiels avec précision. L'auteur n'est pas convaincu des solutions utilisées pour pallier ce problème comme comparer aux normes relatives à des enfants plus âgés ou prendre des normes anciennes susceptibles d'être invalidées par l'effet Flynn.

En investiguant les recherches sur la fidélité du QI mesuré par la WISC pour les hauts potentiels, l'auteur a trouvé des corrélations test-retest sont assez faibles variant de 0,33 à 0,66 soit pour les subtests soit pour l'échelle totale même avec des délais courts. Si l'on regarde parmi les enfants tout-venant et même avec des délais plus longs on a des valeurs de fidélité aux environs de 0,80. L'auteur met dès lors en doute la stabilité du QI mesuré par la WISC-III pour les hauts potentiels. McNemar (1969, cité par Caroff, 2004) explique la sous-estimation de la corrélation entre les 2 mesures par un biais statistique du fait de l'homogénéité du groupe étudié.

Ensuite, en ce qui concerne la validité du QI mesuré par la WISC pour les hauts potentiels. Il a classé les recherches en deux types. Celles relatives à la structure interne et les autres relatives à des critères de validation. La structure factorielle indique, pour la population générale, 2 facteurs stables (CV = compréhension verbale, OP = organisation perceptive), avec un troisième facteur parfois repéré (AC= attention, concentration) mais bien que les 2 premiers facteurs soient retrouvés chez les sujets à haut potentiel, le 3^e facteur lui, ne correspond pas. Certains n'ont qu'un facteur et d'autres 4 (CV, OP, AC, VT= vitesse de traitement).

Caroff (2004) s'est tourné vers les MRI du fait que le niveau de compétence, avec ces modèles, est estimé grâce au profil des réponses aux items et qu'il s'agit d'une évaluation qui ne dépend plus d'un échantillon de référence. Les propriétés de ces modèles pourraient permettre la construction d'une échelle adaptée aux hauts potentiels. Un problème relatif à l'erreur de mesure pourrait néanmoins survenir pour le développement d'une échelle contenant un large panel d'items extrêmes nécessaires à l'identification du haut potentiel. De plus, ces modèles permettent aussi comme présenté avant (cf. [1.5.4](#)) des testing adaptatifs et donc une évaluation avec nettement moins d'items que dans des échelles classiques.

Preckel et Thiemann (2001, 2003 ; cités par Caroff, 2004) ont développé deux versions d'un test de matrices inspiré de Raven (papier-crayon et sur ordinateur) afin de diagnostiquer des hauts potentiels. Le modèle de Rasch s'ajustait parfaitement aux données et pour les deux versions, il y avait une bonne consistance interne ($\alpha = 0,80$).

La corrélation entre le QI des participants (identifié avant) et le score obtenu dans la version « papier-crayon » et était de 0,56.

Caroff (2004) conclut par l'intérêt de poursuivre ce type de recherches où l'on applique les modèles MRI aux tests pertinents pour le diagnostic des hauts potentiels ou encore pour construire de nouvelles échelles de mesure.

1.5.6 Unidimensionnalité de l'échelle d'engagement au travail d'Utrecht

1.5.6.1 Introduction

Les développements en psychologie positive ont remis en question l'approche de certaines recherches et notamment en psychologie du travail. Au lieu de se concentrer sur la notion de déficits, les ressources humaines deviennent un objectif d'étude à part entière comme le concept d'engagement au travail. Saks (2006, p. 163, cité par De Bruin, Hill, Henn, & Muller, 2013) a mis en évidence que l'engagement est une « *construction significative* » sur laquelle la recherche devrait se concentrer.

De Bruin et al. (2013) souhaitent vérifier la dimensionnalité de l'échelle d'engagement au travail d'Utrecht (UWES-17) utilisée pour mesurer l'engagement au travail à l'aide de 3 sous-échelles. La première de sous-échelles est relative à la vigueur (en lien avec le niveau d'énergie élevé - 6 items) ; la deuxième porte sur le dévouement (en lien avec le sens - 5 items) et la troisième concerne l'absorption par le travail (en lien avec le niveau de concentration – 6 items). Cette échelle a été validée et est disponible en 23 langues avec des versions pour les étudiants. Cependant, bien que son usage soit répandu, selon les auteurs, l'unidimensionnalité de cette échelle n'est pas clairement établie par la recherche. Le psychologue doit-il interpréter les 3 scores des sous-échelles comme une seule dimension ? Plusieurs études vont dans le sens d'une seule dimension. Mais quelques-unes arrivent à 3 dimensions en éliminant certains items.

1.5.6.2 Analyse Rasch et résultats

Les 2429 sujets étaient des employés d'une société Sud-africaine qui a fourni toutes les informations nécessaires au bon déroulement de l'étude. Ils ont répondu via un questionnaire en ligne. Chacun des items propose 7 catégories de réponses (scores de 0 à 6). Les coefficients alpha de Cronbach provenant d'études internationales donnent pour les 3 sous-échelles des valeurs entre 0,69 et 0,90.

De Bruin et al. (2013) ont mené une analyse Rasch. Celle-ci a permis d'examiner la position des seuils de chacun des items. 15 items parmi les 17 avaient des seuils désordonnés, avec les catégories de réponse 2, 3 et 4 n'émergeant pas (cf. [annexe 8.1.](#)) Les auteurs ont alors procédé à un recodage afin de passer à 5 catégories de réponses. Les

catégories 0 et 1 n'ont pas été modifiées. La catégorie 2 a été associée avec la catégorie 1 et donc transformée en niveau 1. Les catégories 3 et 4 ont été transformées en la nouvelle catégorie 2. Ensuite les catégories 5 et 6 ont été transformées en nouvelles catégories de niveau 3 et 4. Suite à ce recodage, l'indice de fidélité de séparation est passé de 0,91 à 0,92.

La p-value de l'ajustement global était significative ($< 0,0001$) et 3 items (it. 6, 14, 16) de la sous-échelle absorption présentaient des résidus non ajustés ($> 2,5$). La suppression de ces items a amélioré partiellement l'ajustement global. Différents items parmi les 14 restants (notamment le 4 et le 5) présentaient des résidus fortement négatifs. Certains de ces items présentaient également un misfit significatif (it. 5, 17) mais pour la classe d'aptitude la plus basse (sur 9 classes) qui ne concernait que très peu de sujets. Pour toutes les autres classes, il y avait adéquation avec le score attendu.

Le ciblage de l'échantillon n'était pas adéquat du fait d'une moyenne à 1,284 logits avec une grande partie des sujets avec un niveau d'aptitude à 2 logits et un sommet de la courbe d'information situé à -0,5 logits.

1.5.6.3 Conclusion

Pour les auteurs, les informations relatives aux 3 items « problématiques », du fait qu'ils sont issus de la même sous-échelle, suggèrent que cette sous-échelle n'est pas complètement une composante de l'engagement professionnel. La corrélation entre les items 4 et 5 conduit à une légère surévaluation de la fidélité de l'échelle. De plus, les auteurs suggèrent d'envisager la création d'items plus « difficiles » sur le trait pour améliorer le ciblage, l'échelle étant moins précise pour les niveaux élevés.

Les auteurs sont conscients des limites de leurs études et le fait qu'une reproduction de cette analyse avec un autre échantillon serait indiquée. De plus, les résultats dans différents groupes démographiques n'ont pas été étudiés.

En conclusion, ces diverses recherches mettent en évidence la pertinence de l'analyse Rasch dans les réductions d'échelles sans perte de qualité psychométrique, de

1.6 Utilité d'un outil d'analyse R

1.6.1 Le langage R

Qu'est-ce que R ?

R est un langage « open source » dérivé du langage S (payant) qui a été développé par John Chambers. A ses débuts, en 1976, il s'agissait d'un environnement d'analyses

statistiques. Il contenait alors des librairies en Fortran. En 1988, lors de la version 3, le système a été réécrit à l'aide du langage C. R apparaît en 1991, crée par Ross Ihaka and Robert Gentleman en Nouvelle Zélande. Ils publient en 1996 leur expérience sur leur développement de R dans le « Journal of Computation and Graphical Statistics ». Ils ont créé deux listes d'échanges d'e-mail, l'une appelée « R-help » qui liste les questions sur R et l'autre « R-devel » pour les développeurs de R. Ils ont alors constitué le « R core group » en 1997, constitué de ces personnes, qui contrôle le code source de R et qui sont les seuls à pouvoir le modifier. Ensuite, des personnes ne faisant pas partie de ces groupes ont commencé à suggérer des changements et ont finalement été intégrées dans le « R core group ».

Un des aspects positifs de R en plus d'être « open source » est qu'il peut être utilisé avec tous les systèmes d'exploitation comme Mac, Windows, Linux.

Actuellement, ses fonctionnalités sont réparties dans différents packages (ou librairies). Ce qui permet de n'installer que ce dont les utilisateurs ont besoin spécifiquement. Les développeurs convaincus vous diront qu'il y a beaucoup plus de possibilités que dans les programmes payants de statistiques. Avec en plus le fait que l'utilisateur garde un contrôle sur ce qu'il produit à partir des packages et la possibilité de réaliser de nouveaux développements répondant à ses besoins spécifiques. Il y a de plus une grande communauté de développeurs et des sites comme [stackoverflow](https://stackoverflow.com/)⁹ où l'on peut poser des questions.

La philosophie prégnante est que les utilisateurs sont libres de faire des modifications, des développements et qu'ils peuvent les partager ensuite afin que tout le monde puisse les utiliser. R se télécharge sur CRAN¹⁰ où se trouvaient environ 10 000 packages en 2017.

1.6.2 Outil open source d'analyse R

Kubinger (2000), cite une réplique célèbre de Jürgen Rost : «Was ist aus dem Rasch-Modell geworden? ». En effet, il voulait de cette façon à la fois souligner l'importance du modèle de Rasch, « *apporter la preuve de la nécessité* » de ce modèle et attirer l'attention sur son utilisation encore peu fréquente. Ce constat peut s'expliquer du fait du coût élevé de la plupart des logiciels permettant d'utiliser le modèle de Rasch mais aussi du fait de la nécessité d'être formé pour pouvoir les utiliser que ce soit pour la préparation du jeu de données ou les connaissances nécessaires (techniques et

⁹ <https://stackoverflow.com/>

¹⁰ <https://cran.r-project.org/>

conceptuelles) à leur utilisation et à l'interprétation des résultats. Cela a pour conséquence que seuls quelques experts utilisent le modèle de Rasch et entraîne que ce modèle est peu diffusé et utilisé.

Kubinger (2005) présente l'utilité du modèle de Rasch notamment pour l'étalonnage des tests et souligne que dans les utilisations du modèle en psychologie, les critères d'unidimensionnalité ne sont pas toujours appliqués et soulève certaines situations de significativité qui peuvent être difficiles à appréhender.

Ce module open source d'analyse Rasch en R a donc bien tout son sens. Non seulement, il a l'avantage de lever le problème du coût mais il offre aussi différentes possibilités (cf. [chapitre 2](#)) d'analyse Rasch de façon relativement accessible.

1.6.3 [RMarkdown](#)

Rmarkdown permet d'analyser, partager et reproduire. Il fournit un cadre de création pour la science des données.

RMarkdown permet d'enregistrer et d'exécuter du code et ensuite, de transformer les analyses réalisées avec R en documents, rapports, présentations et tableaux de bord de haute qualité pouvant être partagés.

Les documents RMarkdown sont entièrement reproductibles et prennent en charge des dizaines de formats de sortie statiques et dynamiques

Une interface bloc-notes (notebook) productive est disponible pour associer du texte narratif et du code (« chunks ») afin de produire une sortie au format élégant. Il est possible d'utiliser plusieurs langages, notamment R.

Cette vidéo d'une minute offre un aperçu rapide de ce qui est possible avec RMarkdown¹¹:

<https://vimeo.com/178485416>

¹¹ Le site de référence est <https://rmarkdown.rstudio.com/lesson-1.html>

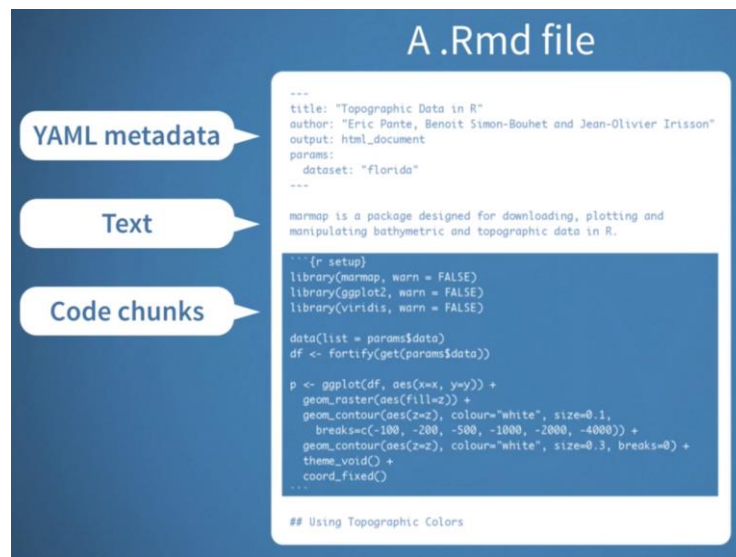


Figure 10 - Capture d'écran de la vidéo mentionnée - RMarkdown. (s.d.). [Vidéo]

2 Développement d'un module d'analyse Rasch en R pour l'interprétation des mesures en psychologie

2.1 Introduction

Afin de répondre au cahier de charges (cf. [annexe 9.1.](#)), nous avons mis en place une solution basée sur R au travers de notebooks R. Ce chapitre parcourt les aménagements nécessaires ainsi que les difficultés rencontrées lors de la construction du module d'analyse R. Afin de faciliter la compréhension, des illustrations spécifiquement créés pour ce mémoire ou captures d'écran de séquences de code ont été ajoutées.

Les séquences de code sont en anglais de même que les fonctions de R citées dans ce mémoire, celles-ci étant en caractères gras.

2.2 Architecture de la solution et explication des choix

Cette section présente l'architecture de la solution élaborée et explicite les choix techniques posés. La solution globale est formée des modules suivants :

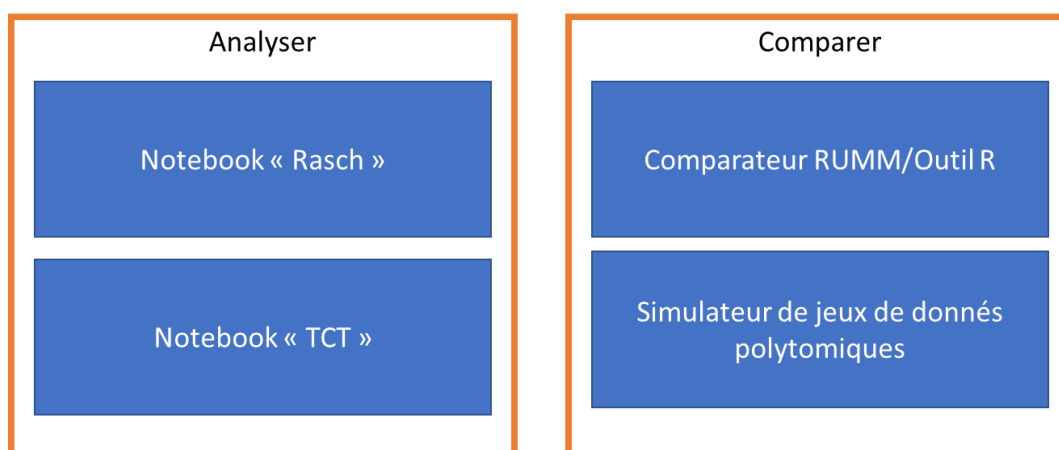


Figure 11 - Modules formant la solution

2.2.1 Notebooks d'Analyse

Les notebooks, rédigé en RMarkdown, sont configuré via un fichier annexe (*params_xxx.R*) contenant les paramètres pour l'étude à mener. L'exécution du notebook produit une sortie finale en HTML dotée d'une table des matières permettant d'accéder rapidement à la section souhaitée.

Il existe deux notebooks clé : le notebook d'analyse Rasch et le notebook d'analyse TCT

2.2.1.1 Notebook Analyse Rasch

Ce notebook contient l'analyse Rasch d'un jeu de données et exploite principalement le package **eRm**. Les défis associés sont examinés en détail ultérieurement (cf. [2.3.1](#)).

2.2.1.2 Notebook Analyse TCT

Ce notebook contient l'analyse en théorie classique des tests et exploite principalement les packages **matrixstats**, **psych** et **car**.

Ses fonctionnalités sont :

- Calcul du totalscore et histogramme associé
- Calcul de la difficulté des items et histogramme associé
- Calcul de la fiabilité des items et histogramme associé
- Calcul de la variance des items et histogramme associé
- Calcul de la discrimination des items et histogramme associé :

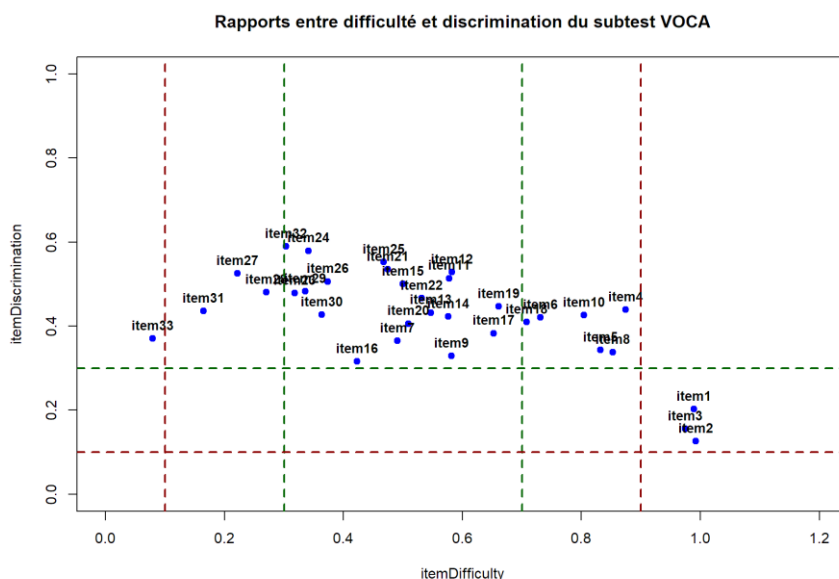


Figure 12 - Visualisation TCC : rapport entre difficulté et discrimination des items du subtest Vocabulaire de la WAIS-III

- Calcul de l'alpha de Cronbach
- Statistiques descriptives du totalscore

- Calcul de SEM
- Divers tests et visualisations pour s'assurer de l'homogénéité de la variance avant de faire une analyse de celle-ci :
 - Levene
 - Brown Forsythe
 - Test de normalité
 - Fit du totalscore sur base des facteurs DIF
 - Visualisations des résidus
 - Boxplot divers sur le fit du score sur base des facteurs DIF
 - Histogrammes des totalscore par facteur DIF

2.2.2 Comparateur des résultats RUMM / R

Cet outil permet de représenter la localisation des items fournis par le notebook d'analyse Rasch à celle fournie par RUMM. Les données des deux outils sont stockées dans des fichiers Excel et représentées graphiquement afin de se faire une idée de la correspondance des valeurs. Une fonction de comparaison a été développée et est simple d'utilisation :

```
plotComparison(myVocaRUMMData, myVocaRData, main="Location RUMM vs Location eRm Voca WAIS III")
```

Figure 13 - Appel de la méthode de comparaison RUMM2030 / R

On obtient alors un graphe représentant les valeurs RUMM avec une étendue dans laquelle les valeurs sont acceptables. Les points sont les valeurs calculées avec le Notebook Rasch.

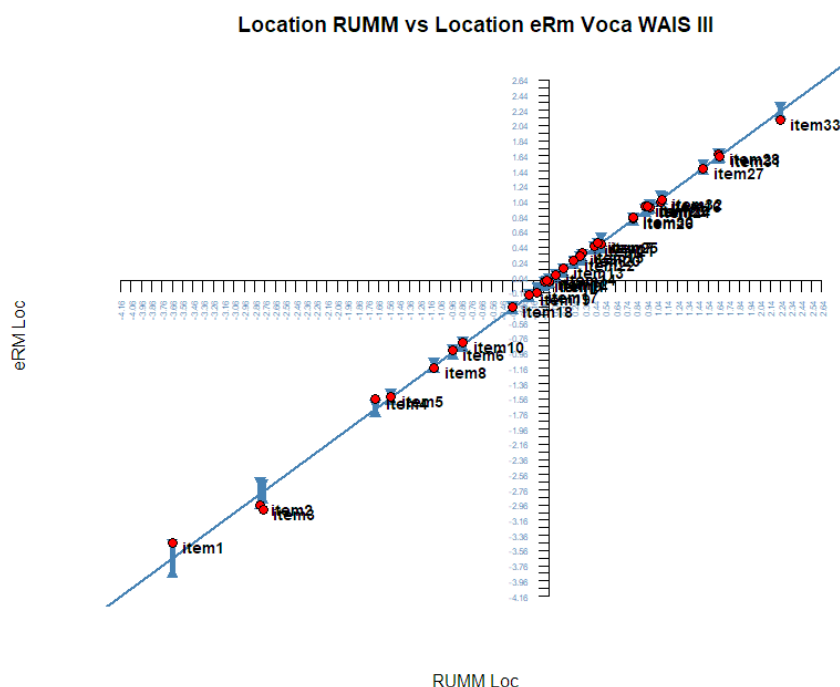


Figure 14 - Exemple de comparaison RUMM / Outil R

2.2.3 Simulateur de jeu de données polytomiques

Afin de pouvoir se faire une idée des résultats fournis par une analyse, il est utile de disposer de jeux de données simulés sur lesquels effectuer une analyse avec les outils, plutôt que de comparer les outils entre-eux. On dispose ainsi d'une référence connue avec laquelle on peut comparer les outils. L'outil a été testé en polytomique avec PCM et on obtient le résultat suivant, où l'on peut constater que le fit est assez correct. L'idée est maintenant de tester le fit que donne RUMM2030 afin de le comparer.

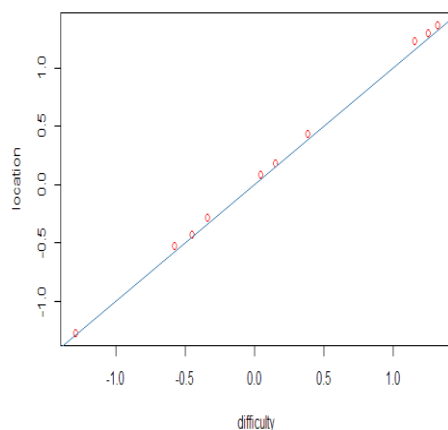


Figure 15 - Fit PCM sur données simulées

2.2.4 RStudio

L'édition d'un notebook et la génération du rapport associé sont effectués au sein de RStudio, l'environnement de développement intégré dédié à R. Entre autres avantages, il permet d'exécuter le rapport en donnant accès aux diverses données calculées, ce qui a permis de détecter et corriger les anomalies plus rapidement.

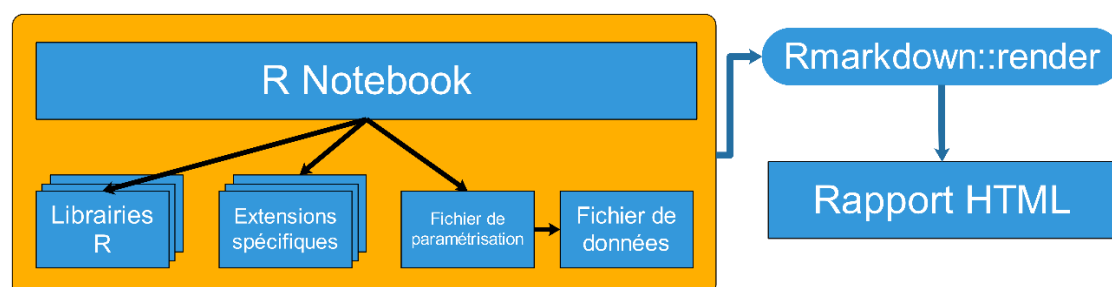


Figure 16 - Aperçu de la solution permettant le développement du module d'analyse Rasch en R

2.2.5 Package eRm

La solution exploite le package **eRm** pour le calcul de modèles de Rasch.

Ce package n'est pas le seul disponible à cette fin, d'autres packages existent. Cependant, il apparaît comme étant celui qui permet d'aborder les divers angles à considérer pour ce mémoire.

En effet, d'autres packages, tels que **mirt**, **ltm**, **lordif**, **TAM** (Robinson, Jonhson, Walton, & MacDermind, 2019), peuvent utiliser d'autres méthodes pour établir les modèles mais sont moins documentés et d'un abord plus complexe. L'objectif étant de couvrir les différentes demandes du cahier des charges, il a été choisi d'exploiter les possibilités de **eRm**, qui restait maîtrisable. Il serait éventuellement utile de prolonger le travail en complétant et/ou étendant les différents aspects à l'aide d'autres packages également centrés MRI.

Le package **eRm** calcule les modèles au travers de l'approche ~~Person Separation Index~~ Conditional Maximum Likelihood (CML).

Cette méthode n'est pas la plus précise de toutes les méthodes mais il faut bien considérer que lors de l'établissement d'une échelle de mesure, elle est normalement suffisante pour situer si les items sont pertinents ou non.

Il faut cependant signaler qu'une mise à jour récente (fin mars 2020) du package eRm vient d'être réalisée alors que le module R était déjà bien avancé. Il faut donc en tenir compte lors de l'installation du module et de R afin de pouvoir utiliser le module.

2.2.6 Performance

Au niveau des temps de calcul, il apparaît également que lorsqu'il y a des données manquantes, l'algorithme peut prendre du temps. Afin de le réduire, il s'est avéré utile d'examiner les possibilités de parallélisation de calcul dans R.

L'introduction du package **doParallel**, qui permet d'utiliser plusieurs cœurs de processeur pour les boucles, a permis de réduire le temps de calcul.

2.2.7 Données

Les données brutes CSV ont été importées dans des fichiers de données au format Excel afin de faciliter leur visualisation préalable et aussi d'introduire facilement des cellules calculées. Ceci permet de limiter les manipulations de données initiales via du code R. Cela facilite également l'identification de l'étendue à utiliser pour identifier les items, les personnes et les données DIF dans le fichier de paramétrisation.

2.2.8 Librairies

Le notebook utilise diverses librairies R (dont **eRm** et **doParallel**) afin de manipuler, calculer, ou sortir des données (cf. [Annexe 5- 10.30.](#)).

2.2.9 Structure d'un Notebook

Un notebook est structuré à l'aide de niveaux de titre RMarkdown (#, ##, ###) qui se retrouvent dans la table des matières du rapport HTML final, facilitant ainsi la navigation. Le notebook est constitué de texte et de chunks. Les chunks sont nommés ce qui permet de les retrouver dans le code au niveau de l'environnement de développement.

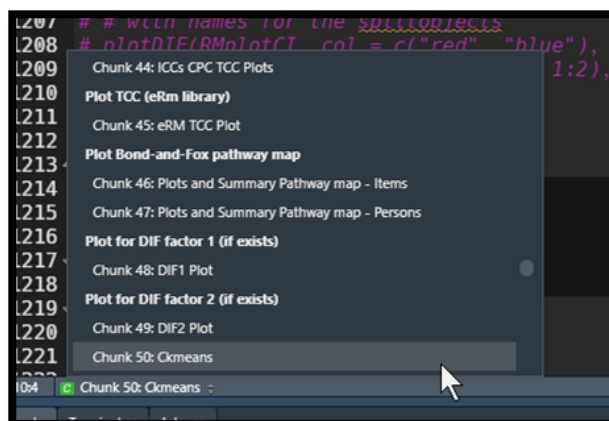


Figure 17 – Niveaux de titre (en gras) et chunks de code

2.2.10 Procédure de production d'un rapport d'analyse

La transformation du notebook en rapport HTML se fait au travers de la librairie R **knitr**, lancée depuis la console R dans RStudio plutôt que de cliquer sur le bouton « knitr html ». En effet, ce bouton ne permet pas d'examiner les données produites.



Figure 18- bouton knitr to HTML

Cependant, il est nécessaire au bon fonctionnement des chunks. Il faut donc l'utiliser au départ. Par contre, la commande lancée depuis la console R de RStudio garde les données dans l'environnement.

```
rmarkdown::render("Polytomique/PolytomicDataPrepAndAnalysis.Rmd")
```

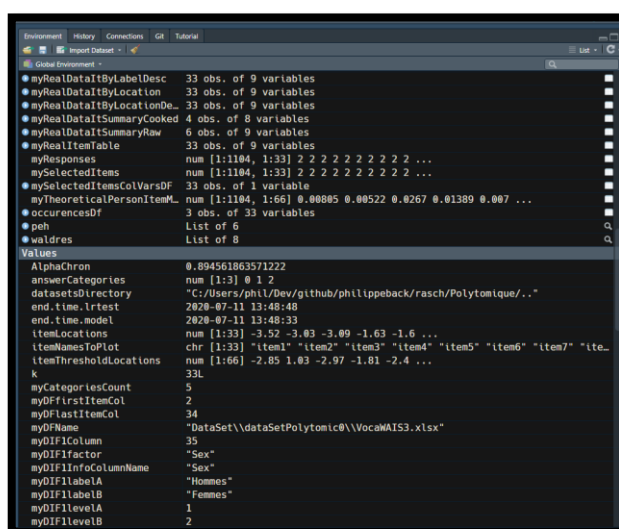


Figure 19 - Données conservées dans l'environnement

RStudio permettant d'exécuter un chunk indépendant, cette démarche a permis de travailler itérativement.

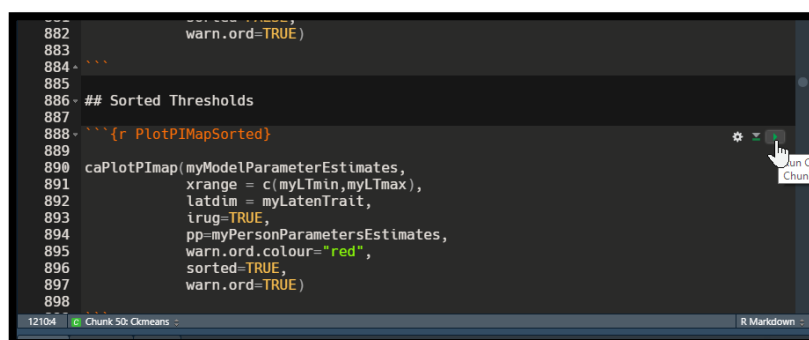


Figure 20 - exécution isolée d'un chunk

Le rapport final est stocké dans le répertoire du notebook, au format HTML.

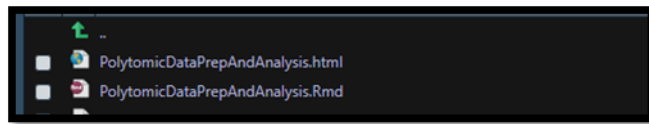


Figure 21 - Notebook (.Rmd) et rapport (.html)

Toutes les figures sont embarquées dans le fichier HTML, qui peut ainsi être facilement envoyé par email ou autre (il convient de le compresser au format ZIP pour réduire un peu sa taille).

On peut facilement ouvrir un rapport depuis RStudio.

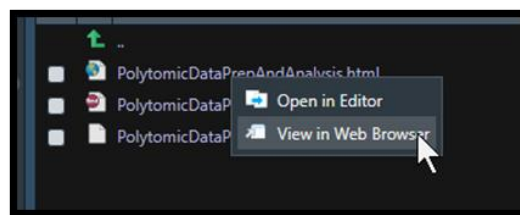


Figure 22- Accès au rapport

Le rapport s'ouvre alors dans le navigateur par défaut et on peut accéder facilement à ses éléments via la table des matières dynamique, ce qui la navigation efficace.

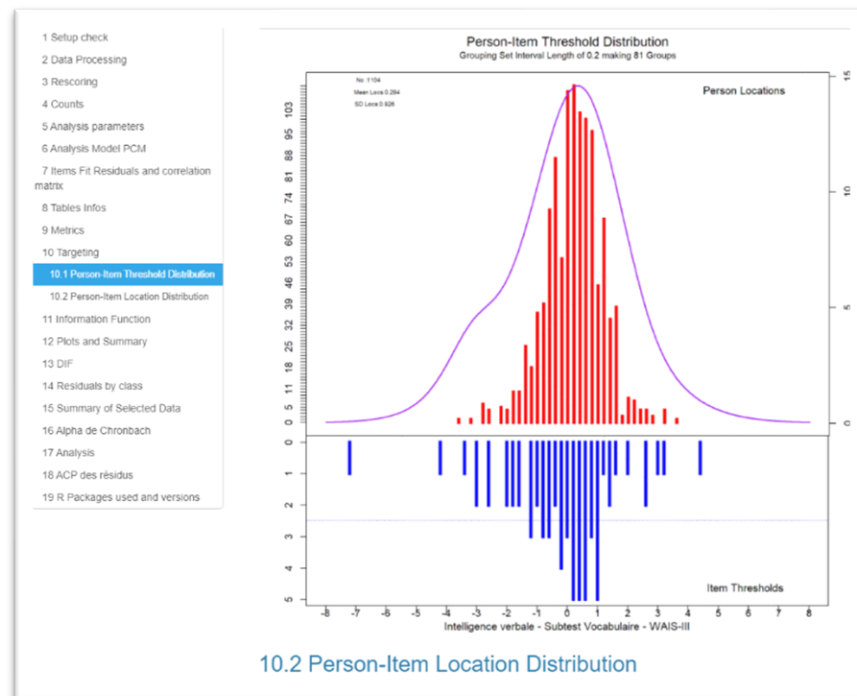


Figure 23 - Rapport au format HTML

2.2.11 Visualisations graphiques

Les visualisations graphiques sont produites au travers de la librairie de base **graphics** de R. Des packages plus avancés tels que **ggplot2** n'ont pas été utilisés. La raison en est que la librairie **eRm** produit ses visualisations au travers de **graphics** et donc, il était pertinent de garder la même option, eu égard au fait que des extensions aux fonctions existantes ont été apportées. De plus, **graphics** a permis d'atteindre les objectifs de visualisation.

2.3 Conception et défis rencontrés

Un certain nombre de défis lors de la mise en œuvre de la solution ont été rencontrés.

Voici leur description ci-après avec la réponse qui leur a été apportée.

Une description détaillée de l'implémentation se trouve en annexe.

2.3.1 Notebook Rasch

Le tableau ci-après identifie les défis rencontrés et la réponse apportée en termes d'implémentation. Des sections en annexe détaillent certains des points clé.

Tableau 3- Défis de conception et solutions apportées (outil Rasch)

<u>Défi</u>	<u>Réponse</u>
Sortie de tableaux dans Rmarkdown	Utilisation de knitr ::kable et knitr :kextra .
Configuration du notebook facile pour l'utilisateur	Utilisation de fichiers de paramètres.
Extension de packages	Extension d'eRm avec les fonctions qui ont été nécessaires afin de produire une version spécifique de caPlotPImap , caPlotCC et caPlotPersonItemHist .
Performance	Utilisation de doParallel .
Gestion des valeurs manquantes et conversion en NA	Développement d'une fonction caReadDataset spécifique pour le chargement de jeux de données depuis des fichiers Excel.
Sélection de la plage de données souhaitée	Convention sur les données.
Elimination des personnes et items non souhaités	Déclaration de ces informations au sein du fichier de paramètres.
Rescoring	◆ Activation du rescoring au travers des paramètres et fourniture d'un tableau de correspondance. ◆ Utilisation de la fonction mapValues du package plyr.
Comptage des éléments	Développement de fonctions spécifiques pour le comptage.
Calcul du fit et des thresholds sur base de l'algorithme choisi	◆ Spécification du type de modèle dans le fichier de paramètres. ◆ Support des modèles RSM et PCM. ◆ Support du modèle RM. ◆ Utilisation selon la configuration des fonctions RM, RSM et PCM selon la configuration et de la fonction thresholds de eRm. ◆ Chronométrage de la durée du fit.

Evaluation de l'adéquation du modèle aux données modélisées et d'aider à la critique du modèle sur une base rationnelle	♦ Utilisation de LRtest de eRm, spécification d'un vecteur de split des personnes adapté. ♦ Interprétation de la pValue du test. ♦ Graphes goodness of fit plotGOF avec bandes et ellipsoïdes de confiance. ♦ Waldtest.
Evaluation de l'unidimensionalité en PCM	Utilisation de Martin-Löf avec MLof .
Extraction des paramètres des personnes	Utilisation de person.parameter .
Extraction des statistiques de fit pour items et personnes	Utilisation de itemfit et personfit .
Calcul des locations des items	Extraction programmée des valeurs de la threshtable depuis le résultat du fit. Assemblage d'une dataframe adaptée.
Extraction des résidus du fit sur les personnes	Utilisation de residuals sur le résultat de person.parameters .
Evaluation du niveau du trait latent des personnes	Utilisation de coef .
Confection du tableau des items	Confection d'une ou plusieurs dataframe permettant de présenter les informations relatives aux items.
Traitement des résidus	Vérification des résidus avec un qqPlot du package car. Etablissement d'une matrice de corrélation des items basée sur les résidus avec légende colorée selon le niveau de souci à l'aide de corrplot. Sortie d'une table de corrélation des résidus.
Présentation des informations des items et des personnes, en ce y compris les statistiques de fit, l'erreur standard, les	Extraction des informations dans des dataframes depuis les résultats de fit PCM, des résultats de person.parameters , du

locations et les locations centralisées pour les items	calcul des locations, du centrage des valeurs. Présentation de ces dataframes dans des tables kable .
Calcul des métriques	Calcul du PSI, du SSD, MSE, G/SeparationRatio et nombre de strates à l'aide de formules spécifiques et utilisation de SepRel de eRm .
Targeting – Person-Item Threshold Distribution	Implémentation complète d'une visualisation spécifique exploitable au travers de la fonction caPersonItemHist qui est une extension de eRm . Cette extension fournit une visualisation similaire à une visualisation de RUMM2030, intégrant histogrammes personnes et items mis en regard, affichage de statistiques, intégration de la courbe d'information.
Targeting – Person-Item Location Distribution	Utilisation paramétrée de la fonction caPlotPersonItemHist pour présenter les locations en lieu et place des seuils.
Fonction d'information	Utilisation de plotINFO de eRm .
Présentation résumés d'une carte des personnes et des items, avec seuils ordonnés ou non	Ecriture d'une version améliorée de plotPIMap appelée caPlotPIMap qui présente les items, seuils et axes de manière plus élégante et lisible que la version de base fournie par eRm .
Courbes caractéristiques d'item en polytomique en ce y compris la courbe sigmoïde avec positionnement éventuel de points correspondant au résumé des DIF	Développement spécifique d'une fonction caPlotICC sur base de plotICC de eRm car il n'y avait pas de support pour les courbes sigmoïdes en polytomique dans eRm . Cette fonction est complexe et ajoute la courbe sigmoïde de l'item, les statistiques, des axes calibrés au niveau des

	graduations, l’affichage du résumé d’items DIF. La mise en œuvre fut complexe eu égard à la logique de calcul et d’organisation de la visualisation de courbes diverses, de la présentation de seuils, de la génération de graphismes multiples dans un seul appel de la fonction.
Visualisation du TCC Global en polytomique	L’affichage de la courbe globale des scores attendus n’est pas disponible. Mise en place d’un plot spécifique en extrayant les données pertinents du fit.
Identification des misfits en termes d’items et de personnes	◆ Utilisation du Bond and Fox Pathway Map avec la fonction plotPWmap de eRm pour les items set les personnes ◆ Rédaction d’instructions d’interprétation.
Support du DIF	◆ Utilisation de facteurs DIF (DIF1,DIF2, cluster) avec l’option splitcr de LRtest de eRm pour sortir des modèles partiels. Visualisation avec plotDIF de eRm. ◆ Visualisation des résidus par facteur de DIF avec qqPlot de car. ◆ Clustering en catégories de compétence à l’aide d’un algorithme Kmeans, à savoir Ckmeans.1d.dp
Statistiques descriptives du jeu de données sélectionné	Utilisation de describe de psych pour chaque item et assemblage dans une dataframe et visualisation avec kable .
Calcul de l’Alpha de Cronbach	◆ Calcul via des formules spécifiques.

	◆ Utilisation de colVars de matrixstats pour le calcul des variances, et rowSums et var de base pour la somme des lignes. ◆ Affichage de la valeur.
Visualisation des items et de personnes triés selon divers critères et statistiques descriptives associées	Utilisation de mixeorder et order sur des dataframes ainsi que de describe de psych pour les statistiques descriptives.
Analyse en composantes principales des résidus	◆ Utilisation de pca de mdatools sur les résidus, visualisation des résidus à l'aide de plot.pca de mdatools, ◆ visualisation des valeurs propres et tracé de lignes de référence avec plot. ◆ Identification des personnes extrêmes qui influencent les composants principaux avec categorize de mdatools. ◆ Visualisation tabulaire des personnes extrêmes.

2.4 Versions

La liste des packages utilisés et leur version ainsi que les versions de R et RStudio sont placées en [annexe 5-10.30](#).

2.5 Sorties pour différents jeux de données et comparaisons avec RUMM2030

Le comparateur est disponible pour comparer les valeurs des locations des items et voir si les résultats sont cohérents entre ce que RUMM2030 donne et ce que l'outil basé sur eRm donne.

Le principe est de représenter les données des locations des items telles que fournies par chaque logiciel et de tracer les erreurs standard des seuils sortis par RUMM et de voir où le point issu de l'outil basé sur eRm se situe.

2.5.1 Correspondance

Si le point est « dans les clous », le module est au moins aussi bon en termes de résultat.

C'est le cas en ce qui concerne l'étude du jeu de données Voca WAISIII :

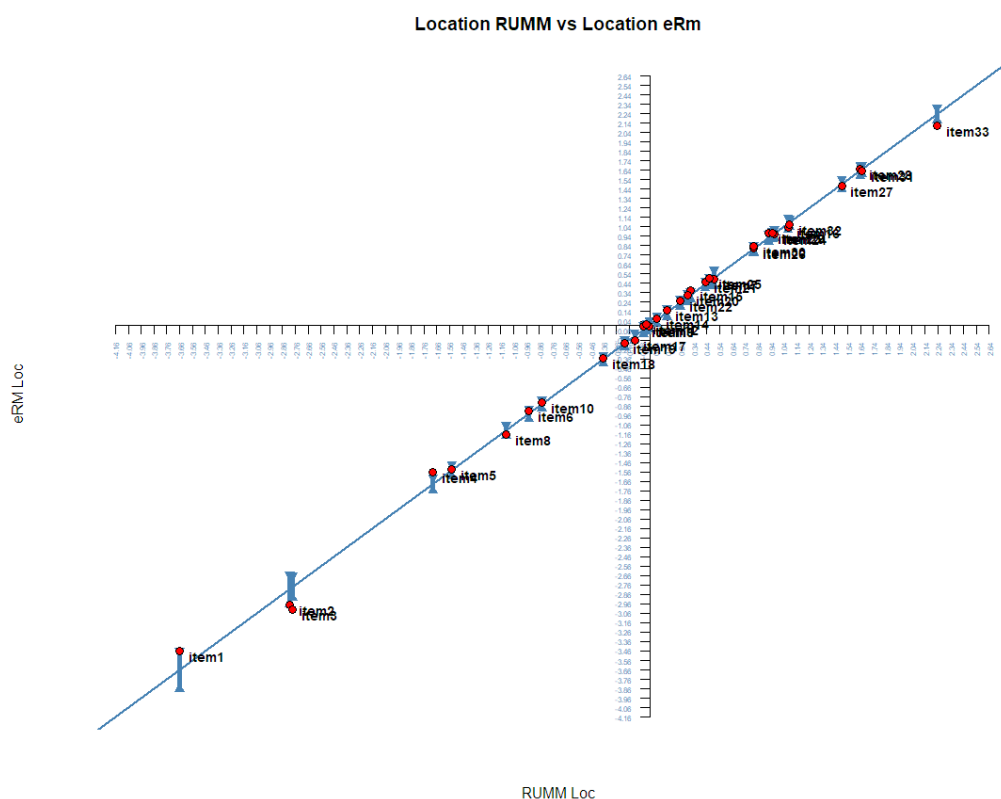


Figure 24 - Comparaison RUMM2030 / Outil R

En termes de statistiques descriptives sur les items, nous obtenons :

RUMM2030

```
> psych::describe(myVocaRUMMData$Location)
```

```
> psych::describe(myVocaRUMMData$Location)
vars n mean sd median trimmed mad min max range skew kurtosis se
X1 1 33 0 1.34 0.24 0.14 1.02 -3.66 2.24 5.9 -0.94 0.5 0.23
Warning messages:
```

Figure 25- Vvaleurs RUMM2030

R/eRm

```
> psych::describe(myVocaRData$Location)
```

```
> psych::describe(myVocaRData$Location)
vars n mean sd median trimmed mad min max range skew kurtosis se
X1 1 33 -0.07 1.34 0.2 0.08 1.05 -3.52 2.05 5.57 -0.98 0.44 0.23
>
```

Figure 26 - Vvaleurs eRm

2.5.2 Non correspondance

En cas de non correspondance, il convient de s'interroger sur le résultat obtenu.

En effet, pour le jeu de données **depress**, on note un décalage sur la droite de l'échelle :

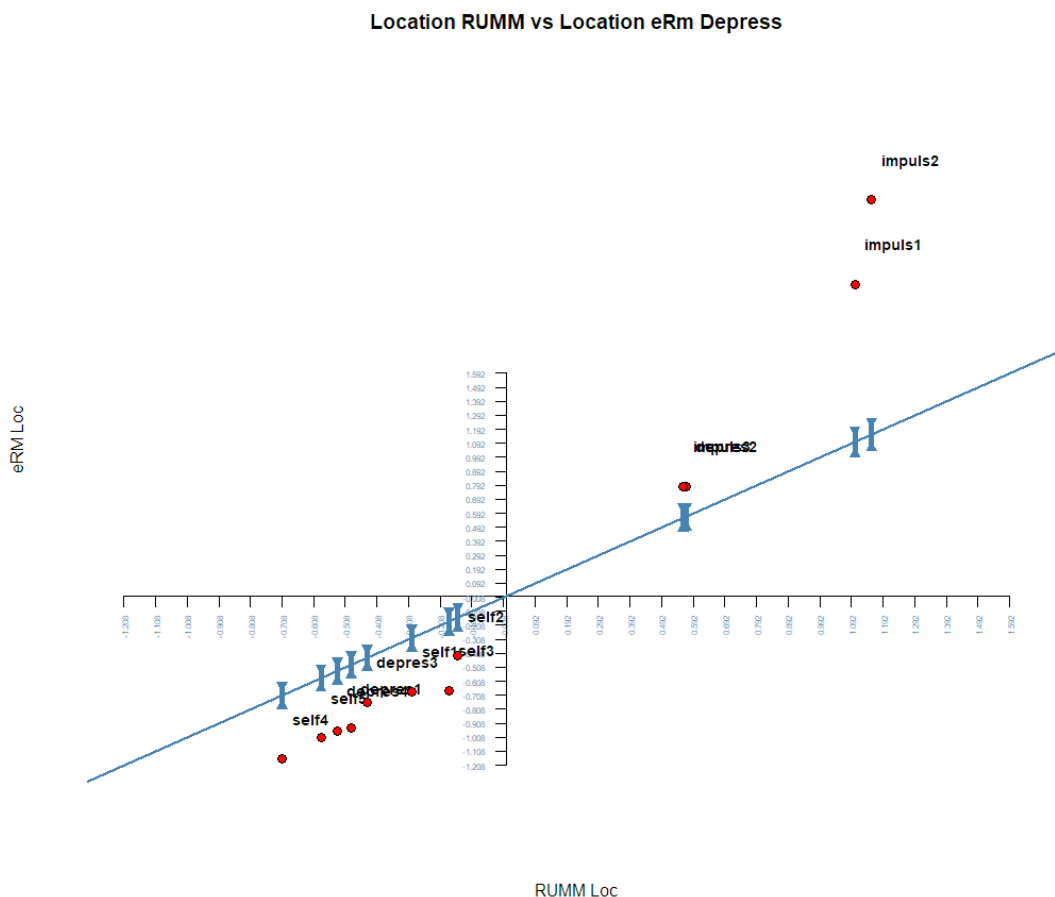


Figure 27 - RUMM vs Outil pour jeu de données depress

En termes de statistiques descriptives sur les items, nous obtenons :

```
> psych::describe(myDepressRUMMData$Location)
vars n mean sd median trimmed mad min max range skew kurtosis se
X1 1 12 0 0.67 -0.24 -0.04 0.47 -0.71 1.16 1.86 0.65 -1.31 0.19
> psych::describe(myDepressRData$Location)
vars n mean sd median trimmed mad min max range skew kurtosis se
X1 1 12 0.92 1.48 0.19 0.72 0.52 -0.28 4.08 4.36 1.07 -0.44 0.43
```

Figure 28- Valeurs RUMM2030 et R – Stats items

Cependant, il convient de noter que les locations respectent l'ordre des locations, elles sont « juste » poussées plus loin vers l'avant ou l'arrière et non pas aléatoirement dans diverses directions. La cohérence générale de l'échelle est conservée entre les deux logiciels. Par contre, on observe que les points ne rentrent pas dans l'intervalle de l'erreur standard.

On peut soupçonner des tels effets de déséquilibre dans l'échelle eu égard au déséquilibre d'items sur la partie gauche de l'échelle par rapport à la partie droite, ce qui peut influencer le fit au détriment de la partie droite.

Ceci étant, on observe que le résultat « Goodness of Fit » (GOF) indique que ces items sont « assez particuliers ».

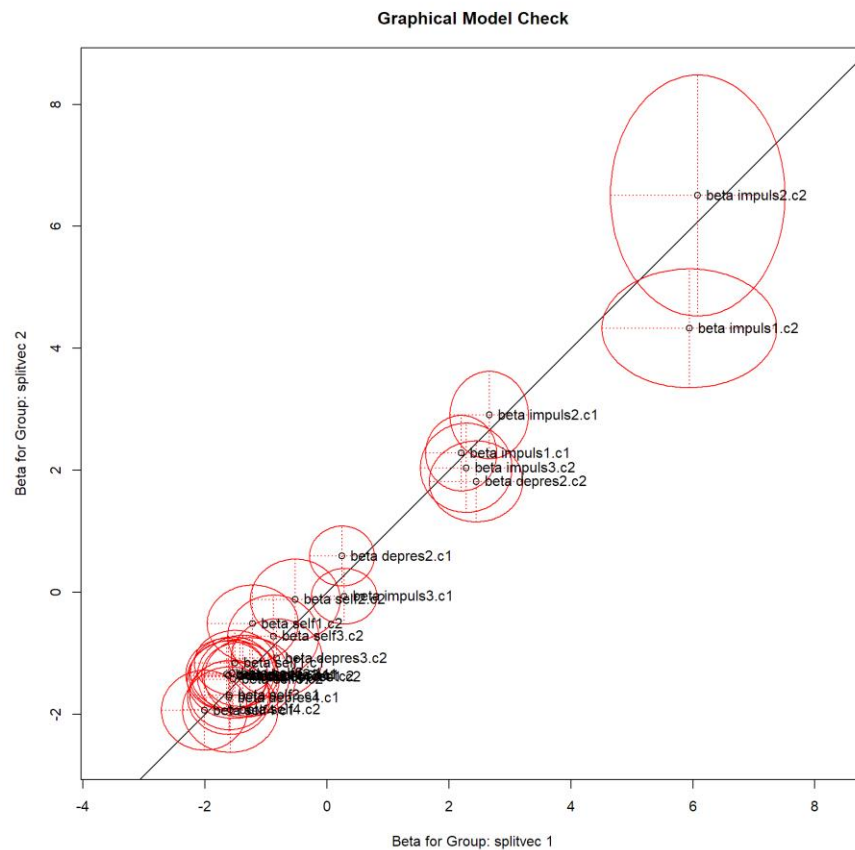


Figure 29 - GOF Outil R sur depress

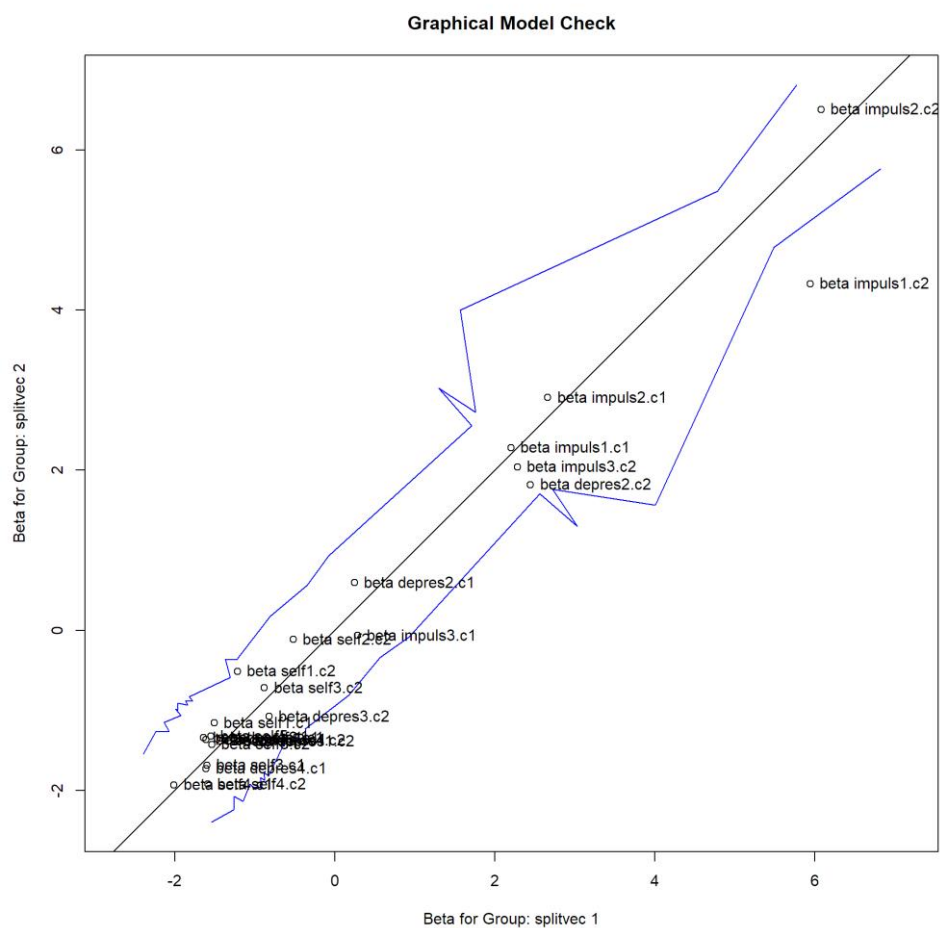


Figure 30 - GOF outil R - confidence bands

La distribution des personnes est aussi particulière au niveau du trait latent.

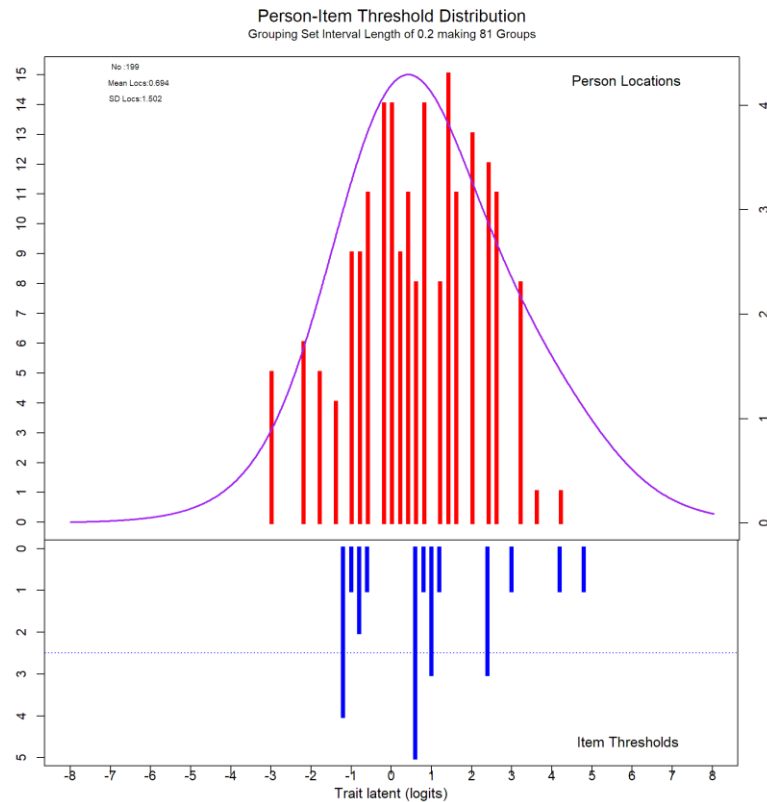


Figure 31 - distribution personnes et items dans depress

2.6 Comparaison entre les modules proposés et RUMM2030

Robinson, Johnson, Walton & MacDermid (2019) ont comparé RUMM2030, eRm et des packages IRT dans R. Il en ressort quelques points clés notamment pour eRm au niveau de R. Selon leur analyse, RUMM2030 et R peuvent fournir des courbes caractéristiques et des person-item threshold distribution maps similaires. Ceci étant, il manque certaines représentations « out of the box » avec eRm ainsi que certaines possibilités orientées Rasch.

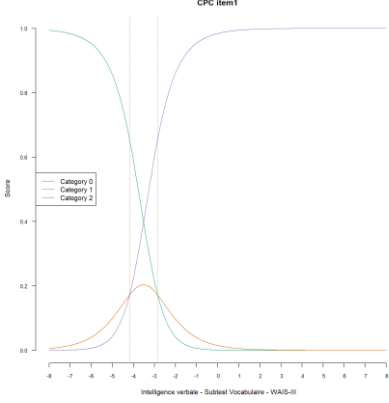
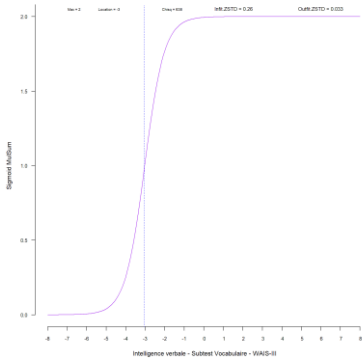
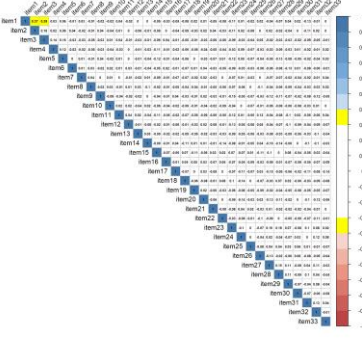
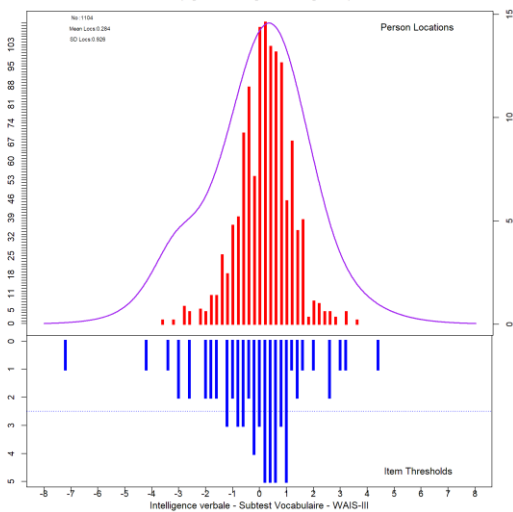
Le module R développé montre qu'il est possible de couvrir certains points équivalents en utilisant les outils open source et en tous cas, de sortir les données fondamentales dans l'un ou l'autre format.

L'avantage d'un outil open source est qu'il est possible de voir certains des processus utilisés et de fournir des extensions offrant les possibilités souhaitées.

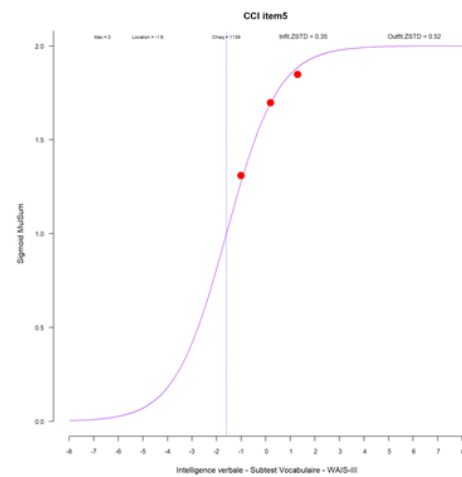
Un des apports notables de l'outil proposé est de fournir des éléments qui permettent de disposer de certaines fonctionnalités de RUMM2030 :

Tableau 4- Apports du module R développé avec plusieurs illustrations

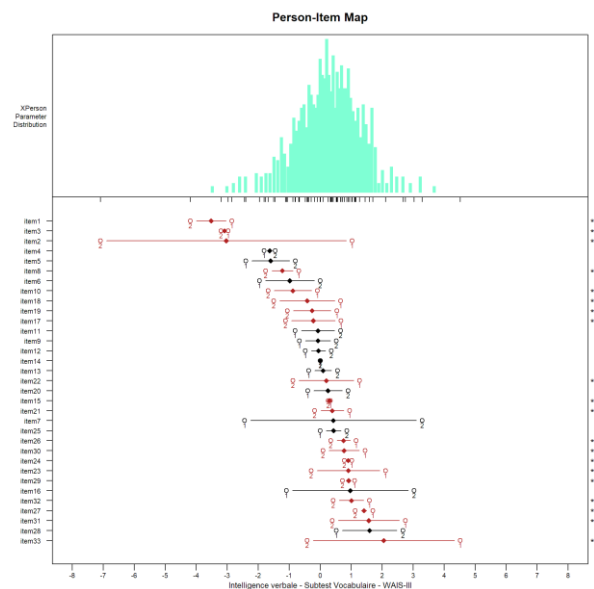
Support du rescoring	
----------------------	--

Courbes caractéristiques d'item avec indication des seuils	
Courbes sigmoïdes pour les items en polytomique (pas supporté de base dans eRm)	
Matrice de corrélation des résidus graphique	
Graphes de targeting : Person-Item Threshold distribution Person-Item Location distribution avec fonction d'information et statistiques intégrées	

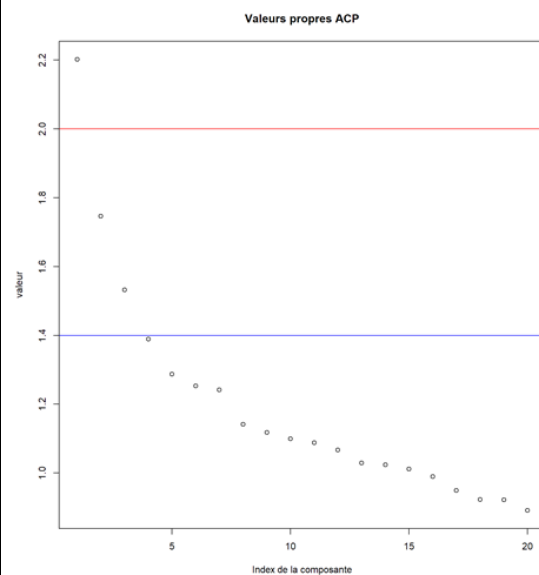
Support du clustering en classes avec affichage graphique de scores vrais vs scores attendus sur les courbes des items



Graphiques améliorés pour les Person-Item maps



ACP sur les résidus



Rapport spécifique fournissant des statistiques en théorie classique des tests sur les jeux de

données pour faire des comparaisons	
Utilisation d'un notebook pour conduire l'analyse et produire le rapport	Ceci permet de facilement reproduire les analyses. Les rapports peuvent être rapidement et facilement mis à disposition en ligne au format HTML (soit sur RPubS ou avec RStudio Connect en entreprise)

Robinson et al. (2019) indique des éléments pour lesquels notre solution apporte des avantages également :

Comparaison des capacités

Table 6 Comparison of the Capabilities of each Software Package

	RUMM2030	R-ltm	R-eRm	R-TAM	R-lordif
Model Fit Statistics	YES	YES	YES	YES	NO
Unidimensionality	YES	YES ^a	YES ^d	YES	NO
Category Thresholds	YES	NO	YES	YES	YES
DIF	YES	YES ^a	YES ^a	YES	YES
Local Independence	YES	YES	YES	YES	NO
Reliability	YES ^b	YES ^c	YES ^e	YES ^f	NO

^a Dichotomous Data Only
^b Person Separation Index
^c Cronbach's Alpha
^d Martin-Loef-Test
^e Conditional maximum likelihood framework
^f EAP and WLE reliability statistics
^g Joint maximum likelihood

fait avec LRTTest

Figure 32 - Comparatif de référence pour eRm - Robinson et al. (2019)

eRm couvre les besoins comme RUMM2030.

Au niveau de la reliability, le module développé en R calcule le PSI.

En ce qui concerne le FDI, le module donne des résultats aussi en polytomique.

Pour l'unidimensionnalité, eRm utilise Martin-Lof test comme alternative pour tester l'unidimensionnalité du test complet. Par contre, cela ne fonctionne pas avec les NA dans les données (limitation de la version actuelle qui semble encore trop récente).

Afin de comparer, on trouve les éléments suivants :

Table 1 RUMM2030 and R Output

			Item Fit Residual ^a		Person Fit Residual ^a		Chi-Square ^b		PSI ^c		UNID T-Test ^d	
			Mean	SD	Mean	SD	Value (DF)	p	With	without		
RUMM 2030	PAIN	Initial	-0.8398	1.7301	-0.4845	0.9770	24.391 (20)	0.4969	0.9218	0.9135	4.6800%	
		Final	-0.5336	1.1746	-0.4492	0.9127	18.5992 (20)	0.5480	0.8781	0.8644	2.3500%	
	SPECIFIC	Initial	-0.6052	2.1366	-0.4538	1.0702	54.0898 (30)	0.0045	0.9220	0.9033	4.7800%	
		Final	-0.6266	1.5595	-0.4497	0.9686	50.3341 (20)	0.0002	0.9016	0.8699	3.6800%	
	USUAL	Initial	-0.4119	1.7207	-0.4544	0.9432	23.1212 (20)	0.2829	0.9208	0.8304	6.1600%	
		Final	-0.3182	1.8380	-0.4642	1.0076	32.1662 (20)	0.0416	0.9217	0.8888	5.0700%	
			Item Fit Residual ^a		Person Fit Residual ^a				Cronbach's Alpha ^c		Martin-Loef ^d	
			Mean	SD	Mean	SD					Value (DF) p	
R	PAIN	Initial	1.63322	1.6195	-0.0447	1.89779			0.922		370.171(599)	1.0000
		Final	1.7807	1.6219	0.0067	1.7558			0.872		239.851(339)	1.0000
	SPECIFIC	Initial	1.6092	0.80491	0.12174	1.37694			0.92		392.039(899)	1.0000
		Final	1.7810	1.0349	0.1273	1.3960			0.8980		246.659(399)	1.0000
	USUAL	Initial	2.73459	1.43031	0.27866	1.84451			0.921		216.992(399)	1.0000
		Final	2.9494	1.9276	0.2744	2.0210			0.9200		ERROR	n/a

^a The fit residuals should have mean of 0 ± 2.5 and a standard deviation of 1 ± 2.5

^b The Chi-Square statistic should be small and statistically non-significant

^c A Person Separation Index (PSI) or Cronbach's Alpha should be > 0.70 to be statistically reliable

^d Unidimensionality is present if the percentage of statistically significant t-tests is $< 5\%$ or if the result of the Martin-Loef test is not significant

Figure 33 - Eléments à comparer entre les outils - Robinson et al. (2019)

3 Comparaison entre l'analyse TCT et le modèle de Rasch

Les analyses pour cette comparaison sont réalisées avec les nouveaux modules d'analyse (TCT et Rasch) développés dans ce mémoire.

3.1 Présentation du subtest Vocabulaire de la WAIS-III

Cette section analyse le subtest vocabulaire de la WAIS-III à l'aide du module R créé pour ce mémoire. Pour ce subtest, le psychologue va présenter des cartes reprenant une liste de 33 mots au sujet en le lisant au même moment. La consigne est de définir le mot oralement. Lorsque le sujet définit correctement le mot, il se voit attribué de 2 points ; s'il le fait partiellement, il reçoit alors 1 point. Si le sujet se trompe ou ne peut définir le mot, il a une note de 0 pour l'item concerné. La tâche commence directement à l'item 4. Si le sujet le réussit ainsi que le suivant (it. 5) alors il reçoit 2 points pour les 3 premiers items. Ceux-ci ont été ajoutés par rapport à la version précédente de la WAIS pour mieux discriminer les sujets avec un handicap mental léger ou modéré (Wechsler, 2000).

Les modifications apportées à la WAIS-III concernaient le fait d'avoir une plus grande étendue des scores. Le score au subtest vocabulaire fait partie de l'échelle verbale et de l'indice de compréhension verbale. De plus, les paramètres des items ont aussi été évalués

avec le modèle de Rasch via le logiciel RASCAL¹² (Hambleton,& Jones, 1993) et classé par ordre de difficulté. Les items présentant un fonctionnement différentiel (sujets belges/français), évalué par la méthode de Mantel-Haenszel ont été écartés pour plusieurs subtests mais cela n'a pas été possible pour le subtest Vocabulaire et donc (autant de chaque). Evolution au scores des subtests verbaux moins liée à l'âge que ceux performance.

3.2 Analyse TCT

3.2.1 Fidélité du subtest

L'alpha de Chronbach (0,8945) indique une très bonne fidélité du subtest Vocabulaire.

3.2.2 Difficulté et discrimination des items

Les occurrences des réponses aux items (figure 92) permettent de repérer des items très faciles (1, 2, 3) et un item très difficile (33). Il est également visible que l'augmentation de la difficulté annoncée au fur et à mesure des items dans le manuel ne soit pas complètement respectée.

Ensuite, l'analyse conjointe des « difficultés » et discriminations des items (figure 93) permet de faire différents constats :

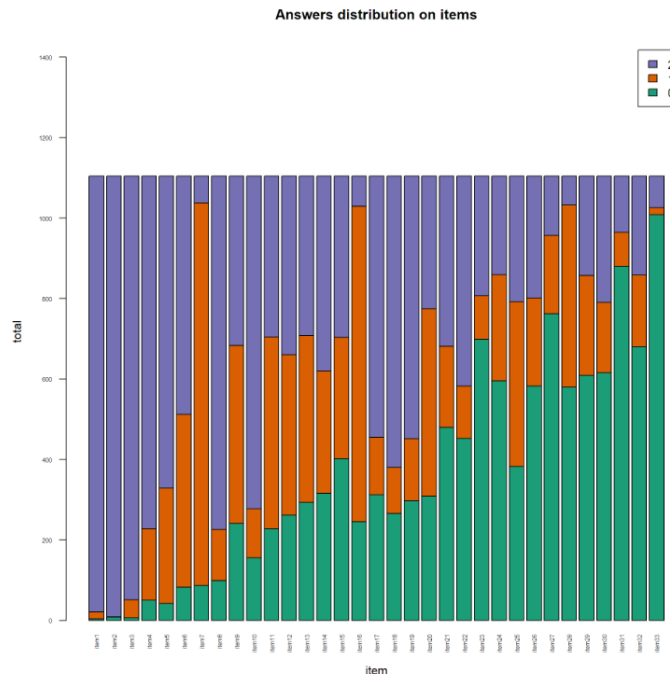


Figure 34 - occurrences des réponses à chaque item du subtest Vocabulaire de la WAIS-III

¹² <https://www.rasch.org/rmt/rmt63g.htm>

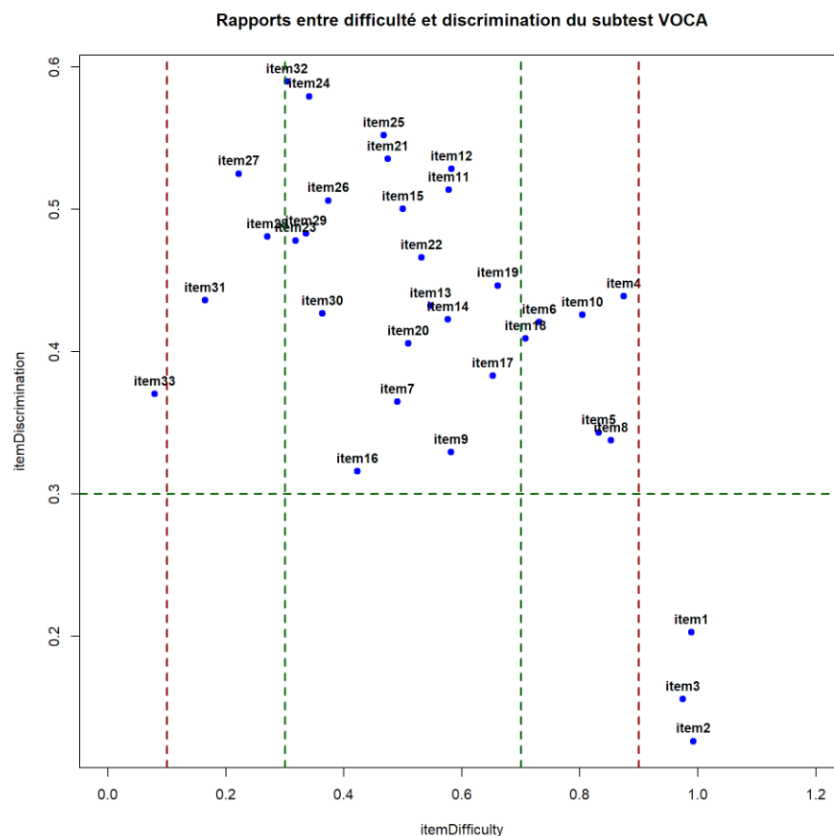


Figure 35 - rapports entre les indices de « difficulté » et discrimination des 33 items du subtest Vocabulaire de la WAIS-III

- ◆ 20 items ont une « difficulté » moyenne et une discrimination supérieure à 0,3. Ce sont des items à conserver (Laveault & Grégoire, 2014).
- ◆ 4 items avec un indice de « difficulté » dans des zones à exclure ($< 0,10$ ou $> 0,90$) :
 - certains avec un indice de discrimination bas : les items 1 (*bateau*), 2 (*fauteuil*), 3 (*Bol*) ;
 - et un avec un bon indice de discrimination : l'item 33 (*conjecture*).

Néanmoins le manuel de la WAIS-III mentionne que les 3 premiers items ont été ajoutés dans l'intention de discriminer les personnes avec déficience mentale. Vu la faible valeur de leur indice de discrimination, il serait intéressant de se questionner si dans la pratique l'objectif est réellement atteint.

Etant donné que les indices de discrimination sont influencés par la difficulté de l'item, il est difficile de ne garder ce type d'item que sur base de leur discrimination éventuellement haute (Laveault & Grégoire, 2014). Cependant au vu de la distribution des scores totaux, on distingue moins de 10% de l'échantillon avec des scores inférieurs à la moyenne. Cet échantillon ne contient probablement pas de nombreux sujets présentant un retard mental.

- ♦ 9 items dans les zones de discussion ($[0,10 ; 0,30[$ ou $]0,70 ; 0,90]$) :

A garder éventuellement en fonction de leur niveau de discrimination ($> 0,30$) qui est ici bon ou excellent.

En conclusion, sur base de ces 2 indices et au vu de certaines intentions présentées dans le manuel de la WAIS-III, seul l'item 33 devrait être supprimé. Néanmoins, par analogie avec les trois premiers items, au vu de ses deux indices, l'item 33 pourrait servir à discriminer les profils à haut potentiels.

Cependant dans cette analyse TCT, les items étant polytomiques, les écarts de progression entre les différentes catégories de chaque item ne sont pas connus.

Ces indices ne donnent aucune informations quant aux distances entre les différentes catégories des items polytomiques de ce subtest.

3.2.3 Ecart-type des items

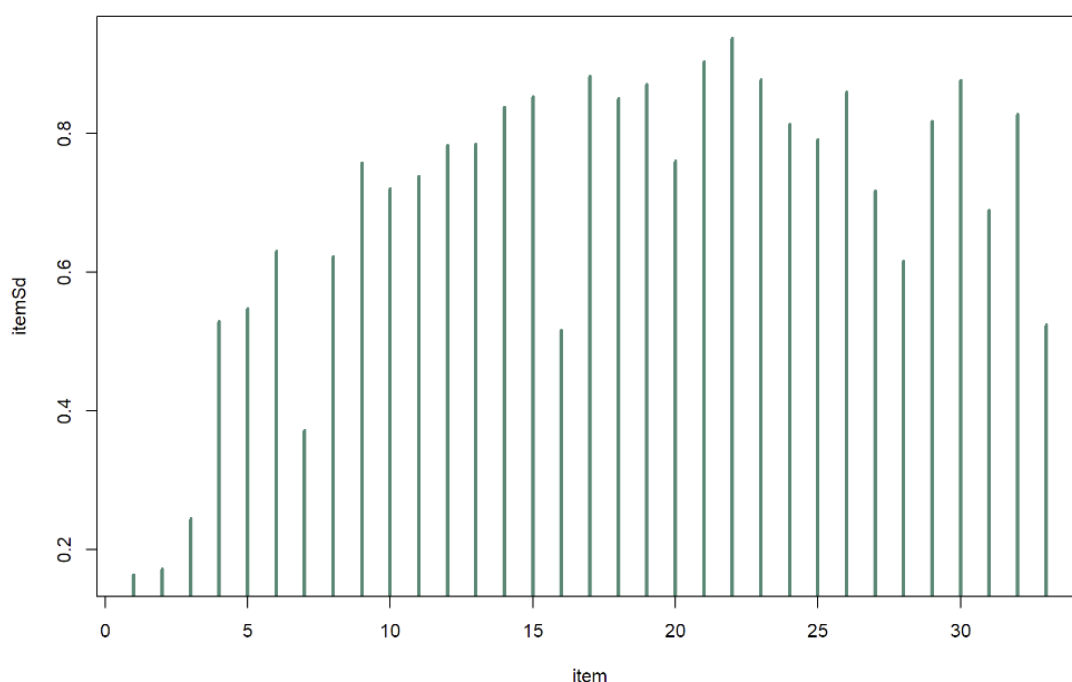


Figure 36- Ecart-type des items du subtest Vocabulaire de la WAIS-III

On constate par exemple qu'il n'y a pas de variabilité pour les premiers items.

3.2.4 Fiabilité et validité des items

Sur base de ces valeurs, on peut décider de ne garder que les items avec un coefficient de fidélité et un coefficient de validité élevés. Les informations sur la validité ne sont pas disponibles avec ces données.

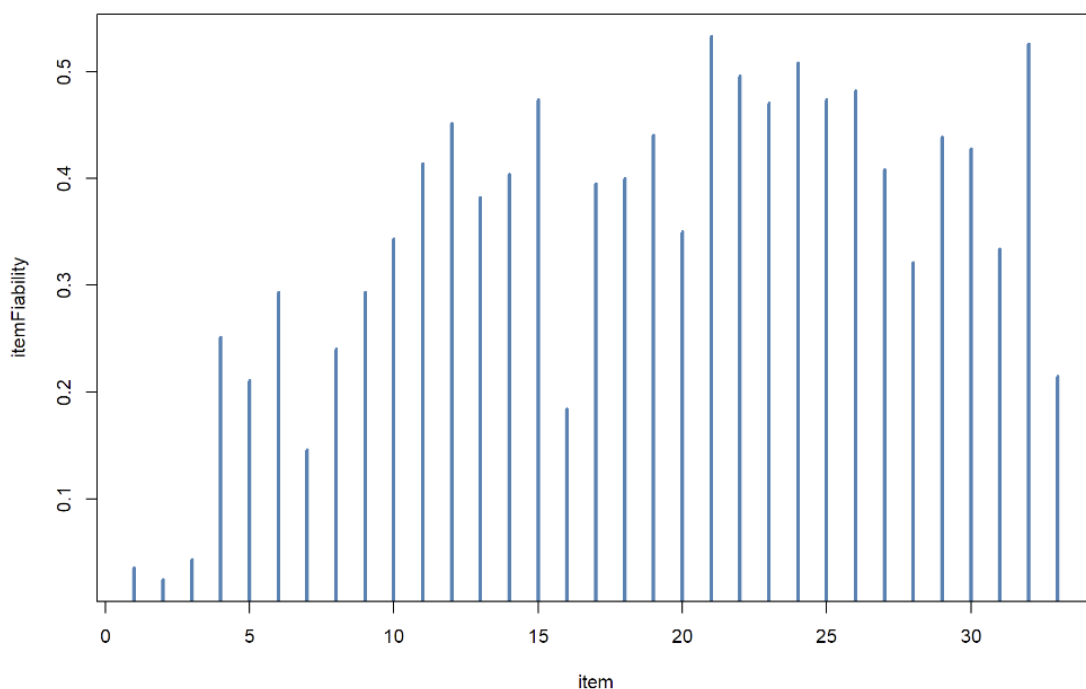


Figure 37- Fidélité des items du subtest Vocabulaire de la WAIS-III

3.2.5 Fonctionnement différentiel en fonction de l'âge et/ou du genre

Afin de procéder à une analyse de variances, il faut tester l'homogénéité des variances des groupes et la normalité de la distribution. Différents tests sont effectués pour ces vérifications.

Le test de Levene est significatif ($3,4 \cdot 10^{-6}$) de même que le test de Brown Forsythe¹³ ($1,3 \cdot 10^{-5}$). De plus, le test de normalité des score Shapiro-test est également significatif ($1,9 \cdot 10^{-3}$). Il n'est donc pas possible de procéder à une ANOVA pour tester ces effets.

Il est néanmoins possible de réaliser un test de Kruskal-Wallis qui est significatif pour l'âge ($1,9 \cdot 10^{-3}$) mais pas le genre ($1,9 \cdot 10^{-3}$). Il y a donc un rejet de l'hypothèse nulle d'absence de différence liée à l'âge.

3.3 Analyse Rasch

3.3.1 Les statistiques générales

La valeur du PSI (0,9018) indique une bonne fidélité et le ratio de séparation G (3,03) permet de calculer le nombre de strates à savoir pour ce subtest, 4 strates d'aptitudes distinctes. L'alpha de chronbach est également calculé (0.8946) et indique également une bonne fidélité.

¹³ <https://www.statsoft.fr/concepts-statistiques/glossaire/t/tests-levne.html>

3.3.2 Ciblage et difficulté des items

Au vu des occurrences (Annexes p), on peut repérer directement quelques items extrêmement faciles (1,2,3,4) ainsi que des items très difficiles (16,28,33).

Sur le graphique « targetting » on voit un bon ciblage de cet échantillon, pas d'effet plancher ni d'effet plafond.

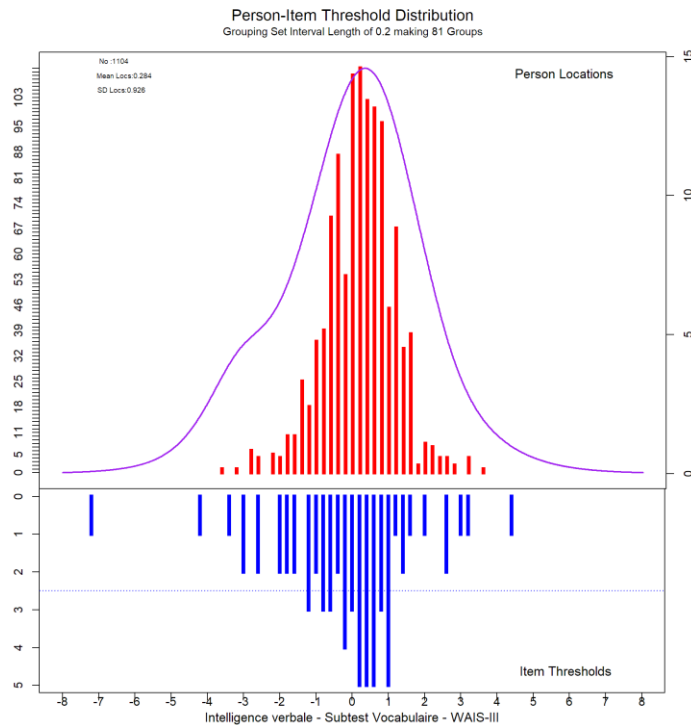


Figure 38 - Targettingthreshold et persons avec courbe d'information

Les moyennes des localisations des items et des personnes sont très proches.

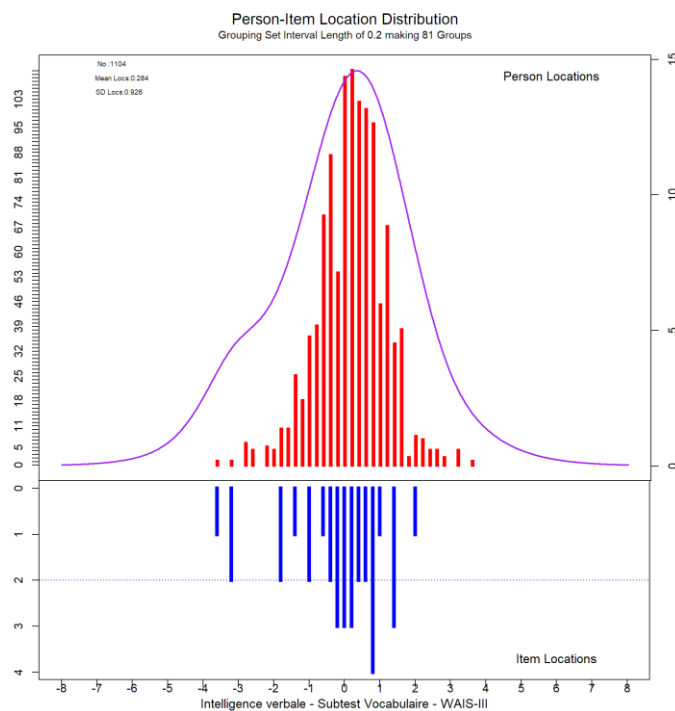


Figure 39 - TargettingItems et persons avec courbe d'information

3.3.3 Critères d'unidimensionnalité

1. Le test Lof n'est pas significatif
2. L'Analyse des catégories non émergentes, de l'ordre des seuils, des distances interseuils et des résidus est repris dans le tableau suivant :

Tableau 4 - Relevé items "à problème" selon les indicateurs

<u>Items avec catégorie non émergente et seuils désordonnés</u>	<u>Distance inter seuils > 5</u>	<u>Distance inter seuils < 1,4</u>	<u>Infit.ZSTD € [-2 ; 2]</u>	<u>Outfit.ZSTD € [-2 ; 2]</u>
<u>20 Items</u> 1,2,3,8,10,15,17,18, 19,21,22,23,24,26,2 7,29,30, 31, 32, 33	<u>1 item</u> 7	<u>27 Items</u> Tous sauf : <u>5,6,11,16,28</u>	<u>14 Items</u> <u>< -2</u> 4, 11, 12, 21, 24, 25, 27, 28, 32	<u>14 Items</u> <u>< -2</u> 4, 11, 12, 21, 24, 25, 27, 28, 32
Pour tous, la catégorie de réponse 1 n'émerge pas			<u>≥2</u> 9, 14, 17, 18, 30	<u>≥2</u> 9, 14, 17, 18, 30

Au vu des nombreux items avec une problématique de résidus et/ou de seuils désordonnés et de catégories non émergentes, un recodage des catégories pourrait être envisager. Ce recodage peut altérer la fidélité.

Si l'on réfléchit à la philosophie de test qui semble prévoir une progression par rapport au trait latent, le recodage n'y correspond pas vraiment que ce soit en rassemblant les catégories 0 et 1 (réponse parfaitement complète pour acquérir le trait) ou 1 et 2 (réponse partiel et on a acquis le trait). Une troisième option est de remplacer ces items.

Vu que beaucoup d'entre eux cumulent des problématiques, cette option est à considérer. Le problème des seuil désordonnés est le non respect l'augmentation du trait le long de l'échelle. De plus les écarts différents entre les catégories de différents items, signifie une progression différente pour une même acquisition de point.

3.3.3.1 Fonctionnement différentiel des items+

Au vu des caractéristiques analysée par (article de Mr Grégoire), différents graphiques ont été réalisés pour analysé un fonctionnemnt différentiel en

fonction des catégories d'âges. Ensuite, l'analyse a été également faite pour le genre. Un exemple est en [annexe 6 \(11.1.2.\)](#)

3.3.4 Corrélation des résidus

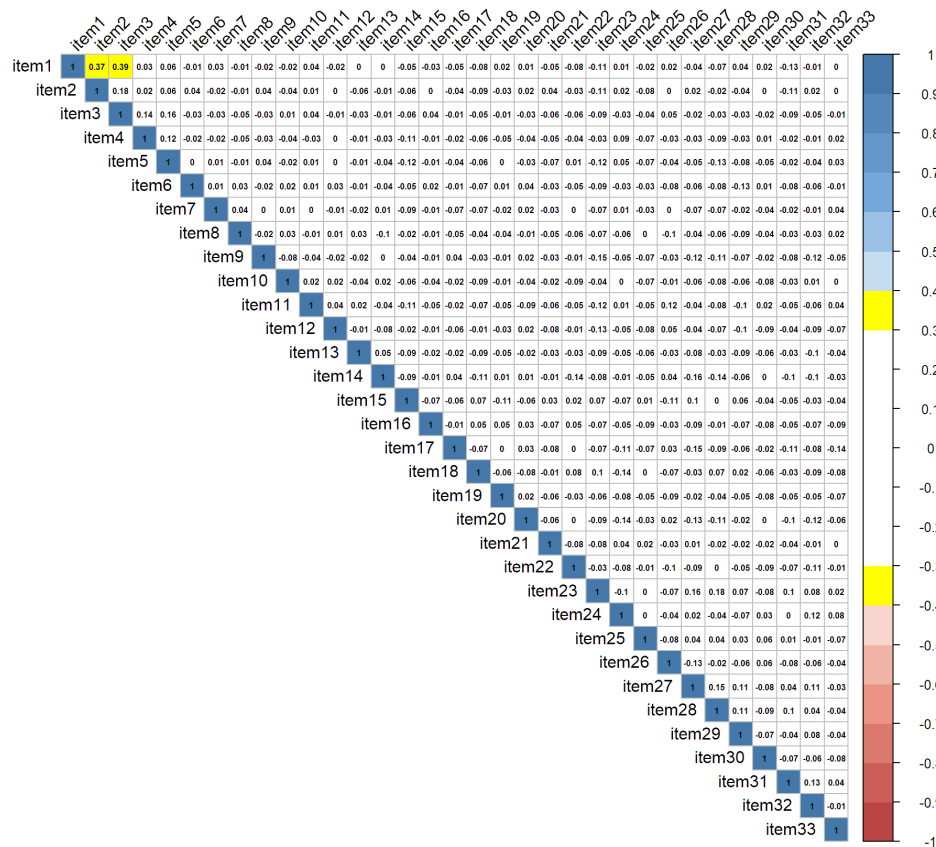


Figure 40 - Matrice des corrélations des résidus

La matrice des corrélations nous indique que les items 2 et 3 sont corrélés à l'item 1.

- ➔ Item 1 candidat à la suppression ?
- ➔ Une réflexion est à mener du fait que ces trois items ont été ajoutés

3.3.5 Redondance des items

Les candidats à l'élimination étant très nombreux, cette analyse n'est probablement pas prioritaire. Néanmoins, si le choix posé est de garder une grande partie des items et de procéder à quelques éliminations, il y a plusieurs items en zone centrale du targetting qui cible les mêmes niveaux de l'échelle.

Pour conclure, au vu des « mauvais » indices de nombreux items, il serait intéressant de questionner m'unidimensionnalité de ce jeu d'item. En effet, certains des mots proposés demande une certaine culture générale acquise soit

vie un bon parcours scolaire soit via un milieu social instruit. De plus certains mots ont une fréquence d'apparition plus grande que d'autres

4 Conclusion

Cette section contextualise les angles clés de ce mémoire et propose quelques pistes pour le futur.

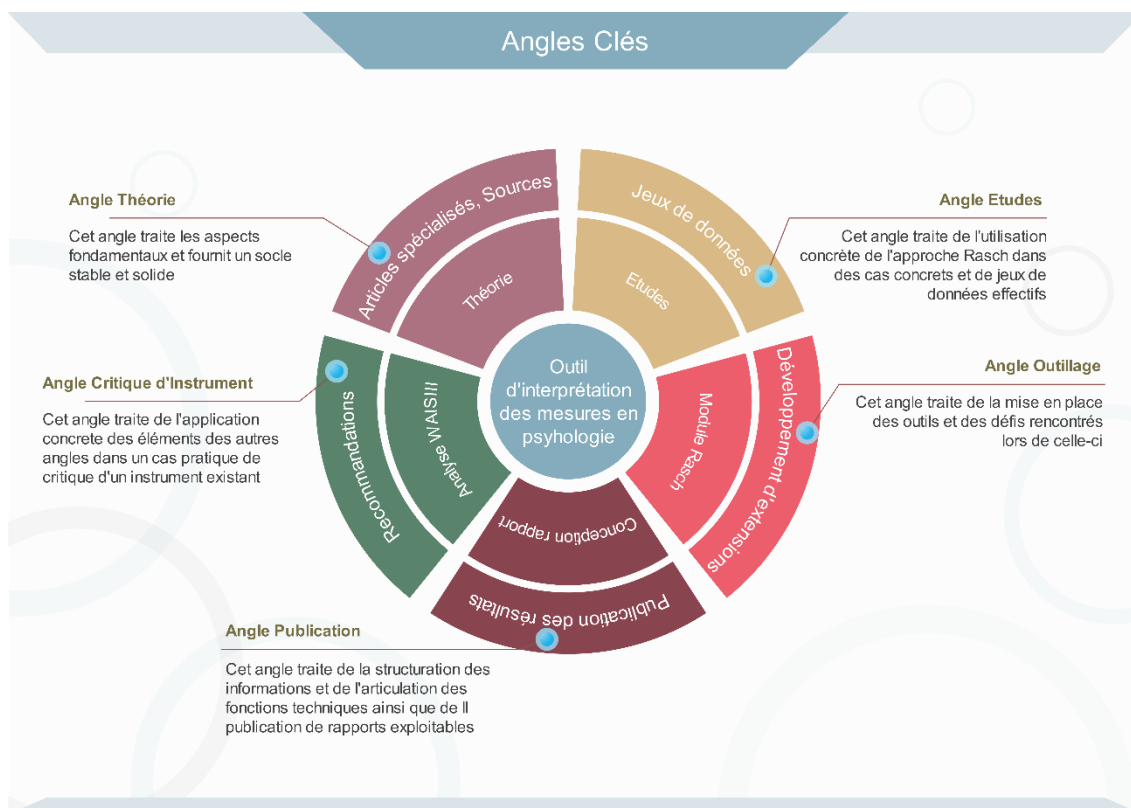


Figure 41 - Angles Clés pour le module d'analyse Rasch en R développé dans ce mémoire

Ce mémoire a permis d'aller en profondeur en ce qui concerne les concepts du modèle de Rasch, que ce soit sur le plan théorique ou par rapport au contraste avec la théorie classique des tests ou encore vis-à-vis de l'outil RUMM2030 et les possibilités du package eRm de R.

Grâce à l'examen de la théorie et des fondamentaux, les enjeux conceptuels, les conditions d'application et les opportunités liés au modèle de Rasch ont été investigués de manière pointue.

La présentation développée de plusieurs recherches en psychologie pour lesquelles le modèle de Rasch a été utilisé, a permis de mettre en avant différents apports utiles de ce

modèle. Parmi ceux-ci, il y a la réduction d'échelle de diagnostic à des formes courtes et psychométriquement valides. Il y a également, l'estimation des difficultés d'items permettant de construire un continuum hiérarchisé avec ceux-ci et d'avoir une vue précise de l'étendue des plages de mesure possibles.

Ensuite, l'indépendance de la mesure fournie par le modèle de Rasch en comparaison avec les résultats du modèle TCT dans certains échantillons ouvre

La confrontation avec des jeux de données réels au travers des développements logiciels a amené à examiner en profondeur les possibilités, les défis, et les limites de eRm.

Des surprises régulières lors des développements ont amené des interrogations assez poussées sur le sens des formules, l'ordinateur ne faisant que reproduire ce qui est « demandé ». Ces questionnements sont la clé pour aller vers ce qui permet d'avancer et la mise en place des bonnes solutions une fois les différentes options explorées.

La mise en place d'un module fournissant les éléments utiles en théorie classique des tests permet de disposer d'informations pertinentes en vue de comparer les résultats de cette approche avec l'approche Rasch.

La réalisation d'extensions permettant de livrer des visualisations ou des informations qui n'étaient pas disponibles au départ montre l'avantage d'une solution basée sur des outils et packages Open Source qui permet de transcender les limites de eRm.

L'examen de la sous échelle de WAIS-III au travers des outils réalisés a été particulièrement instructive pour l'examen critique de cette sous échelle et la production de recommandations sur les limites, soucis et améliorations possibles la concernant.

Le LRtest d'Andersen est ainsi un outil fondamental de travail et a permis d'approfondir l'approche CML et les log likelihood ratios.

En termes de DIF, il a été très intéressant d'examiner le concept, de même que de mettre en place une logique de clustering des niveaux de trait latent.

Afin de présenter les résultats, l'approche utilisée est basée sur la technologie RMarkdown qui s'est avérée particulièrement efficace pour organiser les rapports produits et permet également de partager du code entre les approches Rasch et TCT.

La mise en place de ce rapport nous a aussi fait nous interroger sur comment nommer et structurer les sections du rapport afin qu'il soit lisible et compréhensible par l'utilisateur final. Mentionnons que l'on peut tout à fait imaginer produire des rapports sur base de résultats de questionnaires et de publier les résultats en vue de leur consommation par des

profils en pure consultation. Notons qu'il est possible de calculer toutes les informations avec le rapport principal et de produire des rapports partiels exploitant ces données afin de cibler certains aspects. Ceci constitue un avantage non négligeable de l'approche Open Source.

L'aspect « recherche reproductible » est aussi mis en avant car les résultats peuvent être recalculés par tout qui dispose du code du rapport et de R.

Nous avons aussi découvert de nombreux packages spécifiques tout au long de ce travail, ce qui nous a équipée pour travailler très efficacement sur des jeux de données, que ce soit pour l'approche Rasch ou non. Ceci constitue un autre avantage de l'approche Open Source, à savoir l'appropriation et l'extension des outils selon les besoins.

La mise à disposition d'un outil Open Source basé sur R permet également d'ouvrir le champ en termes d'utilisation d'un outil Rasch facilement exploitable et dont les résultats peuvent être publiés rapidement et partagés, via un seul fichier rapport ou via une publication sur Internet. Ceci contribue au rayonnement et à la diffusion de l'approche Rasch qui se voit actuellement limitée suite aux outils dont l'utilisation reste confidentielle eu égard à leur complexité de prise en main.

Le choix de R est à notre sens pertinent car il constitue la lingua franca des statisticiens. En termes d'horizons futurs, il conviendrait de continuer le développement de l'outil, certaines parties réalisées au départ n'étant pas optimales par eu égard à notre niveau de compétence qui s'est amélioré au fur et à mesure et donc, certaines parties devraient être revues pour plus d'efficacité ou de facilité d'exploitation. Il serait également utile de contribuer les fonctions des extensions aux package eRm afin de les mettre à disposition du plus grand nombre.

Il serait également intéressant de poursuivre des comparaisons entre outils en utilisant le module de simulation afin de disposer de résultats centrés sur des données au profil bien connu. Le comportement des logiciels pourra ainsi être examiné au travers de diverses distribution des personnes et des items.

5 Bibliographie

- Andersen, E. B. (1973). A goodness of fit test for the rasch model. *Psychometrika*, 38(1), 123-140. <https://doi.org/10.1007/bf02291180>
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561-573. <https://doi.org/10.1007/bf02293814>
- Andrich, D. (2004). Controversy and the Rasch Model : A characteristic of incompatible paradigms ? *Medical Care*, 42(1), 7-16. <https://doi.org/10.1097/01.mlr.0000103528.48582.7c>
- Bertrand, R., & Blais, J. G. (2004). *Modèles de mesure : L'apport de la théorie des réponses aux items*. Presses de l'Université du Québec. Consulté à l'adresse https://muse.jhu.edu/book/15412?fbclid=IwAR0x2BFqS_p6MAHr2D901tJmfvKQ_RjHwSSRWxmDULilIxyaZ5eS1WoDlfw
- Boadé, G. (2013). *Robustesse du modèle de Rasch unidimensionnel à la violation de l'hypothèse d'unidimensionnalité*. (2013, juin). (Mémoire, Université de Montréal). Consulté à l'adresse : <https://papyrus.bib.umontreal.ca/xmlui/handle/1866/9942>
- Bryce, T. G. K. (1981). Rasch-Fitting. *British Educational Research Journal*, 7(2), 137-153. <https://doi.org/10.1080/0141192810070203>
- Bond, T. G. (2003). Validity and assessment: a Rasch measurement perspective. *Metodología de las Ciencias del Comportamiento*, 5(2), 179-194. <https://researchonline.jcu.edu.au/1799/1/bondValidity.pdf>
- Bond, T. G., Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences, Third Edition*. Routledge.
- Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, 71(3), 425-440. <https://doi.org/10.1007/s11336-006-1447-6>
- Borsboom D., Zand Scholten A. The Rasch Model and Conjoint Measurement Theory from the Perspective of Psychometrics. *Theory & Psychology* · February 2008; 18; 111. DOI: 10.1177/0959354307086925.
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, 110(2), 203-219. <https://doi.org/10.1037/0033-295x.110.2.203>

- Caroff, X. (2004). L'identification des enfants à haut potentiel : Quelles perspectives pour l'approche psychométrique ? *Psychologie Française*, 49(3), 233-251.
[https://doi.org/10.1016/s0033-2984\(04\)00046-9](https://doi.org/10.1016/s0033-2984(04)00046-9)
- De Bruin, G.P., Hill, C., Henn, C. M., & Muller, K.-P. (2013). Dimensionality of the UWES-17: An item response modelling analysis. *SA Journal of Industrial Psychology/SA Tydskrif vir Bedryfsielkunde*, 39(2), Art. #1148, 1-8.
<http://dx.doi.org/10.4102/sajip.v39i2.1148>
- DeVellis R. F. (2006). Classical Test Theory. *Medical Care*, 44(11), Measurement in a Multi-Ethnic Society, S50-S59.
<https://doi.org/10.1097/01.mlr.0000245426.10853.30>
- Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists (Multivariate Applications Series)*. Mahwah, NJ : Erlbaum, p12.
- Flieller, A. (1994). Méthodes d'étude de l'adéquation au modèle logistique à un paramètre (modèle de Rasch). *Mathématiques et sciences humaines*, 127, 19-47.
http://www.numdam.org/article/MSH_1994__127__19_0.pdf
- Goulet V. (2016). Introduction à la programmation en R. 5^e édition
ISBN 978-2-9811416-6-8 – consulté sur :
https://cran.r-project.org/doc/contrib/Goulet_introduction_programmation_R.pdf
- Guyon, H. (2016). *Variables latentes et processus mentaux : une réflexion épistémologique et méthodologique en Psychologie*. (Thèse : Université Sorbonne, Paris). Consulté à l'adresse <https://www.theses.fr/2016USPCB015>
- Hambleton, R. K., & Jones, R. W. (1993). An NCME Instructional Module on Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38–47.
<https://doi.org/10.1111/j.1745-3992.1993.tb00543.x>
- Hogan, T. P., Parent, N., & Stephenson, R. (2017). *Introduction à la psychométrie, (2^e éd) (French Edition)*. CHENELIERE.
- Huet, K. (2018). Edumétrie et psychométrie, document non publié, Université de Mons.
- Juhel, J. (1991). Les Modèles de Réponse à l'Item en psychométrie cognitive: principes de base et exemples d'application. Communication présentée au Séminaire du GRPD, LPE, Université Rennes 2. Consulté à l'adresse
<https://xn--universit-rennes2-jtb.academia.edu/JacquesJuhel>

- Kubinger, K.D. (2000). Replik auf Jürgen Rost "Was ist aus dem Rasch-Modell geworden?": Und für die Psychologische Diagnostik hat es doch revolutionäre Bedeutung. <https://www.univie.ac.at/Psychologie/diagn/files/Rostrep.pdf>
<https://www.psychlinker.de/link.php?link=22630>
- Kubinger, K. D. (2005). Psychological Test Calibration Using the Rasch Model - Some Critical Suggestions on Traditional Approaches. *International Journal of Testing*, 5(4), 377-394. http://dx.doi.org/10.1207/s15327574ijt0504_3
- Laveault, D., & Grégoire, J. (2014). *Introduction aux théories de tests en psychologie et sciences de l'éducation (Méthodes en sciences humaines) (French Edition)* (3e éd.). Louvain-la-Neuve : De Boeck Supérieur.
- Le Corff, Y., & Yergeau, E. (2017). Psychométrie à l'UdeS. Consulté à l'adresse <http://psychometrie.espaceweb.usherbrooke.ca/>
- Lerdal, A., Opheim, R., Gay, C. L., Moum, B., Fagermoen, M. S., & Kottorp, A. (2017). Psychometric limitations of the 13-item Sense of Coherence Scale assessed by Rasch analysis. *BMC Psychology*, 5-18. <https://doi.org/10.1186/s40359-017-0187-y>
- Linacre, J. M. (s.d.). Comparing and Choosing between "Partial Credit Models" (PCM) and "Rating Scale Models" (RSM). Retrieved from <https://www.rasch.org/rmt/rmt143k.htm>
- Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, 7 (4), 328. Retrieved from <https://www.rasch.org/rmt/rmt74m.htm>
- Linacre, J. (2002). Optimizing Rating Scale Category Effectiveness. *Journal of Applied Measurement*, 3(1),85-106. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.424.2811&rep=rep1&type=pdf>
- Magno, C. (2009). Demonstrating the Difference between Classical Test Theory and Item Response Theory Using Derived Test Data. *The International Journal of Educational and Psychological Assessment*, 1, 1-11.

Mair, P., & Hatzinger, R. (2007a). CML based estimation of extended Rasch models with the eRm package in R. *Psychology Science*, Volume 49, p. 26–43.

Mair, P., & Hatzinger, R. (2007b). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, 20(9), 1–20. <https://doi.org/10.18637/jss.v020.i09>

Mise à jour de ces 2 articles avec un auteur en plus :

Mair, P., Hatzinger, R., & Maier, M. J. (s.d.). Consulté à l'adresse

<https://cran.r-project.org/web/packages/eRm/vignettes/eRm.pdf>

Merbitz, C., Morris, J., & Grip, J. C. (1989). Ordinal Scales and Foundations of Misinference. *Arch Phys Med Rehabil*, 70, 308-312. Consulté à l'adresse [https://www.archives-pmr.org/article/0003-9993\(89\)90151-2/pdf](https://www.archives-pmr.org/article/0003-9993(89)90151-2/pdf)

Penta, M., Arnould, C., & Decruynaere, C. (2005). *Développer et interpréter une échelle de mesure : Applications du modèle de Rasch (PRATIQUES PSYCHOLOGIQUES) (French Edition)*. Sprimont: Mardaga.

Penta, M., & Lories, G. (2018). Analyse de données : modèles de mesure (LPSYS2144), document non publié, Université catholique de Louvain.

Petrillo J., Cano S. J., McLeod L. D., Coon C. D. Using Classical Test Theory, Item Response Theory, and Rasch Measurement Theory to Evaluate Patient-Reported Outcome Measures: A Comparison of Worked Examples. *VALUE IN HEALTH* 18 (2015) 25 – 34. <http://dx.doi.org/10.1016/j.jval.2014.10.005>

Prieto, L., Alonso, J., & Lamarca, R. (2003). Classical test theory versus Rasch analysis for quality of life questionnaire reduction. *Health and Quality of Life Outcomes*, 1, 1-27. <https://doi.org/10.1186/1477-7525-1-27>

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17(5), 1–25. Retrieved from <http://www.jstatsoft.org/v17/i05/>

Robinson M., Johnson A.M., Walton D.M. & MacDermid J.C. (2019). A comparison of the polytomous Rasch analysis output of RUMM2030 and R (ltm/eRm/TAM/lordif). *BMC Medical Research Methodology*. <https://doi.org/10.1186/s12874-019-0680-5>

- Roulin, J.L.. (2018). Savoir, Comprendre, Apprendre. Leçons de Psychométrie (3^{ème} édition). Cours de la formation universitaire à distance Suisse, document non publié, Université Savoie Mont Blanc. Consulté sur :
<http://www.psychometrie.jlroulin.fr/cours/cours-psychometrie.pdf?fbclid=IwAR08f1N3YXsPr63QJvGG5PID96pON5NKp23lRGfav4s2uDOJQb0cb-gKvpY>
- R Markdown. (s. d.). [Vidéo]. <https://rmarkdown.rstudio.com/lesson-1.html>.
<https://vimeo.com/178485416>
- Schiltz, J., & Schiltz, L. (2007). De l'adéquation d'un test informatisé en mathématiques pour élèves à haut potentiel présentant un fléchissement scolaire à l'âge de la puberté. *Faísca*, 12(14), 84-105. <http://hdl.handle.net/10993/5640>
- Shea, T. L., Tennant, A., & Pallant, J. F. (2009). Rasch model analysis of the Depression, Anxiety and Stress Scales (DASS). *BMC Psychiatry*, 9-21.
<https://doi.org/10.1186/1471-244x-9-21>
- Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, 103(2684), 677-680. <https://doi.org/10.1126/science.103.2684.677>
- Tennant, A., Mckenna, S.P., Hagell, P. (2004). Application of Rasch analysis in the development and application of quality of life instruments. *Value Health* 7 Suppl 1, S22-26.
- Vautier, S. (2015). La psychotechnique des aptitudes. Pour différencier une sociotechnique de l'évaluation sans mesurage et une psychologie balbutiante de la compréhension de la performance. *Pratiques Psychologiques*, 21(1), 1-18.
<https://doi.org/10.1016/j.prps.2015.01.005>
- Vautier, S. (2016). "40a. Les échelles de mesure : lire Stevens (1946)", *Épistémologie de la psychologie*, Carnet d'enseignement et de recherche de Stéphane Vautier, document non publié, université Toulouse 2. Consulté à l'adresse
<https://epistemo.hypotheses.org/3696>
- Wechsler, David, (2000). *WAIS-III: Wechsler adult intelligence scale-Third edition*. (3e ed.). ECPA.
- Wenders, A., & Visschers, M. (2016). *Utilisation de tests adaptatifs pour les cours d'informatique* (Mémoire : Ecole polytechnique de Louvain, Université catholique de Louvain). Prom. : Bonaventure, Olivier. Consulté à l'adresse
<http://hdl.handle.net/2078.1/thesis:4580>

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation*, 70(12), 857-860. Consulté à l'adresse <https://www.rasch.org/memo44.htm>

Wright, B. D., & Masters, G. N. (Éds.). (1982). *Rating scale analysis : Rasch measurement*. (2^e éd.). Chicago, Illinois: Mesa Press

6 Annexe 1

6.1 La psychométrie (Roulin, 2018)

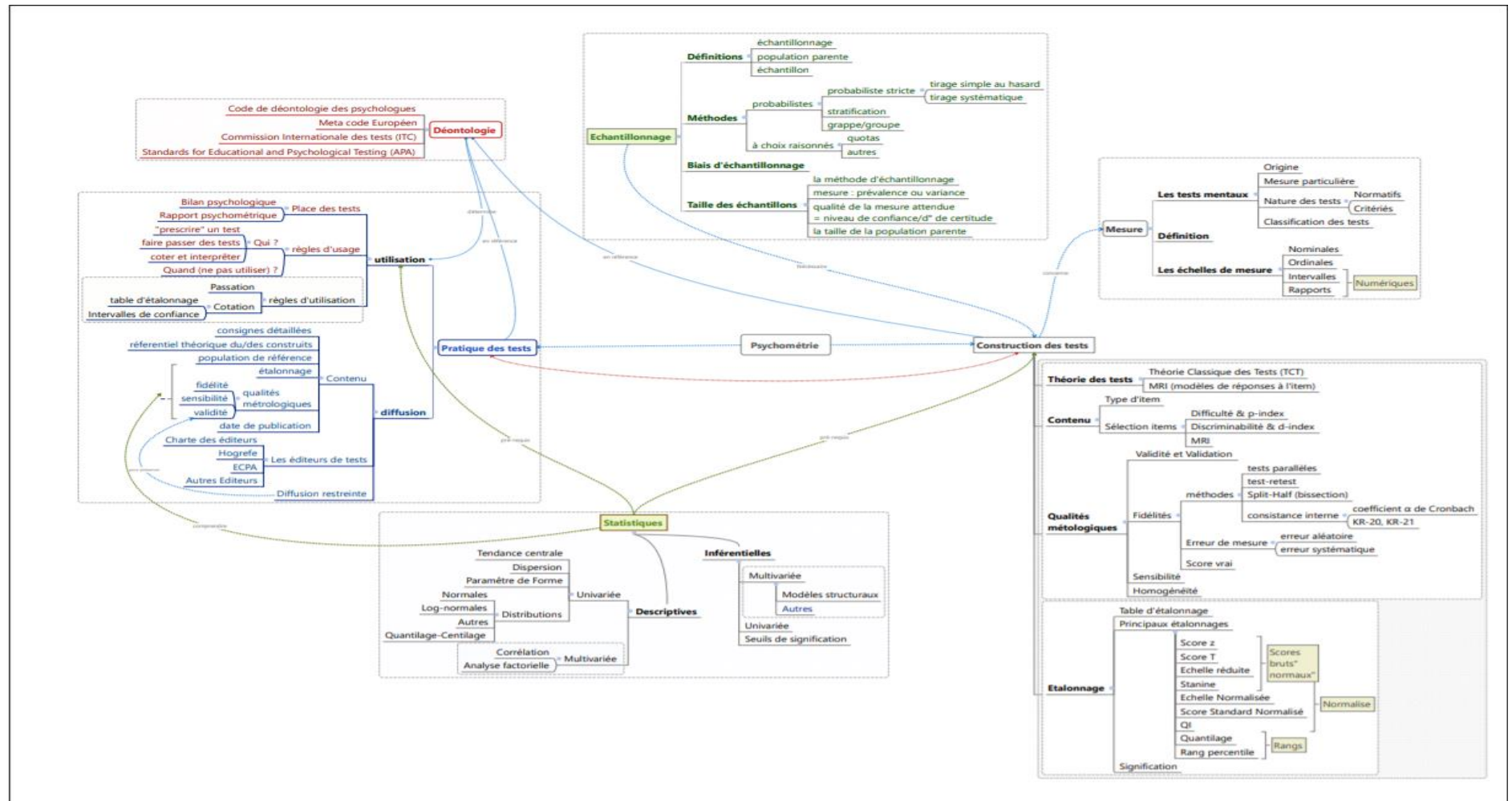


Figure 42 - Mind map psychométrie - Roulin (2018)

7 Annexe 2

7.1 Réduction de l'échelle NHP38

7.1.1 Réduction de l'échelle NHP38 – Modèle TCT

Table 2: Reduced NHP38 version obtained through Classical Test Theory (CTT): the NHP20

Original NHP38 items By dimension			1 st set of criteria for reduction: Discrimination (<i>r</i>) & Endorsement (<i>p</i>)		2 nd set of criteria for reduction: Factor analysis*			NHP20
Dimension	No. items	α	Items deleted as $r < 0.40$	Items deleted as $P < 0.20$	Items deleted as communality < 0.30	Items deleted as main loading < 0.40	Items deleted as difference between similar loadings ≤ 0.1	No. Items remaining
Energy (EN)	3	.76	-	-	-	-	EN2, EN3	1
Pain (P)	8	.90	-	P2	-	-	P4, P5	5
Emotional Reactions (EM)	9	.82	EM8	-	EM5, EM7	-	-	6
Sleep (SL)	5	.81	SL1	-	-	-	-	4
Social Isolation (SO)	5	.78	-	SO2, SO3 SO4, SO5	SO1	-	-	0
Physical Mobility (PM)	8	.83	-	PM1, PM3 PM8	-	-	PM6	4
Summary Index	38	.95	(28 items remaining, $\alpha = 0.94$)					20 ($\alpha = .92$)

Figure 43- tableau des résultats de l'analyse TCT relative à la réduction du questionnaire de qualité de vie NHP38 – Prieto et al., (2003)

7.1.2 Réduction de l'échelle NHP38 – Modèle Rasch

Table 3: Reduced NHP version obtained through Rasch Analysis: the NHP22

PHYSICAL SCALE				
ITEMS	ANCHORED MEASURE	SE	INFIT MNSQ	OUTFIT MNSQ
PM3-UNABLE TO WALK	76.1	.62	.82	.96
PM1-WALKING LIMITED	64.1	.51	1.03	1.35
P2-AWFUL PAIN	59.7	.47	.89	.92
PM6-HARD TO DRESS	55.7	.45	.83	.72
P8-PAIN SITTING	53.1	.44	.93	.90
PM5-HARD TO REACH	47.5	.42	.83	.74
EN3-OUT OF ENERGY	44.7	.41	.99	.99
P3-CHANGE PAIN	41.5	.41	.96	.97
P4-WALK PAIN	39.1	.41	.92	.91
PM2-HARD TO BEND	34.8	.41	.90	.93
PM4-STAIRS HARD	25.2	.43	.99	1.50
PSYCHOLOGICAL SCALE				
ITEMS	ANCHORED MEASURE	SE	INFIT MNSQ	OUTFIT MNSQ
SO5-PEOPLE HARD	68.6	.55	1.00	1.17
SO2-CONTACT HARD	60.6	.48	.96	1.03
SO4-I'M A BURDEN	58.6	.46	.98	1.02
EM4-DAYS DRAG	56.1	.44	.90	.89
EM6-NO CONTROL	55.0	.44	.90	.88
EM9-DEPRESSED	48.1	.41	.86	.81
EM2-JOY FORGOTTEN	44.9	.41	1.06	1.11
SL3-CAN'T SLEEP	44.0	.40	.93	.88
SL5-SLEEPS BADLY	40.8	.40	.89	.86
SL4-SLOW TO SLEEP	39.2	.40	1.02	1.08
EM1-GETTING ME DOWN	36.8	.43	1.20	1.34

Figure 44 - tableau des résultats de l'analyse Rasch relative à la réduction du questionnaire de qualité de vie NHP38 – Prieto et al., (2003)

7.1.3 Contenu comparé des deux versions NHP réduites

Table 4: Content of the reduced NHP versions

Original NHP38 dimensions	Classical Test Theory reduction NHP20				Rasch reduction NHP22	
	Emotional	Physical	Sleep	Pain	Physical	Psychological
Energy						
EN1 I'm tired all the time	X					
EN2 Everything is an effort						
EN3 I soon run out of energy					X	
Pain						
P1 I have pain at night				X		
P2 I have unbearable pain					X	
P3 I find it painful to change position				X	X	
P4 I'm in pain when I walk					X	
P5 I'm in pain when I'm standing						
P6 I'm in constant pain				X		
P7 I'm in pain when going up/down stairs		X				
P8 I'm in pain when I'm sitting				X	X	
Emotional Reactions						
EM1 Things are getting me down	X					X
EM2 I've forgotten how to enjoy myself	X					X
EM3 I'm feeling on edge	X					
EM4 These days seem to drag	X					X
EM5 I lose my temper easily these days						
EM6 I feel as if I'm losing control	X					X
EM7 Worry is keeping me awake at night						
EM8 I feel that life is not worth living						
EM9 I wake up feeling depressed	X					X
Sleep						
SL1 I take tablets to help me sleep						
SL2 I'm waking in the early hours ...			X			
SL3 I lie awake for most of the night			X			X
SL4 It takes me long time to get to sleep			X			X
SL5 I sleep badly at night			X			X
Social Isolation						
SO1 I feel lonely						
SO2 I'm finding it hard to contact people						X
SO3 I feel there is nobody I am close to						
SO4 I feel I am a burden to people						X
SO5 I'm finding hard to get on with people						X
Physical Mobility						
PM1 I can only walk about indoors					X	
PM2 I find it hard to bend		X			X	
PM3 I'm unable to walk at all					X	
PM4 I have trouble getting up/down stairs		X			X	
PM5 I find it hard to reach for things		X				
PM6 I find it hard to dress myself					X	
PM7 I find it hard to stand for long					X	
PM8 I need help to walk about outside						

(X) indicates the items included in each dimension of the reduced questionnaires. Items common to the NHP20 and NHP22 questionnaires are shown in **italics**.

Figure 45 - liste des items du NHP38 et des versions courtes NHP20 (TCT) et NHP22 (Rasch) (Prieto et al., 2003)

7.1.4 Rasch analysis

Table 6: Distribution of scores and summary Classical Test Theory (CTT) and Rasch analysis results for the NHP20 and NHP22

	NHP20					NHP22		
	Total score	Emotional	Physical	Sleep	Pain	Total score	Physical	Psychological
Number of items	20	7	5	4	4	22	11	11
Principal components results								
Loadings of the first component ^a	-	0.70	0.68	0.60	0.72	-	0.92	0.92
Distribution of scores								
Valid observations	7,243	7,382	7,442	7,455	7,452	7,559	7,558	7,557
Mean	35.36	30.45	44.72	40.62	28.24	31.04	29.42	28.21
Standard deviation	29.33	31.46	38.10	39.34	36.50	23.84	28.06	27.40
25 th Percentile	10	0	0	0	0	8.77	0	0
50 th Percentile	30	14.29	40	25	0	28.52	23.56	24.22
75 th Percentile	55	57.14	80	75	50	46.94	48.18	46.46
% 0 score	10.8	30.1	27.0	33.0	49.9	15.7	28.1	26.9
% 100 score	2.3	5.5	17.3	20.3	12.0	1.7	3.0	3.0
CTT analysis results								
Item-total correlation (range)	0.45–0.65	0.51–0.62	0.57–0.71	0.51–0.75	0.65–0.68	0.46–0.65	0.47–0.68	0.47–0.64
Reliability								
Cronbach's α	-	0.82	0.83	0.88	0.87	-	0.88	0.87
Linear combination	0.94	-	-	-	-	0.93	-	-
Rasch analysis results								
Person separation	2.17	0.74	0.32	0.00	0.00	2.08	1.39	1.24
Person reliability	0.82	0.35	0.09	0.00	0.00	0.81	0.66	0.61

^a One component accounts of 67.5% of the total variance for the NHP20, and 85% of the total variance for the NHP22.

Figure 46- (Prieto et al., 2003)

8 Annexe 3

8.1 Echelle d'engagement au travail d'Utrecht

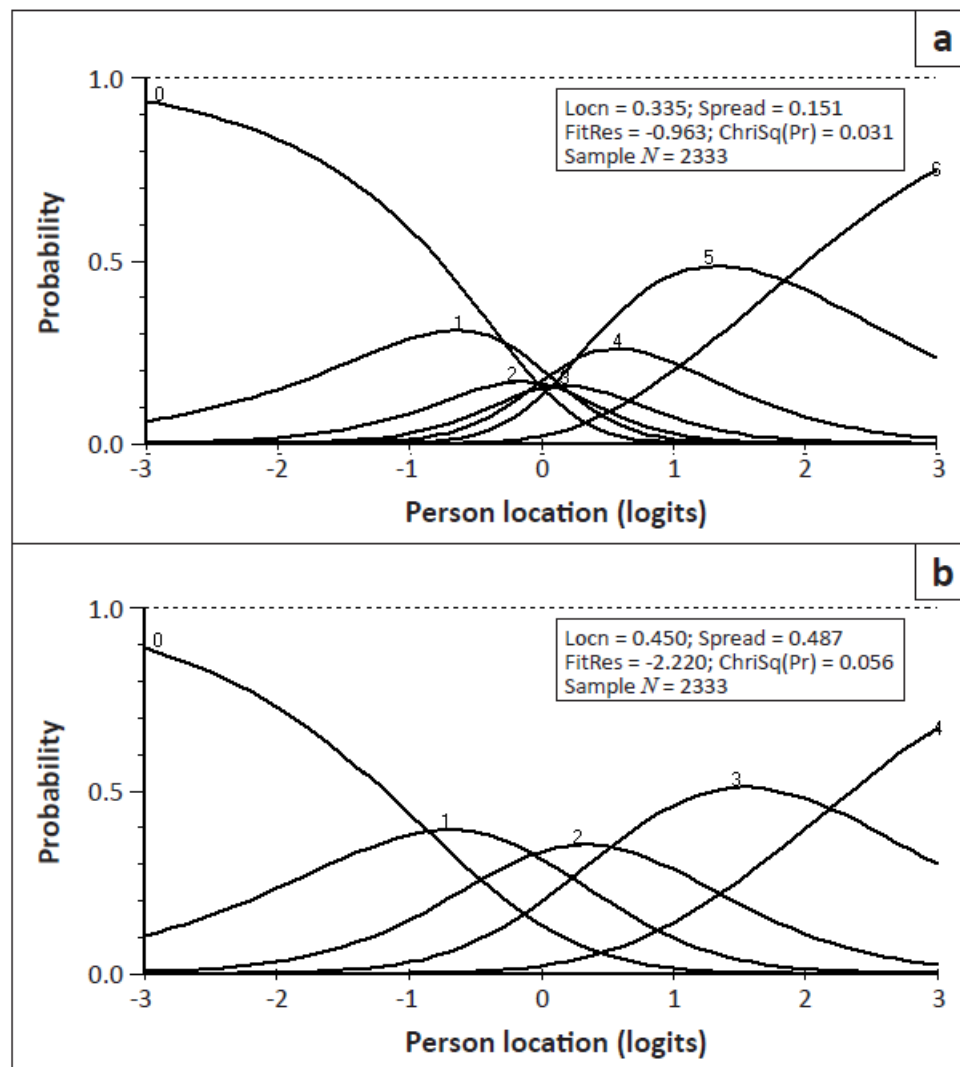


FIGURE 1: (a) Disordered option characteristic curves and thresholds, (b) Ordered option characteristic curves and thresholds after collapsing categories.

Figure 47- Courbes des probabilités des différentes catégories des items polytomiques de l'UWES-17 avant et après recodage (De Bruin et al., 2013)

9 Annexe 4

9.1 Cahier de charge

Priorite	Critère	Satus DICHOTOMIQUE	Status POLY TOMIQUE
Critères généraux			
	permettre la suppression d'un ou plusieurs items de l'analyse		
	valider toutes les analyses avec RUMM sur une meme jeu de donnees		
Priorité 1			
	statistiques moyennes pour les items		
	statistiques moyennes pour les personnes		
	alpha de Chronbach, person separation reliability, strata, d'apres: https://www.rasch.org/rmt/mt63i.htm		
	affichage des statistiques de l'item sur le graphes de chaque ICC		
	lister les items par ordre de numero, localisation, de chi2, en croissant ou decroissant		
	lister les personnes par ordre de numero, localisation, de chi2, en croissant ou decroissant		
	presente la courbe d'information de l'epreuve		
	presente la category probability curve (CPC) pour les items polytomiques en y affichant les stats de l'item comme sur l'ICC, dans ce cas il serait utile d'aligner verticalement l'ICC et la CPC		
	pour les items polytomiques, presenter un tableau des localisation des suiels contenant aussi les stats d'ajustement des seuils, affiche de maniere centralisee ou decentralisee au choix de l'utilisateur		
	permettre le choix d'un modele partial credit ou d'un modele rating scale pour les jeux de donnees polytomiques		
	matrice de correlation des residus standardises a chaque item, ave si possible la possibilite de mettre en evidence les correlation superieure a un seuil, p.ex. >0.3 et <0.3		
Priorité 2			
	graphe de targeting contenant la distribution des sujets, des items et des seuils, avec la courbe d'information si possible		
	graphes ICC et CPC avec les donnees observees par classes d'intervalle (a definir: 3, 4, 5, 6 etc), en evitant que plusieurs personnes de meme aptitude se trouvent dans les classes d'intervalles differentes		
	ANOVA a 2 facteurs sur les residus standardises en fonction du facteur DIF (e.g. le sexe ou l'hopital) et des classes d'intervalles pour analyser le FDI		
Priorité 3			
	indice d'ajustement global		
	tableau de pourcentage $M \Rightarrow C$ et $C \Rightarrow M$		
	pour les items polytomiques		
	ACP sur les residus standardises		
Priorité 4			
	threshold map affichant la reponse des sujets par dessus la reponse attendue a chaque item en fonction de la localisation		

Figure 48- cahier de charge du projet de développement du module d'analyse Rasch en R

10 Annexe 5 - Détails techniques d'implémentation de la solution

Cette annexe développe les aspects techniques utilisés pour solutionner les défis rencontrés.

10.1 Défi : Utilisation de RMarkdown, sortie de tableaux

Au-delà de l'apprentissage des détails de base de RMarkdown, la sortie de tableaux posait souci. Afin de sortir des tableaux présentables, diverses possibilités ont été investiguées.

L'utilisation des fonctionnalités de base de RMarkdown pour sortir des tableaux s'avérait peu pratique. Par conséquent, l'utilisation de la fonction **knitr::kable** a été préférée car elle permet de sortir des données tabulaires efficacement.

En effet, il suffit de soumettre une structure de données tabulaires à la fonction, en n'oubliant pas de configurer les options du chunk de code avec « echo=FALSE » et « results='asis' » afin que le rendu de rapport ne touche plus à la sortie qui est déjà en HTML.

```
```${r comment='', echo=FALSE, results='asis'}
knitr::kable(myDataPersSummaryCooked,
caption = "Statistics of Persons",
floating.environment="sidewaystable")
```
```

10.2 Défi: Configuration du notebook pour les paramètres

Afin de permettre d'utiliser la solution de manière générique et afin que l'utilisateur n'ait pas à modifier le code du notebook à proprement dit, la paramétrisation a été externalisée au sein d'un fichier de paramètres. Il suffit dès lors de référencer ledit fichier de paramètres au sein du notebook. Excepté cette référence et l'entête, il n'y a donc rien à coder pour les utilisateurs futurs dans le notebook.

Ceci étant, il faut que le notebook soit capable de localiser le fichier et donc, il a été nécessaire de mettre en place la configuration de la « working directory » pour ne pas avoir à spécifier de chemins absolus dans le code.

```

1 ---
2 title: "Data Preparation and Analysis"
3 author: "Catherine Absil"
4 date: "07/07/2020"
5 output:
6   html_document:
7     number_sections: yes
8     theme: cerulean
9     toc: yes
10   word_document: default
11 editor_options:
12   chunk_output_type: console
13 ---
14
15 ```{r C1-Environment Cleanup, include=FALSE}
16 rm(list=ls())
17 ```
18
19 ```{r setup, include=FALSE}
20
21 # For debugging https://support.rstudio.com/hc/en-us/articles/205612627-Debugging-with-RStudio#debugging-in-r-markdown-documents
22
23 myWorkDirectory=getwd()
24 # Force all chunks to use the same workdirectory
25 knitr::opts_chunk$set(root.dir = myWorkDirectory)
26 datasetsDirectory=paste(myWorkDirectory, "..", sep="/")
27
28 # Mettre en commentaire ceux qu'on n'utilise pas
29 #params="params_base_dichotomic.R"
30 # params="params_depress_polytomic.R"
31 #params="params_rheumato_polytomic.R"
32 #params="params_stroke_polytomic.R"
33 params="params_wais3_polytomic.R"
34

```

Figure 49- Entête et configuration du répertoire de travail

Les lignes 23 et 25 permettent de configurer le répertoire de travail comme étant le répertoire où se trouve le notebook Rmd.

Ensuite, le fichier de paramètres est identifié (ligne 33) et on peut ainsi le charger dynamiquement à l'aide du paramètre « code » d'un chunk. L'utilisation de « include=TRUE » (ou « include=FALSE ») permettant de (ne pas) faire apparaître les détails techniques dans le rapport. Le « .. » permet de localiser les paramètres au niveau racine du projet, le notebook lui-même étant situé dans un sous-répertoire.

```

62 # Setup check
63
64 working directory is `r myWorkDirectory`
65 datasets are in `r datasetsDirectory`
66
67 ```{r Parameters Definition, code = readLines(paste(myWorkDirectory, "..",
68   params, sep="/")), include=FALSE}
69

```

Figure 50- Localisation du fichier paramètres

Le fichier de paramètres en lui-même contient les éléments paramétrés de l'analyse.

```

1 #
2 # Parameters for Rasch model
3 #
4 #
5 # nom du fichier de donnees avec son extension
6 myDFName <- "DataSet\\dataSetPolytomic0\\VocaWAIS3.xlsx"
7
8 # Fichier de donnees items
9
10 myXlRealItTable <- "DataSet\\dataSetOut\\info_real_itemTable_VocaWAIS3-Generated.xlsx"
11 myXlPersTable <- "DataSet\\dataSetOut\\info_PersonTable_VocaWAIS3-Generated.xlsx"
12
13 #myModelType <- "Dichotomic"
14 myModelType <- "Polytomic" # Pour affichage du type de modèle
15
16 #myModelUsed <- "RSM"
17 myModelUsed <- "PCM"
18
19 myPerformLRtest <- TRUE
20
21 myDFfirstItemCol <- 2 # numero de la colonne contenant l
22 myDFlastItemCol <- 34 # numero de la colonne contenant l
23 myPersonDeletions <- c()
24 # Attention, item indexes are relative. So, first is 1
25 myItemDeletions <- c() #c(14,15,16,52) # liste des items a
26 # c() = ne supprime aucun item
27 # c(3) = supprime le troisieme item dans l'ordre du dataframe de reponses
28 # c(3,6,9) = supprime le 3e, le 6e et le 9e item dans l'ordre du fichier de donnees
29
30 myLatentTrait <- "Trait latent" # Libellé du trait latent dans les graphes
31 myLTmin <- -8 # Localisation minimum (logits) de
32 myLTmax <- +8 # Localisation maximum (logits) de
33 myICCLabel <- "Score" # Libellé de l'ordonnée des CCI/CC
34
35 answerCategories <- c(0,1,2) #Categories de réponses c(0,1) in dichotomic
36 # facteurs de fonctionnement différentiel des items (FDI) pour le facteur 1 (FDI-1)
37 myDIFItodo <- TRUE # TRUE si le test du FDI-1 doit et
38 # FALSE si le test du FDI-1 ne doit pas etre realise
39 myDIFIColumn <- 35 # colonne du fichier de donnees c
40
41 myDIF1InfoColumnName <- "Sex"
42 myDIF1factor <- "Sex" # Nom du facteur FDI-1
43 myDIF1levelA <- 1 # valeur du FDI-1 pour le premier
44 myDIF1labelA <- "Hommes" # libelle du FDI-1 pour le premier
45 myDIF1levelB <- 2 # valeur du FDI-1 pour le deuxieme
46
39:2 (Top Level) R Script

```

Figure 51- Fichier paramètres relatif aux Données du subtest Vocabulaire de la WAIS-III qui doit être complété par l'utilisateur du module d'analyse Rasch en R

10.3 Défi : Extension de packages

Il s'avère que le package **eRm**, bien qu'étant l'un des plus maintenus des packages TRI, souffre de manques lors de la génération de graphiques pour certaines options dont nous avons besoin. Par exemple, la génération des graphes ICC et CPC pour le polytomique.

Il a été nécessaire de trouver une approche pour apporter ces fonctionnalités à **eRm**.

A cette fin, le code R de **eRm** a été exploré afin de comprendre s'il était possible de lui apporter des modifications. Certaines parties sont en Fortran90, ce qui aurait nécessité l'intervention d'un informaticien spécialiste de ce langage. Néanmoins, il s'avère que la structure du package est telle qu'il est possible d'apporter des extensions.

Par conséquent, des extensions nécessaires à la réalisation du module ont ainsi pu être apportées. L'approche est la suivante :

Les extensions sont référencées au sein du notebook via des chunks.

```

69- .....{r Utility Functions, code = readLines(paste(myWorkDirectory, "../caFunctions.R",sep="/"), include=FALSE}
70- .....{r caPlist_Internal, code = readLines(paste(myWorkDirectory, "../caPlist.internal.R",sep="/"), include=FALSE}
71- .....{r caPlotPImap, code = readLines(paste(myWorkDirectory, "../caPlotPImap.R",sep="/"), include=FALSE}
72- .....{r caPlotICC, code = readLines(paste(myWorkDirectory, "../caPlotICC.Rm.R",sep="/"), include=FALSE}
73- .....{r caPlotICC, code = readLines(paste(myWorkDirectory, "../caPlotICC.R",sep="/"), include=FALSE}
74- .....{r caPlotPersonItemHist, code = readLines(paste(myWorkDirectory, "../caPlotPersonItemHist.R",sep="/"), include=FALSE}
75- .....{r caPlotPersonItemHist, code = readLines(paste(myWorkDirectory, "../caPlotPersonItemHist.R",sep="/"), include=FALSE}
76- .....{r caPlotPersonItemHist, code = readLines(paste(myWorkDirectory, "../caPlotPersonItemHist.R",sep="/"), include=FALSE}
77- .....{r caPlotPersonItemHist, code = readLines(paste(myWorkDirectory, "../caPlotPersonItemHist.R",sep="/"), include=FALSE}
78- .....{r caPlotPersonItemHist, code = readLines(paste(myWorkDirectory, "../caPlotPersonItemHist.R",sep="/"), include=FALSE}
79- .....{r caPlotPersonItemHist, code = readLines(paste(myWorkDirectory, "../caPlotPersonItemHist.R",sep="/"), include=FALSE}
80- .....{r caPlotPersonItemHist, code = readLines(paste(myWorkDirectory, "../caPlotPersonItemHist.R",sep="/"), include=FALSE}
81- .....{r caPlotPersonItemHist, code = readLines(paste(myWorkDirectory, "../caPlotPersonItemHist.R",sep="/"), include=FALSE}
82- .....{r caPlotPersonItemHist, code = readLines(paste(myWorkDirectory, "../caPlotPersonItemHist.R",sep="/"), include=FALSE}
83- .....{r caPlotPersonItemHist, code = readLines(paste(myWorkDirectory, "../caPlotPersonItemHist.R",sep="/"), include=FALSE}
84- .....{r caPlotPersonItemHist, code = readLines(paste(myWorkDirectory, "../caPlotPersonItemHist.R",sep="/"), include=FALSE}
85- .....{r caPlotPersonItemHist, code = readLines(paste(myWorkDirectory, "../caPlotPersonItemHist.R",sep="/"), include=FALSE}
86- .....{r caPlotPersonItemHist, code = readLines(paste(myWorkDirectory, "../caPlotPersonItemHist.R",sep="/"), include=FALSE}

```

Figure 52 - Référencement des extensions

On observe ici les extensions d'**eRm** qui ont été nécessaires afin de produire une version spécifique de **caPlotPImap**, **caPlotCC** et **caPlotPersonItemHist**. On observe 2 fichiers pour **caPlotICC** car le second est une redirection vers le premier. Les détails de ces éléments sont développés dans les sections pertinentes.

10.4 Défi : Performance

Le calcul des modèles et l'exécution de LRT est devenue longue (au moins 20 minutes) dans les cas où l'on trouve des « NA » dans les données. , Certaines possibilités de réduction du temps d'exécution ont été examinées

Sous Windows, il existe le package **doParallel** qui permet de paralléliser les boucles dans le code.

```

library(doParallel)
cl <- makeCluster(detectCores(), type='PSOCK')
registerDoParallel(cl)

## turn parallel processing off and run sequentially again:
#registerDoSEQ()

```

Figure 53- Mise en place du package par **doParallel**

Ce package permet d'utiliser les divers cœurs du processeur pour les boucles. Une légère amélioration du temps de calcul a été observée en fonction du type d'ordinateur ou de processeur utilisés.

10.5 Défi : Gestion des valeurs manquantes et conversion en NA

Certains jeux de données contiennent des étoiles « * » pour indiquer une valeur manquante. Il convenait donc de les transformer en « NA » afin de pouvoir travailler avec ledit jeu de données.

Une fonction de chargement du jeu de données a donc été développée depuis le fichier Excel qui les contient afin de convertir les « * » en « NA ».

```

1 library(xlsx)
2
3 # Read contents from first Excel worksheet as a dataframe
4 caReadDataset <- function(datasetDir, inputFileName) {
5   fullFilePath <- paste(datasetDir, inputFileName, sep = "/")
6   normalizedFullFilePath <- normalizePath(fullFilePath, winslash = "\\", mustWork = NA)
7   contentsDf = read.xlsx(normalizedFullFilePath,
8                           sheetIndex=1,
9                           header = TRUE)
10
11   # Replace *s with NAs
12   is.na(contentsDf) <- contentsDf == "*"
13
14   # Coerce columns to numeric (since *s were making them char)
15   contentsDf <- sapply( contentsDf, as.numeric )
16   return(contentsDf)
17 }
18
19 caCountForItemNamedAndValue<- function(df,itemname,value) {
20   return(sum(df[,itemname] == value, na.rm=TRUE))
21 }
22
23 caOccurrencesOfAnswersForItemNamed <- function(df, answersVector, itemname) {
24   return(sapply(answersVector,caCountForItemNamedAndValue, df=msi, itemname=itemname))
25 }
26

```

Figure 54 -caFunctions.R - transformation des * en NA

Si les NA sont représentés par un autre caractère, il suffit de le modifier dans cette routine.

10.6 Défi : Selection de la plage de données souhaitée

La plage de données souhaitée est sélectionnée à l'aide d'une étendue de valeurs, permettant d'identifier les colonnes contenant les items. Les données sont supposées disposer d'un en-tête et commencer à la ligne 2.

Comme le fichier lu est au format Excel, les données sont lues jusqu'à la première ligne vide. Les personnes sont en lignes et la première colonne contient l'identifiant de la personne. Les données sont lues au sein d'une dataframe R.

10.7 Défi : Elimination des personnes et items non souhaités

Certains items et personnes du jeu de données doivent pouvoir être éliminés des données lues en entrée. Ces informations sont déclarées au sein du fichier de paramètres.

```

myDFfirstItemCol <- 2          # numero de la colonne contenant les reponses
                                au premier item
myDFlastItemCol <- 13         # numero de la colonne contenant les
                                reponses au dernier item
myPersonDeletions <- c(17,62,70,89,175) # numéro de ligne de la personne au
                                sein de la dataframe
myItemDeletions <- c()        # numéro de colonne de l'item au sein de la dataframe
# liste des items a supprimer par numero d'ordre, p. ex.:
# c() = ne supprime aucun item
# c(3) = supprime le troisieme item dans l'ordre du dataframe de reponses
# c(3,6,9) = supprime le 3e, le 6e et le 9e item dans l'ordre du fichier de
                                donnees

```

```

# Attention with item deletions, we are using relative indexes then.
myResponses = myData[,myDFfirstItemCol:myDFlastItemCol]
rownames(myResponses) <- myData[, "Person"] ## !!! Must have a column named "Person"

# person deletions if needed
if (length(myPersonDeletions) == 0) {
  mySelectedItems <- myResponses
} else {
  #TODO: Rownames are not staying synched with the original
  mySelectedItems <- myResponses[-myPersonDeletions,]
}

# item deletions if needed
if (length(myItemDeletions) == 0) {
  # Do nothing mySelectedItems <- myResponses
} else {
  #TODO: Rownames are not staying synched with the original
  mySelectedItems <- mySelectedItems[,-myItemDeletions]
}

```

Figure 55 - Selection des données pertinentes

mySelectedItems représente les données conditionnées sur base des suppressions.

10.8 Défi : Rescoring

Le rescoring permet de transformer certaines valeurs des cellules vers d'autres sur base d'une table de correspondance. Cette table est déclarée au sein des paramètres et est appliquée aux données. Les paramètres déterminent si le rescoring doit avoir lieu ou pas. La réalisation du rescoring s'effectue à l'aide de la fonction **mapvalues** (librairie R **plyr**)

```

myRescoring <- TRUE
myRescoringFrom <- c(0,1,2,3,4)
myRescoringTo <- c(0,1,1,2,2)

```

Figure 56- Déclaration du rescoring

```

# Rescoring
Rescoring: 'myRescoring'
```{r rescoring}
require(plyr)

if (myRescoring) {
 dfRescoring <- as.data.frame(mySelectedItems)
 for (i in names(dfRescoring)) {
 replaced <- mapvalues(dfRescoring[[i]], myRescoringFrom, myRescoringTo)
 dfRescoring[[i]] <- replaced
 }

 print("Input Extract (first 5)")
 head(mySelectedItems, 5)

 mySelectedItems <- as.matrix(dfRescoring)

 print("Rescored Extract (first 5)")
 head(mySelectedItems, 5)
}
```

```

Figure 57 - Rescoring avec **plyr** et **mapvalues**

10.9 Défi : Comptage des éléments

Il est nécessaire de disposer d'une vue sur le nombre d'occurrences des différentes réponses aux items au sein du jeu de données. A cette fin, une approche qui permet de compter les types de réponse par item et de les sortir via une table **kable** a été mise en place. Cette table permet de se faire une idée de la distribution des réponses notamment pour les cas extrêmes.

```
caCountForItemNamedAndValue<- function(df,itemname,value) {
  return(sum(df[,itemname]== value, na.rm=TRUE))
}

caOccurencesOfAnswersForItemNamed <- function(df, answersVector, itemname) {
  return(sapply(answersVector,caCountForItemNamedAndValue, df=msi, itemname=itemname))
}
```

Figure 58 - Fonctions de support au comptage (caUtils.R)

Dans le code de comptage, il convient d'observer que **answerCategories** vient du fichier de paramètres et doit contenir les valeurs telles qu'après rescoring.

```
172. # Counts
173.
174. '''{r counts}'''
175. msi<-data.frame(mySelectedItems)
176. occurencesDf <- data.frame(sapply(colnames(msi),caOccurencesOfAnswersForItemNamed, df=msi, answersVector=answerCategories))
177. rownames(occurencesDf)<-answerCategories
178.
179. totalOfItemNamed <- function(itemName) {
180.   return(sum(occurencesDf[,itemName]))
181. }
182.
183. occurencesTotals <- sapply(names(occurencesDf),totalOfItemNamed)
184. '''
185. '''{r occurencesTable, comment='Occurences Table', echo=FALSE, results='asis'}'''
186. knitr::kable(occurencesDf,
187.   caption = "Occurences",
188.   floating.environment="sidewaystable")
189.
190. knitr::kable(occurencesTotals,
191.   caption = "Occurences Totals",
192.   floating.environment="sidewaystable")
193.
194. '''
195. '''
```

Figure 59 - Comptage des éléments et somme totale pour verification

| Occurences | | | | | | | | | | | | | | | | | | |
|-------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| | item1 | item2 | item3 | item4 | item5 | item6 | item7 | item8 | item9 | item10 | item11 | item12 | item13 | item14 | item15 | item16 | item17 | item18 |
| 0 | 3 | 8 | 6 | 50 | 42 | 82 | 87 | 99 | 241 | 155 | 228 | 262 | 293 | 316 | 402 | 245 | 312 | 265 |
| 1 | 18 | 1 | 45 | 178 | 287 | 430 | 951 | 127 | 442 | 122 | 476 | 398 | 415 | 304 | 301 | 784 | 143 | 115 |
| 2 | 1083 | 1095 | 1053 | 876 | 775 | 592 | 66 | 878 | 421 | 400 | 444 | 396 | 484 | 401 | 75 | 649 | 724 | |
| Occurences Totals | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | x |
| item1 | | | | | | | | | | | | | | | | | | 1104 |
| item2 | | | | | | | | | | | | | | | | | | 1104 |
| item3 | | | | | | | | | | | | | | | | | | 1104 |

Figure 60 - Print screen des totaux et compte de vérification

10.10 Défi : Calcul du fit et des thresholds sur base de l'algorithme choisi

Sur base du type de modèle à « fitter »¹⁴ dans le fichier de paramètres, on détermine les paramètres estimés du modèle sur base des données sélectionnées (dans `mySelectedItems`). Les paramètres sont estimés sur base des réponses au travers d'une estimation CML (Conditional Maximum Likelihood) qui est l'approche du package **eRm** dont proviennent ces fonctions.

En ce qui concerne le modèle de type RM (dichotomique) :

- Une capture des estimations de paramètres du modèle est réalisée au travers de la fonction `RM`
- Il n'y a pas de calcul pas de seuils (thresholds)

En ce qui concerne les modèles de type RSM et PCM (polytomique):

- On applique la fonction portant le même nom (RSM ou PCM) pour obtenir les paramètres du modèle ;
- On détermine les seuils à l'aide de la fonction `thresholds` de **eRm** (`myItemThresholds`) ;
- On isole les valeurs desdits seuils dans une dataframe plus facile à exploiter par la suite (`myItemThresholdsTable`).

Le calcul peut être long. Afin de se faire une idée du temps de calcul, on garde trace du temps passé à l'aide d'appels à **Sys.time**.

```
start.time.model <- Sys.time()

# TODO : ajouter test type modèle

# estimate parameters for selected items according to dichotomous model
if (myModelUsed == "RM") {
  # Note: for RM model, thresholds does not work
  myModelParameterEstimates = RM(mySelectedItems)
  # myItemThresholdsTable <- cbind(myModelParameterEstimates$betapar, myModelParameterEstimates$betapar) * -#
  # betapars are easiness parameters
  # rownames(myItemThresholdsTable) <- substring(rownames(myItemThresholdsTable), first=6, last=9999)
}

# estimate parameters for selected items according to polytomous model
if (myModelUsed == "RSM"){
  myModelParameterEstimates = RSM(mySelectedItems)
  myItemThresholds <- thresholds(myModelParameterEstimates)
  myItemThresholdsTable <- data.frame(myItemThresholds$threshtable[[1]])
} else if (myModelUsed == "PCM"){
  myModelParameterEstimates = PCM(mySelectedItems)
  myItemThresholds <- thresholds(myModelParameterEstimates)
  myItemThresholdsTable <- data.frame(myItemThresholds$threshtable[[1]])
} else {
  stop("Model must be RSM or PCM")
}

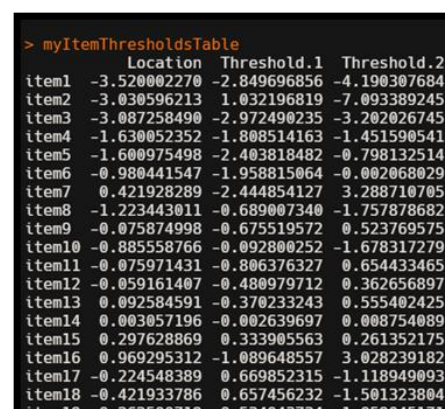
end.time.model <- Sys.time()
time.taken.model <- round(end.time.model - start.time.model,2)
```

Figure 61 - Estimation des paramètres

¹⁴ « fitter » pour model fit

On obtient les paramètres beta des items dans le modèle. Les seuils, si applicables, sont repris dans le tableau des seuils.

En ce qui concerne les personnes, il faut utiliser une autre fonction, détaillée plus loin (**person.parameters**).



```
> myItemThresholdsTable
```

| | Location | Threshold.1 | Threshold.2 |
|--------|--------------|--------------|--------------|
| item1 | -3.520002270 | -2.849696856 | -4.190307684 |
| item2 | -3.030596213 | 1.032196819 | -7.093389245 |
| item3 | -3.087258490 | -2.972490235 | -3.202026745 |
| item4 | -1.630052352 | -1.808514163 | -1.451590541 |
| item5 | -1.600975498 | -2.403818482 | -0.798132514 |
| item6 | -0.980441547 | -1.958815064 | -0.002068029 |
| item7 | 0.421928289 | -2.444854127 | 3.288710705 |
| item8 | -1.223443011 | -0.689007340 | -1.757878682 |
| item9 | -0.075874998 | -0.675519572 | 0.523769575 |
| item10 | -0.885558766 | -0.092800252 | -1.678317279 |
| item11 | -0.075971431 | -0.806376327 | 0.654433465 |
| item12 | -0.059161407 | -0.480979712 | 0.362656897 |
| item13 | 0.092584591 | -0.370233243 | 0.555402425 |
| item14 | 0.003057196 | -0.002639697 | 0.008754089 |
| item15 | 0.297628869 | 0.333905563 | 0.261352175 |
| item16 | 0.969295312 | -1.089648557 | 3.028239182 |
| item17 | -0.224548389 | 0.669852315 | -1.118949093 |
| item18 | -0.421933786 | 0.657456232 | -1.501323804 |

Figure 62- Tableau des locations et seuils des items

10.11 Défi : Réalisation du LR test de Andersen

La fonction **LRtest** permet d'évaluer l'adéquation du modèle aux données modélisées et d'aider à la critique du modèle sur une base rationnelle. C'est un test global examinant tous les items simultanément. Il est puissant en ce qui concerne les violations des critères d'application du modèle de Rasch. Il peut détecter le biais dans les items (aussi nommé DIF pour « Differential Item Functioning », c'est-à-dire qu'un item va fonctionner différemment pour des individus ayant cependant un trait latent de même niveau. Citons l'âge et le genre comme caractéristiques des personnes pouvant répondre à cet aspect). L'approche CML étant celle du package **eRm**, ce test est applicable pour tester notre modèle.

10.11.1 Idée de base

L'idée de base autour de l'utilisation du test est que les paramètres estimés devraient aussi bien fonctionner pour l'un ou l'autre sous-groupe des données. Par exemple, si on sépare la population entre « les scores inférieurs ou égaux à la moyenne » et « les scores supérieurs strictement à la moyenne », il ne devrait pas y avoir de différence significative entre les paramètres estimés du modèle.

S'il y a une différence significative, cela indique alors que les paramètres estimés ne donnent pas un modèle utilisable et pertinent.

On peut imaginer séparer la population en groupes divers et variés et toujours retrouver le même ensemble de paramètres (invariance des paramètres par rapport aux groupements).

10.11.2 Mise en oeuvre

On passe le modèle estimé à la fonction **LRtest**, ainsi qu'un critère de découpe de la population. Ce test du rapport de vraisemblance (Likelihood Ratio Test) considère le découpage des personnes en sous-groupes. Par défaut, le critère de découpage (**splitcr**) est la médiane (median) ou la moyenne (mean) dans la fonction **LRtest** du package **eRm**.

LRtest peut sortir des alertes (warnings).

On voit dans la sortie que certains items posent souci. Il s'agit dès lors d'examiner pourquoi. Par exemple dans le cas examiné, item1, item 2 et item 4 sont à 2 pour presque tous les sujets. Ce sont des items très faciles. On peut directement repérer des items très faciles ou très difficiles dans le tableau des occurrences établi préalablement.

Occurrences

| | item1 | item2 | item3 | item4 | item5 | item6 | item7 | item8 | item9 | item10 | item11 | item12 | item13 | item14 | item |
|---|-------|-------|-------|-------|-------|-------|-------|-------|-------|--------|--------|--------|--------|--------|------|
| 0 | 3 | 8 | 6 | 50 | 42 | 82 | 87 | 99 | 241 | 155 | 228 | 262 | 293 | 316 | 4 |
| 1 | 18 | 1 | 45 | 178 | 287 | 430 | 951 | 127 | 442 | 122 | 476 | 398 | 415 | 304 | 3 |
| 2 | 1083 | 1095 | 1053 | 876 | 775 | 592 | 66 | 878 | 421 | 827 | 400 | 444 | 396 | 484 | 4 |

Figure 63 - Items faciles ou difficiles repérés dans le tableau des occurrences

Si on teste sur la moyenne ou la médiane, on retrouve les groupes sous \$low et \$high dans le resultat.

Le résultat de **LRtest** (lrres), contient, entre autres, la pvalue associée au test, ainsi que le nombre de degrés de liberté (df) du modèle testé l(lrres\$pvalue, lrres\$df).

Le calcul du **LRtest** peut être long, surtout pour des jeux de données contenant des NA et/ou beaucoup d'items. C'est pour cela que son calcul est rendu optionnel par une variable du jeu de paramètres (myPerformLRtest). Néanmoins, il est primordial de l'effectuer pour être au clair avec la qualité de l'ajustement.

Afin de simplement exécuter le **LRtest**, on peut, outre « mean » et « median », utiliser un vecteur de découpage (**splitcr**) de longueur identique au nombre de personnes dans la sélection avec une valeur 1 ou 2 aléatoire.

Ceci permet de s'assurer qu'il est possible d'exécuter **LRtest** sur l'échantillon si on ne peut pas utiliser « mean » ou « median » suite à des données ne le permettant pas.

Le **LRtest** est utile pour les calculs des DIF, cette version avec un vecteur « aléatoire » étant une vérification technique pour identifier les items à problème « évidents ».

A toutes fins utiles, le **LRtest** génère des modèles de fit spécifiques à chaque sous groupe fourni dans le splitcr (lrres\$fitobj)

Au niveau des résultats, on retrouve aussi les erreurs standards du fit ventilées par groupe au sein de `lrres$selist`. Les paramètres estimés par groupe se trouvent eux dans `$betalist` et `$etalist`. On peut ainsi facilement accéder aux détails au niveau des groupes. Par exemple, pour les erreurs standard d'estimation, on utilisera `lrres$selist[[1]]` pour le groupe 1 sur base d'un groupement spécifique demandé via un `splittr`.

Figure 64 - Analyse d'un subset par rapport au fit

```
> splitvec <- sample(1:2, nrow(myResponses), replace=TRUE)
> lrres <- LRtest(myModelParameterEstimates, splittr=splitvec)
Warning in LRtest.Rm(myModelParameterEstimates, splittr = splitvec) :
The following items were excluded due to inappropriate response patterns within subgroups:
item2
Full and subgroup models are estimated without these items!
> lrres$selist
$'1'
[1] 1.03799961 0.90400161 0.62152743 0.50751247 0.24680844 0.23624846 0.24153738 0.24105150 0.17915301 0.18667914 0.16372138 0.24823873 0.20195450 0.10360121 0.12533362 0.14209183
[17] 0.17192953 0.15195148 0.12385271 0.14383872 0.12689959 0.14004227 0.11966018 0.13726265 0.12432707 0.13120905 0.11845621 0.12887922 0.11134517 0.21129529 0.15217254 0.12958671
[33] 0.16985410 0.13276334 0.15193865 0.13872642 0.11246889 0.14011269 0.12283263 0.12655523 0.14925428 0.12252656 0.14355998 0.12919900 0.11299798 0.13734479 0.11095834 0.13624308
[49] 0.11817792 0.12995278 0.12083369 0.15538971 0.09905868 0.18220434 0.11335565 0.14022825 0.13381985 0.12850791 0.16589856 0.15441172 0.12566724 0.13595154 0.33588471 0.18649673
$'2'
[1] 0.85398532 0.77334753 0.64872821 0.61544275 0.22784244 0.21956573 0.24325872 0.24095509 0.17713250 0.18445579 0.16479665 0.24348969 0.19553411 0.16890517 0.11724333 0.13643898
[17] 0.10627751 0.15012031 0.11979628 0.13939826 0.11470440 0.13278364 0.11210501 0.13264710 0.11797443 0.12700601 0.11200960 0.12633852 0.11200333 0.10467613 0.14526713 0.12331770
[33] 0.15704382 0.12795773 0.14156642 0.12476850 0.11115655 0.13456431 0.12764453 0.12047289 0.14263647 0.11897597 0.15989198 0.12629866 0.11031519 0.13221418 0.10621893 0.13200592
[49] 0.11947204 0.12775716 0.11903806 0.15189179 0.09599667 0.21655045 0.11352745 0.12924036 0.12136992 0.12591112 0.16484208 0.15918319 0.12435288 0.13132099 0.33550717 0.20551713
```

Figure 65 - Exemple des erreurs standard de fit avec 2 groupes (exemple technique)

Ces données sont extraites de l'estimation des paramètres du modèle total (`myModelParameterEstimates`).

10.11.3 Interprétation de la pvalue

La p-value que donne le `LRtest` donne des indications sur la qualité du modèle. Le likelihood ratio est effectué entre 2 populations et devrait être proche de 1. Attention LR ratio, c'est $LR(pop1)/LR(pop2)$.

10.11.4 Evaluation graphique du Goodness of Fit

La qualité du fit du modèle aux données peut être vérifié graphiquement. On utilise alors la fonction **plotGOF** du package **eRm**. Un graphique GOF représente les paramètres beta estimés sur l'axe *x* et l'axe *y*.

L'axe *x* correspond à un sous groupe A et l'axe *y* à un sous groupe B de la population (par exemple, découpage basé sur les scores ou au hasard d'un échantillonnage, ou encore d'un critère pour un DIF).

Il existe 3 variantes du `plotGOF`, soit basique, soit avec des ellipses, soit avec des bandes de confiance. Si un item est significativement éloigné de la droite $y = x$, cela signifie qu'il n'y a pas d'homogénéité des paramètres bêta entre les sous-groupes, ce qui est un problème. On passe le résultat de `LRtest` à `plotGOF` pour obtenir le résultat graphique.

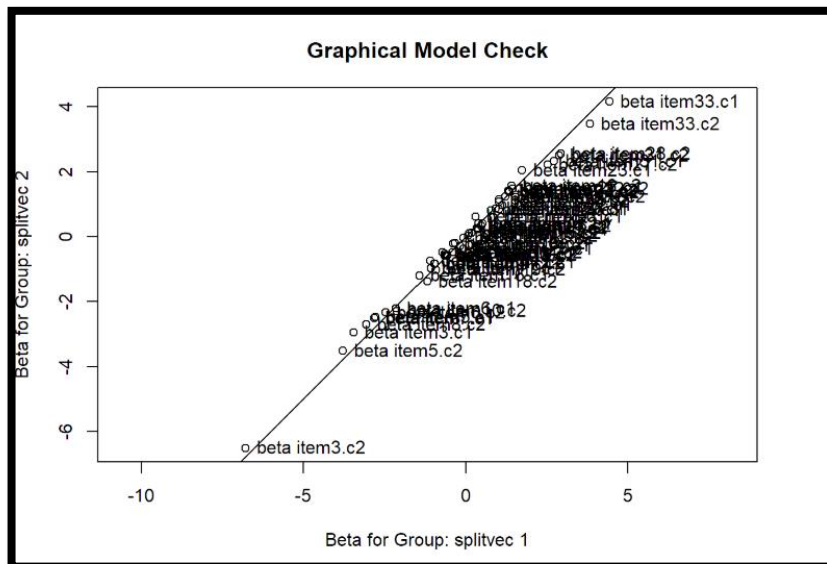


Figure 66 - Exemple de fit correct

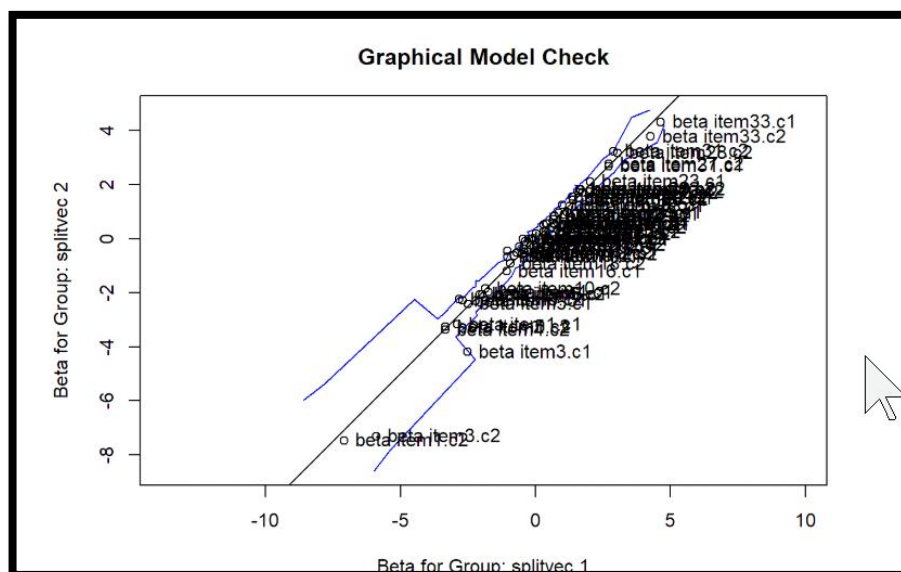


Figure 67 - Exemple de « fit » à problème avec bandes de confiance (item 1 et 3)

10.12 Défi : Person Parameters

Il convient d'estimer les paramètres pour les personnes de la population (theta).

A cette fin, on exploite la fonction `person.parameters` à qui l'on passe le modèle de fit.

```
myPersonParametersEstimates = person.parameter(myModelParameterEstimates)
plot(myPersonParametersEstimates)
myPersonParametersEstimatesConfint <- confint(myPersonParametersEstimates)
```

Figure 68 - Calcul des paramètres des personnes et de l'intervalle de confiance de ceux-ci

Cette fonction détermine les paramètres des personnes selon une approche du maximum de vraisemblance (maximum likelihood).

La fonction **plot** appliquée à un résultat de `person.parameters` représente les valeurs des paramètres des personnes en fonction des scores des personnes.

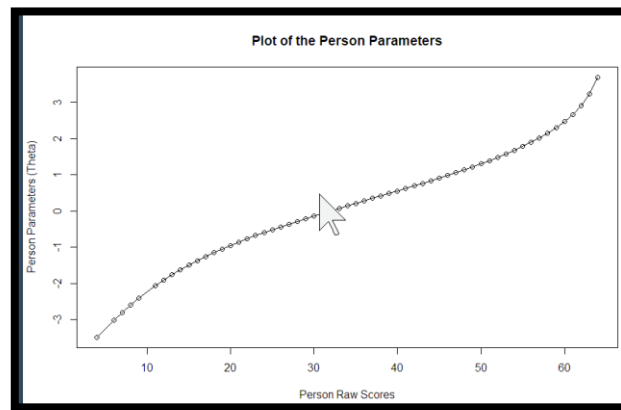


Figure 69 - Plot de `person.parameters`

Une erreur est associée à chaque entrée. La structure des données est la suivante :

| | | |
|---------------------------|----------------------------------|---|
| myPersonParametersEsti... | list [18] (S3: ppar) | List of length 18 |
| X | double [1104 x 33] | 2.2 2.2 2.2 2.2 2.2 2.2 2.2 2.2 2.2 1.1 1.2 2.2 2.2 2.2 1.2 2.2 2.2 1.1 2.1 1.1 1.1 ... |
| X01 | double [1104 x 66] | 0.0 0.0 0.0 1.1 1.1 1.1 1.0 0.0 0.0 0.1 1.1 1.1 1.0 0.0 0.0 0.1 1.1 1.1 1.1 1.0 ... |
| X.ex | double [1104 x 33] | 2.2 2.2 2.2 2.2 2.2 2.2 2.2 2.2 2.2 1.1 1.2 2.2 2.2 2.2 1.2 2.2 2.2 1.1 2.1 1.1 1.1 ... |
| W | double [66 x 65] | -1.1 0.0 0.0 -1.0 1.0 0.0 -1.0 0.1 0.0 -1.0 0.0 1.0 -1.0 0.0 1.0 -1.0 0.0 ... |
| model | character [1] | 'PCM' |
| loglik | double [1] | -1060.489 |
| loglik.cml | double [1] | -22598.94 |
| npar | double [1] | 59 |
| iter | double [1] | 16 |
| betapar | double [66] | 2.85 7.04 -1.03 6.06 2.97 6.17 ... |
| thetapar | list [1] | List of length 1 |
| se.theta | list [1] | List of length 1 |
| theta.table | list [1104 x 3] (S3: data.frame) | A data.frame with 1104 rows and 3 columns |
| Person Parameter | double [1104] | 0.6232 1.0600 -0.5972 0.0715 0.7643 -0.2128 ... |
| NAgroup | integer [1104] | 1 1 1 1 1 ... |
| Interpolated | logical [1104] | FALSE FALSE FALSE FALSE FALSE FALSE ... |
| pred.list | list [1] | List of length 1 |
| hessian | list [1] | List of length 1 |
| mpoints | integer [1] | 1 |
| pers.ex | integer [0] | |
| gmemb | integer [1104] | 1 1 1 1 1 ... |

Figure 70 - Structure du résultat de `person.parameters`

On voit la table des paramètres des personnes et aussi le champ « interpolated » qui permet de savoir si un paramètre a été interpolé suite à la rencontre d'une valeur « NA » dans les réponses. On peut sortir les intervalles de confiance des paramètres des personnes à l'aide de la fonction **confint**.

```
> myPersonParametersEstimatesConfint
$NAgroup1
```

| | 2.5 % | 97.5 % |
|------|-------------|-------------|
| 1101 | 0.10526418 | 1.14107012 |
| 1102 | 0.51573514 | 1.60426197 |
| 1103 | -1.15779163 | -0.03657671 |
| 1104 | -0.44552332 | 0.58857334 |
| 1105 | 0.24053715 | 1.28800039 |
| 1106 | -0.74226735 | 0.31674406 |
| 1107 | 0.80658492 | 1.96921378 |

Figure 71 - Intervalle de confiance des paramètres des personnes

Il convient alors d'examiner les résidus de ces estimations de paramètres des personnes.

```
myPersonParametersEstimatesResiduals <- residuals(myPersonParametersEstimates)
```

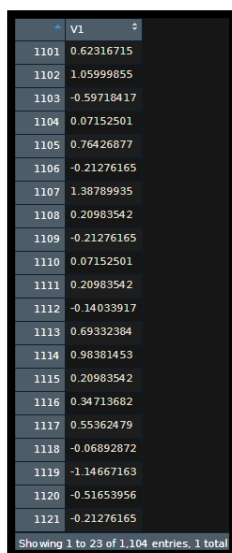
Figure 72 - Extraction des résidus pour les paramètres des personnes

10.13 Défi : Evaluation du niveau du trait latent des personnes

Sur base de l'estimation des paramètres des personnes, on peut sortir la valeur du trait latent pour chaque personne. On utilise à cette fin la methode **coef** qui retourneun vecteur des estimations des paramètres des personnes.

```
myPersonAbilityEstimates <- coef(myPersonParametersEstimates)
```

Figure 73 - Evaluation des valeurs du trait latent des personnes



| | V1 |
|------|-------------|
| 1101 | 0.62316715 |
| 1102 | 1.05999855 |
| 1103 | -0.59718417 |
| 1104 | 0.07152501 |
| 1105 | 0.76426877 |
| 1106 | -0.21276165 |
| 1107 | 1.38789935 |
| 1108 | 0.20983542 |
| 1109 | -0.21276165 |
| 1110 | 0.07152501 |
| 1111 | 0.20983542 |
| 1112 | -0.14033917 |
| 1113 | 0.69332304 |
| 1114 | 0.98381453 |
| 1115 | 0.20983542 |
| 1116 | 0.34713682 |
| 1117 | 0.55362479 |
| 1118 | -0.06892872 |
| 1119 | -1.14667163 |
| 1120 | -0.51653956 |
| 1121 | -0.21276165 |

Showing 1 to 23 of 1.104 entries, 1 total

Figure 74 - Exemple de valeurs calculées pour le trait latent des personnes

On pourra alors grouper les personnes par niveau de trait latent sur cette base.

10.14 Défi : Calcul de l'itemfit et de personfit

On peut obtenir les statistiques sur les items et les personnes via les fonctions **itemfit** et **personfit** de **eRm**. On passe dans les deux fonctions le résultat obtenu via **person.parameters**.

```

> myItemfit <- itemfit(myPersonParametersEstimates)
> myItemfit

Itemfit Statistics:
      Chisq  df p-value Outfit MSQ Infit MSQ Outfit t Infit t Discrim
item1  487.494 1103   1.000   0.442   0.812  -1.804  -0.804   0.282
item2  637.578 1103   1.000   0.578   1.047   0.033   0.258   0.179
item3 1201.364 1103   0.020   1.088   1.080   0.458   0.594   0.220
item4  832.069 1103   1.000   0.754   0.878  -2.840  -2.017   0.479
item5 1139.276 1103   0.218   1.032   1.017   0.354   0.354   0.387
item6 1091.495 1103   0.592   0.989   0.983  -0.238  -0.410   0.444
item7  974.949 1103   0.998   0.883   0.900  -1.583  -1.474   0.385
item8 1151.303 1103   0.152   1.043   1.119   0.421   1.937   0.359
item9 1319.370 1103   0.000   1.195   1.159   4.645   4.361   0.351
item10 1059.092 1103   0.825   0.959   1.019  -0.378   0.377   0.451
item11  992.931 1103   0.992   0.899   0.908  -2.690  -2.703   0.546
item12  972.776 1103   0.998   0.881   0.908  -2.977  -2.727   0.557
item13 1133.520 1103   0.255   1.027   1.034   0.677   1.005   0.456
item14 1270.122 1103   0.000   1.150   1.073   3.086   2.120   0.447
item15 1044.038 1103   0.897   0.946   0.970  -1.192  -0.907   0.515
item16 1127.666 1103   0.296   1.021   1.013   0.465   0.297   0.337
item17 1429.524 1103   0.000   1.295   1.162   3.892   4.161   0.402
item18 1220.086 1103   0.008   1.105   1.128   1.255   3.025   0.422

```

Figure 75- Exemple d'item fit

```

> myPersonfit <- personfit(myPersonParametersEstimates)
> myPersonfit

Personfit Statistics:
      Chisq  df p-value Outfit MSQ Infit MSQ Outfit t Infit t
1101  29.288  32   0.605   0.888   0.888   0.899   0.28  -0.40
1102  34.138  32   0.366   1.034   0.941   0.56  -0.17
1103  25.245  32   0.796   0.765   0.949   0.29  -0.12
1104  24.336  32   0.832   0.737   0.911  -0.12  -0.34
1105  21.537  32   0.919   0.653   0.927   0.11  -0.26
1106  32.074  32   0.463   0.972   1.158   0.18  -0.73
1107  27.059  32   0.715   0.820   1.105   0.60  0.46
1108  14.375  32   0.997   0.436   0.545  -0.50  -2.42
1109  12.850  32   0.999   0.389   0.466  -0.99  -2.83
1110  24.797  32   0.814   0.751   0.914  -0.10  -0.32
1111  28.513  32   0.644   0.864   0.991   0.11  0.03
1112  17.169  32   0.985   0.520   0.632  -0.63  -1.78
1113  26.441  32   0.762   0.789   1.028   0.21  0.20
1114  27.197  32   0.708   0.824   0.931   0.38  -0.22
1115  50.868  32   0.018   1.541   1.003   0.83  0.09
1116  21.961  32   0.908   0.665   0.852  -0.10  -0.64
1117  30.701  32   0.522   0.900   0.906   0.30  0.01
1118  36.687  32   0.260   1.112   1.322   0.39  1.37
1119  26.385  32   0.746   0.800   1.126  -0.04  0.49
1120  20.860  32   0.935   0.632   0.824  -0.47  -0.67
1121  29.182  32   0.614   0.882   1.041   0.25  0.25
1122  15.715  32   0.993   0.476   0.661  -0.78  -1.28

```

Figure 76- Exemple de personfit

Ces informations permettent ensuite de déterminer si les items et les personnes sont pertinents pour la détermination de l'échelle.

10.15 Défi : Confection du tableau des items

Afin de pouvoir présenter les items (et les trier sur des critères utiles à l'analyse), on a besoin d'assembler les diverses valeurs déterminées au sein d'une table synoptique.

Ceci est effectué en récupérant les diverses données depuis les résultats intermédiaires et aboutit à une structure telle que la suivante :

```

> myRealItemTable

      Location Infit.MNSQ Infit.ZSTD Outfit.MNSQ Outfit.ZSTD Chisq df pvalue Label
item1  -3.520    0.812    -0.804    0.442    -1.804  487.5 1103  1.000 item1
item2  -3.031    1.047    0.258    0.578    0.033  637.6 1103  1.000 item2
item3  -3.087    1.080    0.594    1.088    0.458 1201.4 1103  0.020 item3
item4  -1.630    0.878    -2.017    0.754    -2.840  832.1 1103  1.000 item4
item5  -1.601    1.017    0.354    1.032    0.515 1139.3 1103  0.218 item5
item6  -0.980    0.983    -0.410    0.989    -0.238 1091.5 1103  0.592 item6
item7   0.422    0.900    -1.474    0.883    -1.583  974.9 1103  0.998 item7
item8  -1.223    1.119    1.937    1.043    0.421 1151.3 1103  0.152 item8
item9  -0.076    1.159    4.361    1.195    4.645 1319.4 1103  0.000 item9
item10 -0.886    1.019    0.377    0.959    -0.378 1059.1 1103  0.825 item10
item11 -0.076    0.908    -2.703    0.899    -2.690  992.9 1103  0.992 item11
item12 -0.059    0.908    -2.727    0.881    -2.977  972.8 1103  0.998 item12
item13  0.093    1.034    1.005    1.027    0.677 1133.5 1103  0.255 item13
item14  0.003    1.073    2.120    1.150    3.086 1270.1 1103  0.000 item14
item15  0.298    0.970    -0.907    0.946    -1.192 1044.0 1103  0.897 item15
item16  0.969    1.013    0.297    1.021    0.465 1127.7 1103  0.296 item16
item17 -0.225    1.162    4.161    1.295    3.892 1429.5 1103  0.000 item17
item18 -0.422    1.128    3.025    1.105    1.255 1220.1 1103  0.008 item18
item19 -0.263    1.063    1.655    1.127    1.793 1244.0 1103  0.002 item19
item20  0.249    1.055    1.578    1.058    1.520 1168.5 1103  0.083 item20
item21  0.387    0.932    -2.096    0.916    -1.542 1011.5 1103  0.977 item21
item22  0.196    1.068    1.995    1.039    0.619 1147.1 1103  0.173 item22
item23  0.907    0.977    -0.610    0.951    -0.559 1049.8 1103  0.872 item23
item24  0.002    0.872    -5.720    0.720    -5.072  816.2 1103  1.000 item24

```

Figure 77 - Tableau synoptique des items

On pourra exploiter cette structure plus avant dans la suite.

10.16 Défi : Matrice de corrélation des résidus

Afin de visualiser les résidus issus du fit du modèle, on souhaite les représenter sous forme d'une matrice de corrélation. Au niveau des données, il convient de disposer des valeurs des résidus et ensuite d'en établir les corrélations.

Les valeurs des résidus sont obtenues via la fonction **residuals** sur les estimations des paramètres des personnes. Ces résidus sont ensuite corrélés via la fonction **cor**.

Au niveau de la représentation graphique, on s'oriente vers le package **corrplot** qui permet de dessiner des matrices de corrélation avec la fonction **corrplot**.

Cette version présente les données en triangle (type= « upper »), avec des nombres dans les cases. Un code couleur est associé suivant le coefficient de corrélation.

```
```{r Residuals Corr Matrix}
library(corrplot)
corrplot(myPersonParametersEstimatesResidualsCorr,
 type="upper",
 #order="hclust",
 tl.col="black",
 tl.srt=45,
 method="number")
```
```

Figure 78 – Chunk relatif à la matrice de corrélation des résidus

On souhaite représenter l'échelle de manière à ce que les valeurs entre -0.3 et 0.3 soient en blanc, les valeurs entre 0.3 et 0.4 ainsi que entre -0.4 et -0.3 soient en jaune et le reste en bleu et rouge. Une palette de couleurs spécifique a ainsi été confectionnée pour l'appliquer à la matrice. La logique de cette matrice est d'être composée de 20 entrées, ce qui fait 10 sections en positif et 10 sections en négatif, ce qui correspond à la division en dixièmes de l'unité utilisée par la matrice en

```
corrplot(myPersonParametersEstimatesResidualsCorr,
  type="upper",
  # order="hclust",
  tl.col="black", #text label color
  tl.srt=45,      #text label rotation
  addCoef.col="black", # coeff de corrélation color
  method="color",
  # Combiner avec le niveau de significativité
  sig.level = 0.01, insig = "blank",
  # Cacher les coefficients de corrélation sur la diagonale
  diag=TRUE, # Montrer la diagonale
  col=paletteCouleurs0(20), #couleurs pour -0.3, 0.3
  addgrid.col="gray",
  #addrect=TRUE,
  number.cex=0.65
)
```

Figure 79 - Tracé de la matrice avec la palette spécifique

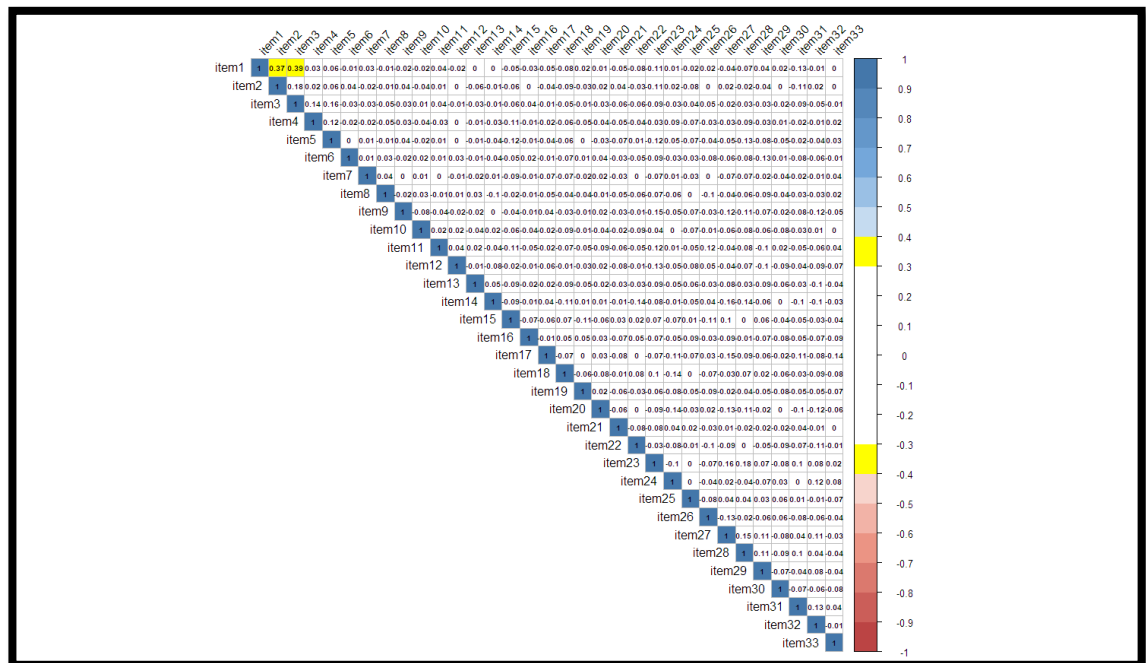


Figure 80 - Matrice de corrélation avec palette spécifique

Cette matrice colorée permet de facilement de détecter les items suspects. Une table des corrélations des résidus est sortie avec **kable**.

```
## Table Correlation Residus
{r comment='Correlation Residuals Table', echo=FALSE, results='asis'}
knitr::kable(myPersonParametersEstimatesResidualsCorr, caption = "Correlation Residus", floating.environment="sidewaystable")
# Tables Infos
```

Figure 81 - Table de corrélation des résidus

10.17 Défi : Targeting

10.17.1 Targeting des personnes

L'établissement du targeting des personnes vise à représenter le compte de personnes par rapport à un niveau de trait latent. Ceci peut s'obtenir via un histogramme sur les données de `myPersonTable$location`.

```

####{r Targeting Person}

peh<-hist(myPersonTable$Location,
          xlim=c(myLTmin,myLTmax),
          breaks=seq(myLTmin,myLTmax,0.2),
          col="green",
          angle=45,
          border="blue",
          axes=FALSE,
          main="Targeting persons",
          xlab=myLatenTrait,
          ylab="Count")

axis(side = 1, at = seq(myLTmin,myLTmax, 1.0))
axis(side = 2, at = seq(0,max(peh$counts),1))
box()

```

Figure 82 -

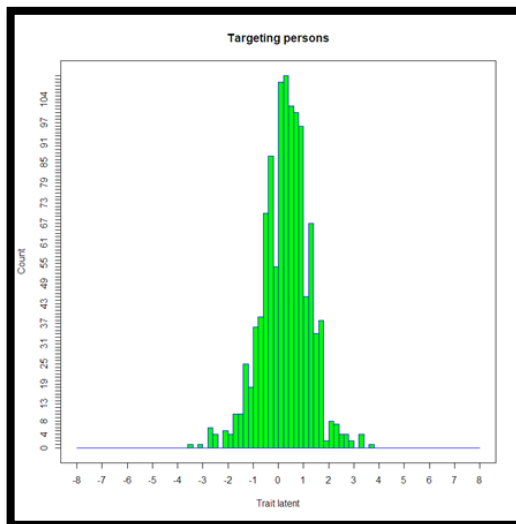


Figure 83 -

10.17.2 Targeting des items

Le principe est le même pour les items.

A noter, `as.vector(sapply(myItemThresholdsForHist, c))` qui vise à aplanir toutes les données du dataframe des seuils.

```

## Targeting Items Thresholds

####{r Targeting Items Thresholds}
myItemThresholdsForHist <- myItemThresholdsTable[, ! names(myItemThresholdsTable) %in%
c("Location"), drop = F]

ith<-hist(
  as.vector(sapply(myItemThresholdsForHist, c)), # flatten dataframe
  xlim=c(myLTmin,myLTmax),
  breaks=seq(myLTmin,myLTmax,0.2),
  col="green",
  angle=45,
  border="blue",
  axes=FALSE,
  main="Targeting itemsThresholds",
  xlab=myLatenTrait,
  ylab="Count"
)

axis(side = 1, at = seq(myLTmin,myLTmax, 1.0))
axis(side = 2, at = seq(0,max(ith$counts),1))
box()

```

Figure 84 -

Ceci étant, ce n'est pas comparable avec les sorties RUMM2030 qui présentent des graphiques combinés. Le package **eRm** ne propose pas cette visualisation, il a donc été nécessaire de créer cette disposition superposée sur base d'une composition de graphiques.

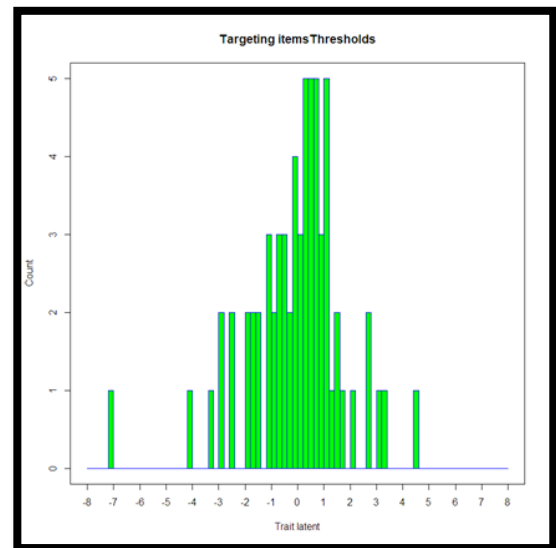


Figure 85 -

10.17.3 Person Item Threshold Distribution

Cette visualisation composée s'obtient au travers d'une fonction développée spécifiquement à cette fin : **caPlotPersonItemHist**.

```

```{r PersonItemHist Threshold Distribution}
personLocations <- myPersonTable$Location

Notez le !names...
myItemThresholdsForHist <- myItemThresholdsTable[, !names(myItemThresholdsTable) %in%
c("Location"), drop = F]

itemThresholdLocations <- as.vector(sapply(myItemThresholdsForHist, c))

caPlotPersonItemHist(personLocations,
 itemThresholdLocations,
 myLTmin, myLTmax, 0.2,
 myLatentTrait,
 "Person Locations",
 "Item Thresholds",
 TRUE,
 FALSE,
 "Person-Item Threshold Distribution"
)
```

```

Figure 86 -

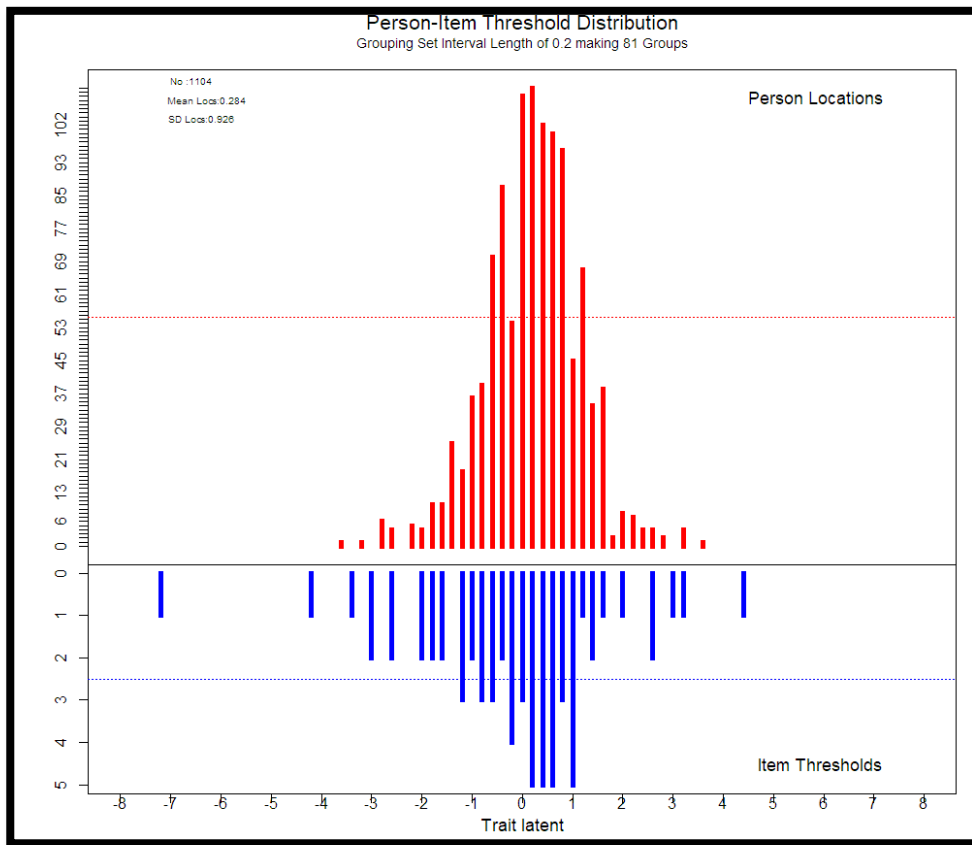


Figure 87 -

10.17.4 Person-Item Location Distribution

Cette visualisation composée s'obtient également au travers de `caPlotPersonItemHist` mais en passant les item locations.

```
##{r PersonItemHist Location Distribution}

personLocations <- myPersonTable$Location

myItemLocationsForHist <- myItemThresholdsTable[, names(myItemThresholdsTable)
%in% c("Location"), drop = F]

itemLocations <- as.vector(sapply(myItemLocationsForHist, c))

caPlotPersonItemHist(personLocations,
  itemLocations,
  myLTmin, myLTmax, 0.2,
  myLatenTrait,
  "Person Locations",
  "Item Locations",
  TRUE,
  FALSE,
  "Person-Item Location Distribution"
)
```

Figure 88 -

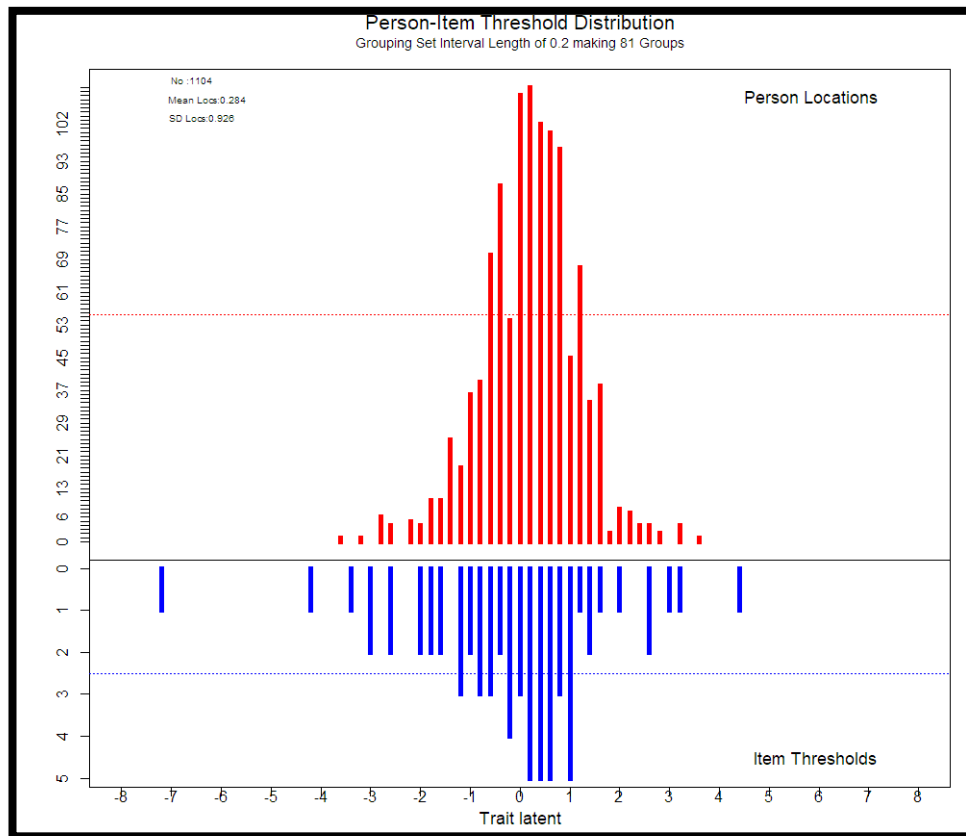


Figure 89 -

Cette fonction se trouve dans le fichier **caPlotPersonItemHist.R**

Le principe est d'établir un layout des figures de 2 lignes sur 1 colonne ainsi que d'avoir un espace 2 fois plus grand en haut qu'en bas. Le dessin dans le layout commence par le bas. On commence par la partie concernant les items. Ensuite, l'histogramme est calculé avec **hist** sans le dessiner (« plot=FALSE »), afin de disposer des valeurs des différents « bins ».

Sur base d'une zone de dessin de départ vide, on va dessiner les axes et libellés des items.

Dans un second temps, les points sont dessinés (avec l'option permettant de les dessiner comme des bâtons : type= « h », lwd=5 (line width)). Ils doivent être dessinés de manière à ce qu'ils aient la tête en bas, ce qui implique quelques calculs d'ajustement.

Une option existe pour afficher les statistiques des éléments. Si cette option est activée, la fonction dessine les statistiques (mean, sd, ...). Des lignes horizontales sont ajoutées dont une pour l'axe des logits. Une fois les items tracés, la même méthode est utilisée pour les personnes. On termine par donner un titre à la visualisation, rappelant son paramétrage.

10.18 Défi : Fonction d'information

Le dessin de la fonction d'information s'obtient via la fonction `plotINFO` de `eRm`.

```
```{r Information function}
TODO faire pareil pour les divers seuils

plotINFO(myModelParameterEstimates,
 type = "item",
 theta = seq(myLTmin, myLTmax,
 length.out = 1001),
 xlab=myLatenTrait)
```
```

Figure 90-

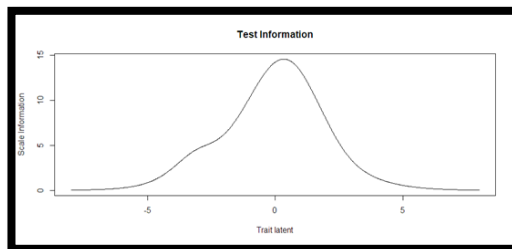


Figure 91-

10.19 Défi : Person et Item map (développement spécifique)

On peut sortir une carte des personnes et des items avec les seuils non triés.

Cette visualisation est réalisée au travers d'une routine développée spécifiquement sur base de la routine `plotPImap` de `eRm` car il manquait des éléments (exemple libellé du trait latent, mise en couleur, autre forme des points...)

```
```{r PlotPIMapUnsorted}

irug = afficher les seuils en barrettes
warn.ord.color Warning ordre seuils en PCM
sorted = seuils ordonnés et items
warn.ord asterisques pour seuils non ordonnés

caPlotPImap(myModelParameterEstimates,
 xrange = c(myLTmin, myLTmax),
 latdim = myLatenTrait,
 irug=TRUE,
 pp=myPersonParametersEstimates,
 warn.ord.colour="red",
 sorted=FALSE,
 warn.ord=TRUE)
```
```

Figure 92-

Une option permet de sortir les seuils triés.

```
```{r PlotPIMapSorted}

caPlotPImap(myModelParameterEstimates,
 xrange = c(myLTmin, myLTmax),
 latdim = myLatenTrait,
 irug=TRUE,
 pp=myPersonParametersEstimates,
 warn.ord.colour="red",
 sorted=TRUE,
 warn.ord=TRUE)
```
```

Figure 932-

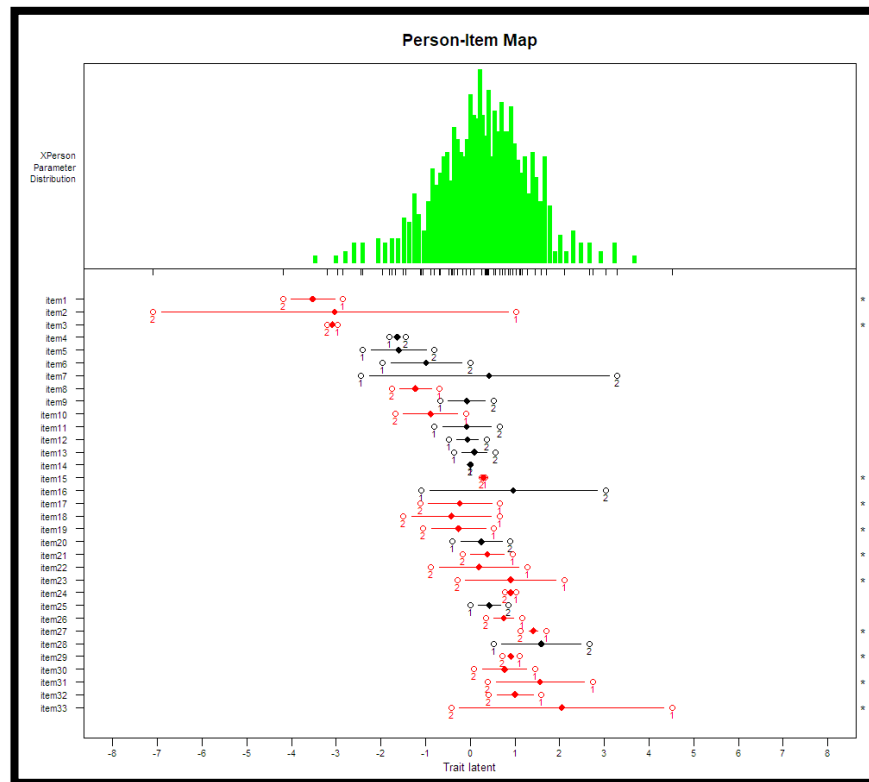


Figure 94-

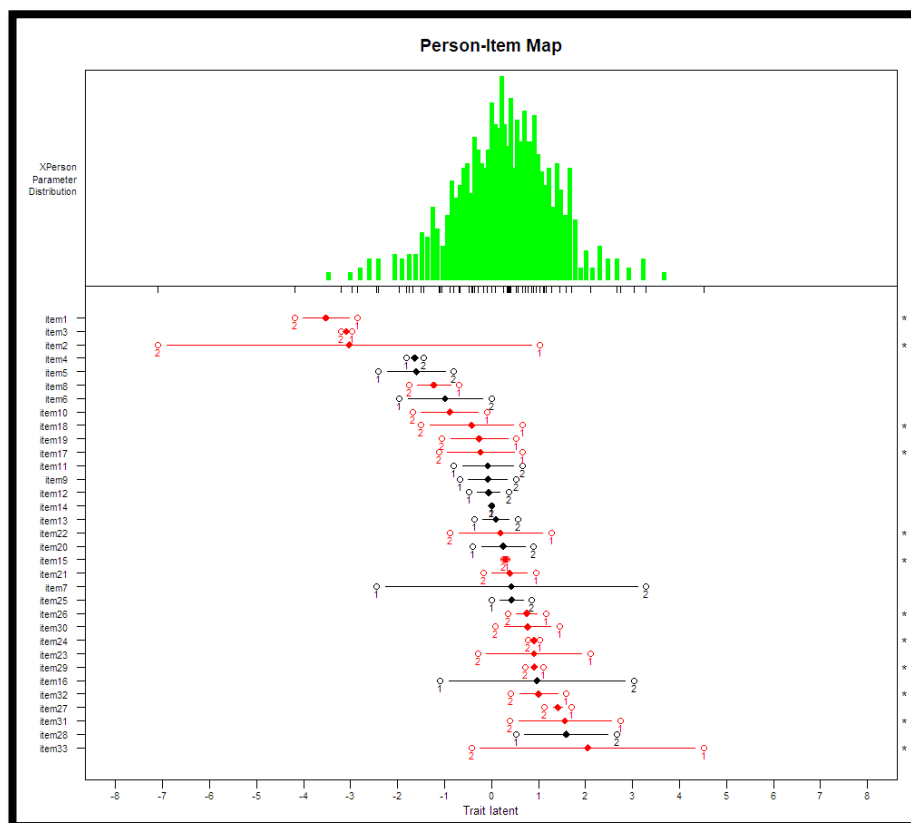


Figure 95-

10.20 Défi : Plot ICC

En ce qui concerne les visualisations des courbes ICC en polytomique, le support de **eRm** n'est pas au point. Il a donc été nécessaire de réaliser un développement spécifique qui a demandé des recherches assez conséquentes afin de parvenir à, et pour chaque item :

- Tracer la visualisation ICC de l'item, avec :
 - Les courbes des catégories
 - Des couleurs correctes pour celles-ci
 - Le tracé des seuils au niveau des intersections des courbes
 - Le légendage des éléments du graphique

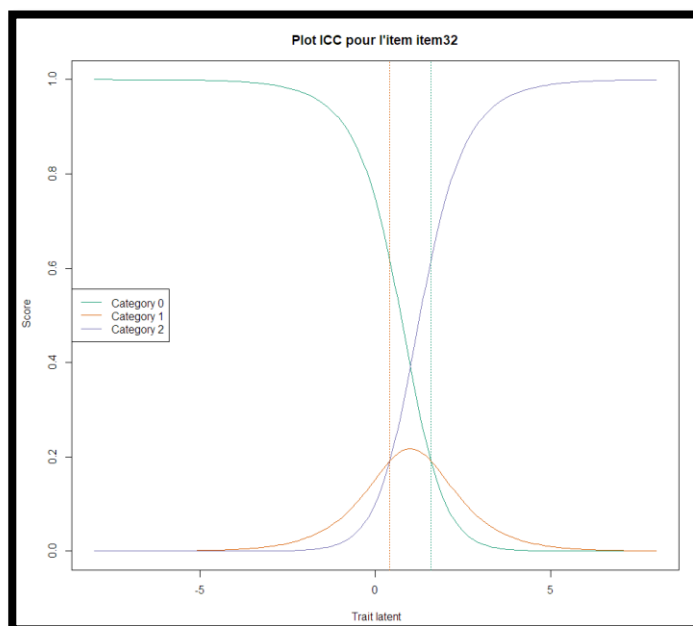


Figure 96-

- Tracer la visualisation TCC de chaque item, avec :
 - La courbe dessinée sur base de la formule permettant d'assembler les courbes de l'item
 - Le tracé de la location de l'item
 - L'affichage des statistiques de l'item, avec un formatage correct des décimales

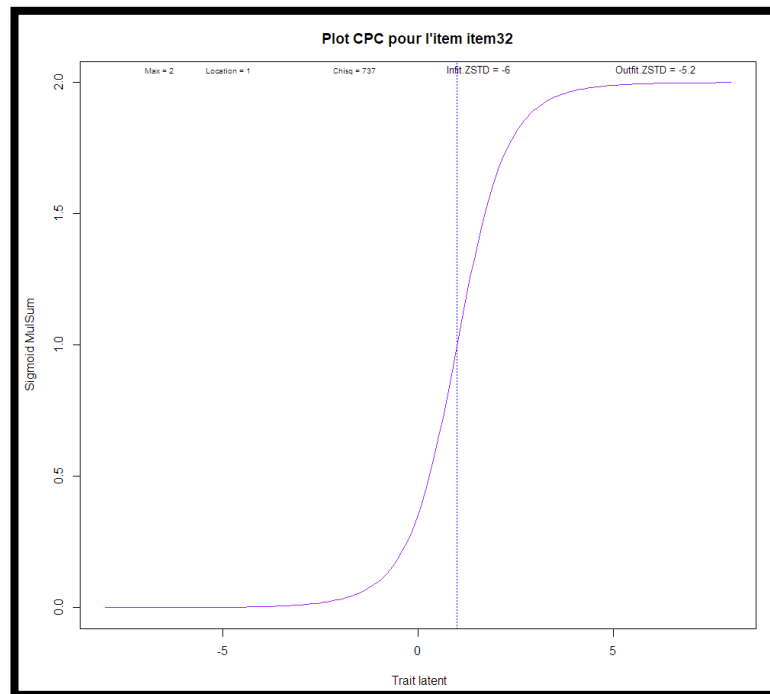


Figure 97 -

Afin de pouvoir suivre le déroulement de cette routine et son extension (on part de la routine **plotICC** de **eRm**), on utilise **log4r** qui permet d’afficher des informations sur son déroulement dans la console. Les appels de **log4r** sont toujours présents dans le code mais désactivés. La routine calcule la somme des courbes au sein d’une variable **sigmoidData**.

10.21 Défi : Visualisation du TCC Global

La visualisation du TCC global en polytomique peut s’effectuer avec un plot spécifique. L’utilisation de la routine de **eRm** ne fonctionne pas en polytomique.

```

```{r eRm| TCC Plot}
par(mfcol=c(1,1))

plot(x = myPersonParametersEstimates$pred.list$`1`$y,
 y = myPersonParametersEstimates$pred.list$`1`$x,
 xlab = myLatenTrait,
 ylab = myICCLabel,
 type = "o",
 col="purple",
 xlim = c(myLTmin,myLTmax),
 ylim = c(
 min(myPersonParametersEstimates$pred.list$`1`$x), # yes, x
 max(myPersonParametersEstimates$pred.list$`1`$x)),
 main="TCC")
```

```

Figure 98-

On obtient une courbe du type suivant :

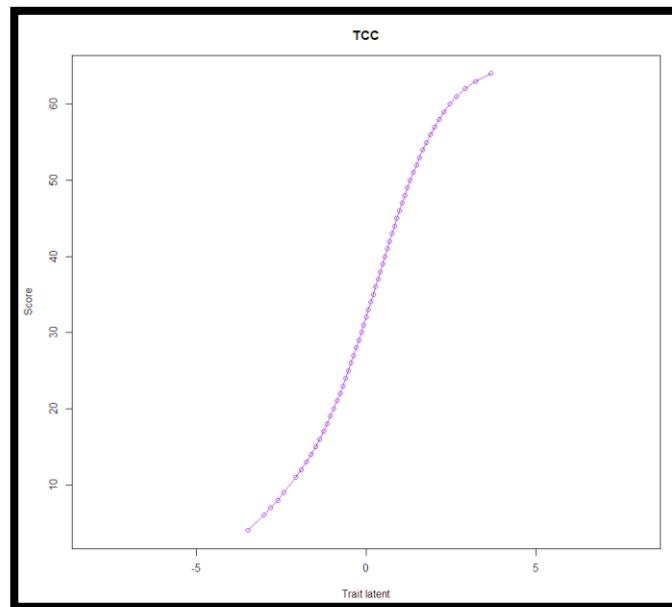


Figure 99-

10.22 Défi : Visualisation Plot Bond-and-Fox pathway map (items, personnes)

On peut sortir ces deux visualisations (items et personnes) afin de voir graphiquement les items et personnes hors cadre.

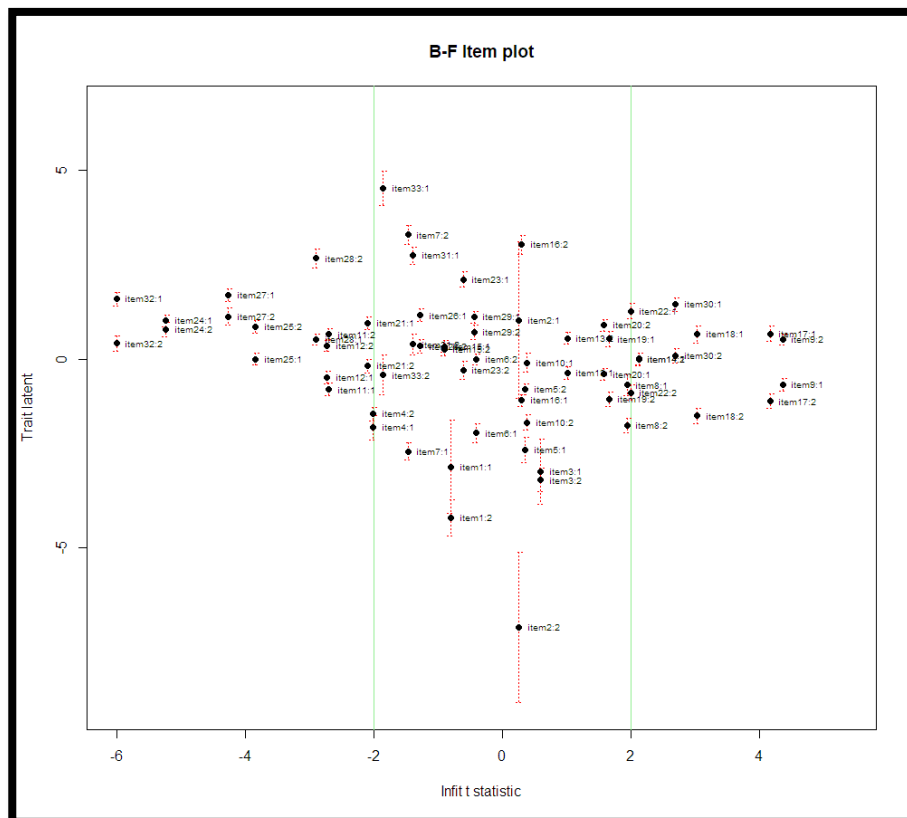


Figure 100-

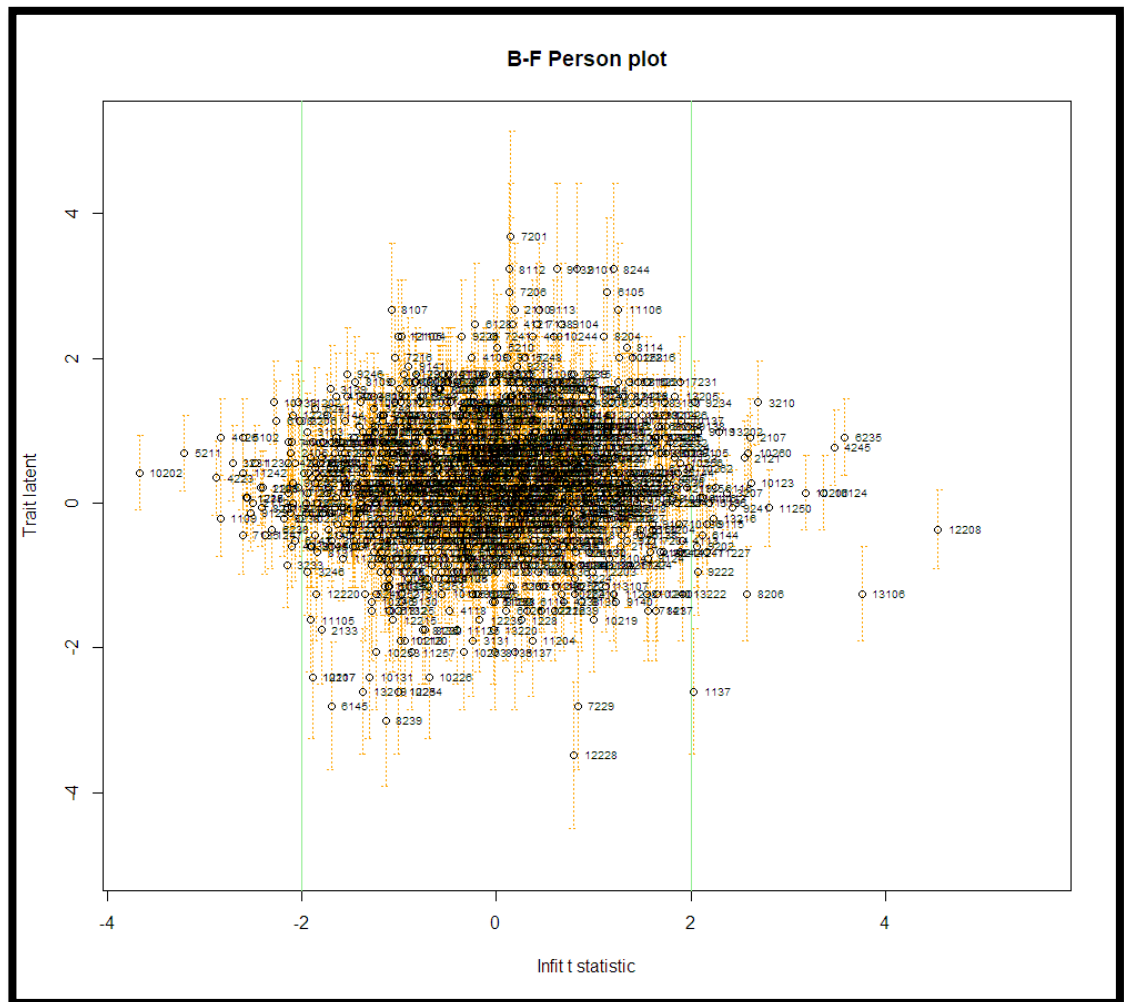


Figure 101-

10.23 Défi : Calcul des métriques PSI, SSD, MSE, G/Separation Ratio,

Le calcul de ces éléments est disponible dans **eRm** via la fonction **SepRel**.

```
## Metrics
### PSI
`r myPersonSepRel$sep.rel `

### SSD
`r myPersonSepRel$SSD.PS `

### MSE
`r myPersonSepRel$MSE`

### G / Separation Ratio
Value `r SepRatio`

Alternative Value `r SepRatioBis`
```

Figure 102-

10.24 Défi: Clustering en categories

Le groupement des personnes en un nombre donné de catégories en termes de niveau de trait latent est demandé. Afin de rencontrer ce besoin, les possibilités de « clustering 1D » ont été examinées. Il s'avère que le groupement par catégories au travers d'histogrammes ne fonctionne pas car les « buckets » doivent être prédéfinis.

L'approche **K-means** a alors été envisagée car elle permet de grouper les éléments en K categories de manière dynamique. La fonction **kmeans** de R est une option mais la solution fournie par l'algorithme n'est pas nécessairement toujours la même, ce qui est un souci.

Des options plus intéressantes ont été alors considérées pour un clustering k-means 1D, dont le package **Ckmeans.1d.dp** qui permet une meilleure reproductibilité des résultats, en sus de proposer plus de possibilités au niveau des informations sur les groupes.

Les métriques qui nous intéressent sont la moyenne (et si besoin, la médiane, non implémentée mais il suffit d'utiliser le partitionnement par rapport à la médiane si besoin).

On peut ainsi grouper les personnes en k groupes (ici k=5).

```
myCategoriesCount <- 5
k <- myCategoriesCount
myCategoriesResult <- Ckmeans.1d.dp(myPersonAbilityEstimates, k)
```

Figure 103 - Clustering

On peut obtenir alors un partitionnement en 5 groupes.

La routine permet éventuellement de donner un poids aux éléments, ceci ne nous est pas utile et chaque élément se voit assigner un poids de 1. La routine donne également la position des centres des clusters (représenté par les lignes verticales noires) ainsi que le nombre d'éléments de chaque groupe.

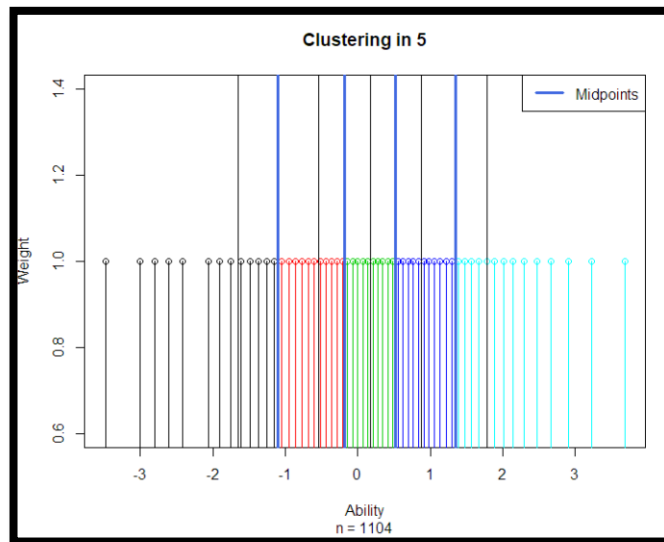


Figure 104 – Partitionnement

On peut disposer des bornes des groupes à l'aide de la fonction **ahist** du package (midpoints sur la figure).

Afin de pouvoir partitionner le jeu de données sur base de ce clustering, nous avons accès à un vecteur contenant le numéro de la catégorie. Ce vecteur ayant la longueur de la sélection de personnes que nous traitons. On peut donc l'utiliser comme critère DIF via l'option **splitter** du DIF.

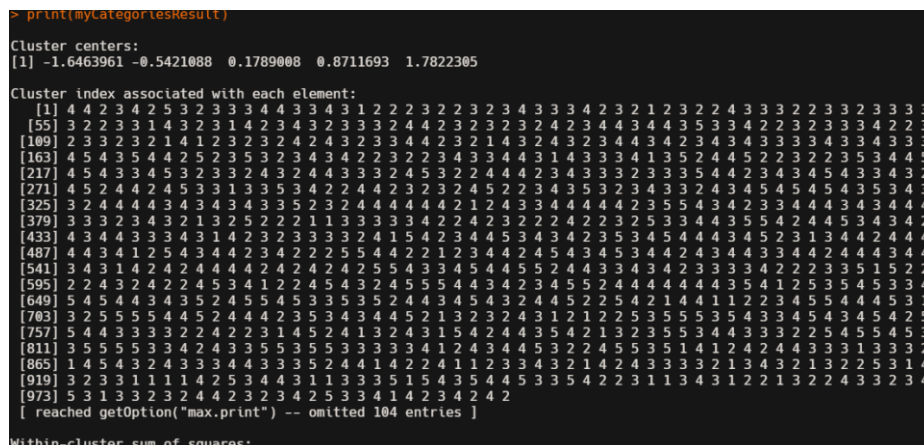


Figure 105 - Centres des groupes et vecteur associant un groupe à chaque personne

10.25 Défi : Visualisations DIF

En ce qui concerne le DIF, on peut configurer deux facteurs DIF dans le fichier de paramètres, en spécifiant l'endroit où ils se trouvent dans le jeu de données, leurs niveaux muni d'un libellé.

```

# facteurs de fonctionnement différentiel
myDIF1todo <- TRUE
# FALSE si le test du FDI-1 ne doit pas e
myDIF1Column <- 35

myDIF1InfoColumnName <- "Sex"
myDIF1factor <- "Sex"
myDIF1levelA <- 1
myDIF1labelA <- "Hommes"
myDIF1levelB <- 2
myDIF1labelB <- "Femmes"

# facteurs de fonctionnement différentiel
myDIF2todo <- TRUE
# FALSE si le test du FDI-2 ne doit pas e

myDIF2InfoColumnName <- "Categorie Age"
myDIF2Column <- 37
myDIF2factor <- "ageCat"
myDIF2levelA <- 0
myDIF2labelA <- "Les Jeunes"
myDIF2levelB <- 1
myDIF2labelB <- "Les Vieux"

```

Figure 106-

En ce qui concerne la catégorie d'âge, on réalise la transformation de l'âge en catégorie directement dans le fichier de données. Comme il est au format Excel, il est simple d'utiliser un IF pour ce faire. Ceci permet de ne pas devoir obliger l'utilisateur à aller écrire un code de transformation de ses données en R.

Pour chaque facteur DIF (1 et 2), un LRtest est effectué afin de disposer d'un résultat permettant d'effectuer la partition de la population des personnes.

```

```{r Calcul du DIF}

test for DIF factor 1
if (myDIF1todo == TRUE) {
 myDIF1test <- LRtest(myModelParameterEstimates, splitcr = myData[,myDIF1Column])
}

test for DIF factor 2
if (myDIF2todo == TRUE) {
 myDIF2test <- LRtest(myModelParameterEstimates, splitcr = myData[,myDIF2Column])
}

...

```

Figure 107-

Le résultat de ces LRtest est ensuite utilisé dans le cadre de la fonction **plotDIF** de **eRm**.

```
plotDIF(myDIF1test,
 main = paste("DIF selon",myDIF1factor),
 xlim = c(myLTmin,myLTmax),
 xlab = myLatenTrait,
 leg = TRUE, legpos = "topright",
 distance = .2, # inter line distance
 # item.subset = c("item32","item33"),
 col = c("red", "blue","green","pink","orange","black","yellow"),
 splitnames = c(paste(
 myDIF1levelA,
 myDIF1labelA),
 paste(
 myDIF1levelB,
 myDIF1labelB)))
```

Figure 108-

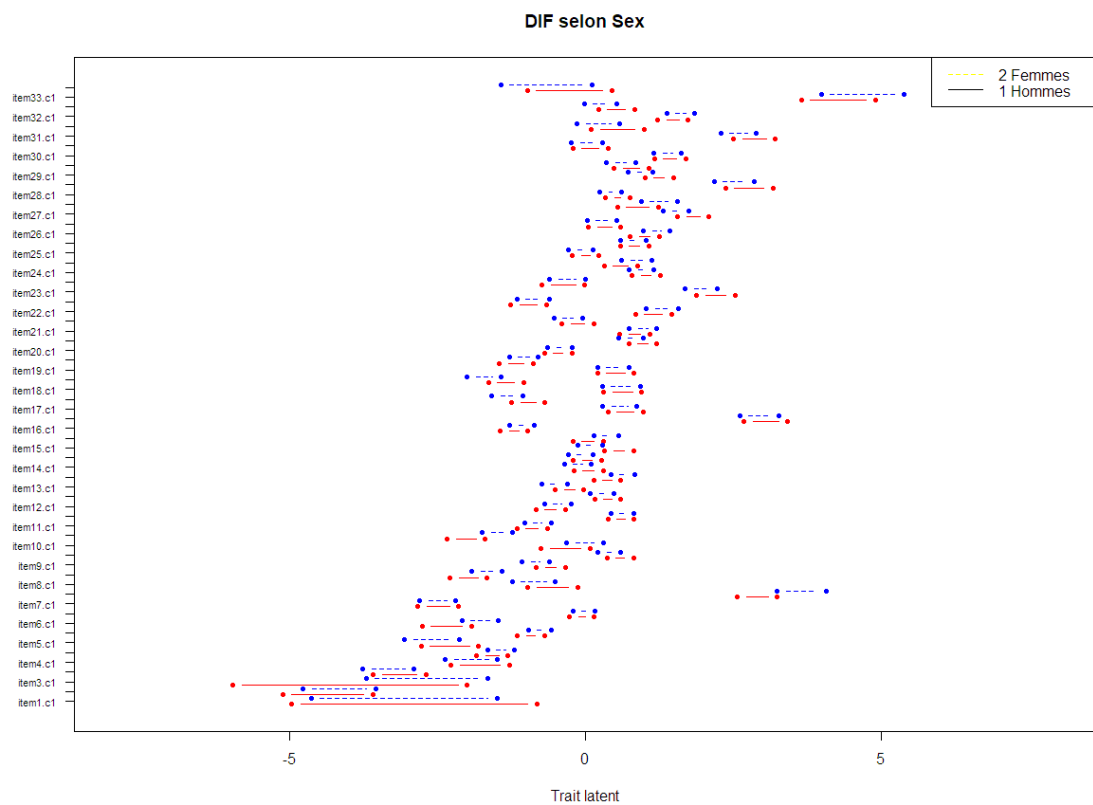


Figure 109-

```
if (myDIF2todo == TRUE) {
 plotDIF(myDIF2test,
 main = paste("FDI selon",myDIF2factor),
 xlim = c(myLTmin,myLTmax),
 xlab = myLatenTrait,
 leg = TRUE, legpos = "topright",
 distance = .2, # inter line distance
 # item.subset = c("item32","item33"),
 col = c("red", "blue","green","pink","orange","black","yellow"),
 splitnames = c(paste(myDIF2levelA,myDIF2labelA),
 paste(myDIF2levelB,myDIF2labelB)))
}
```

Figure 110

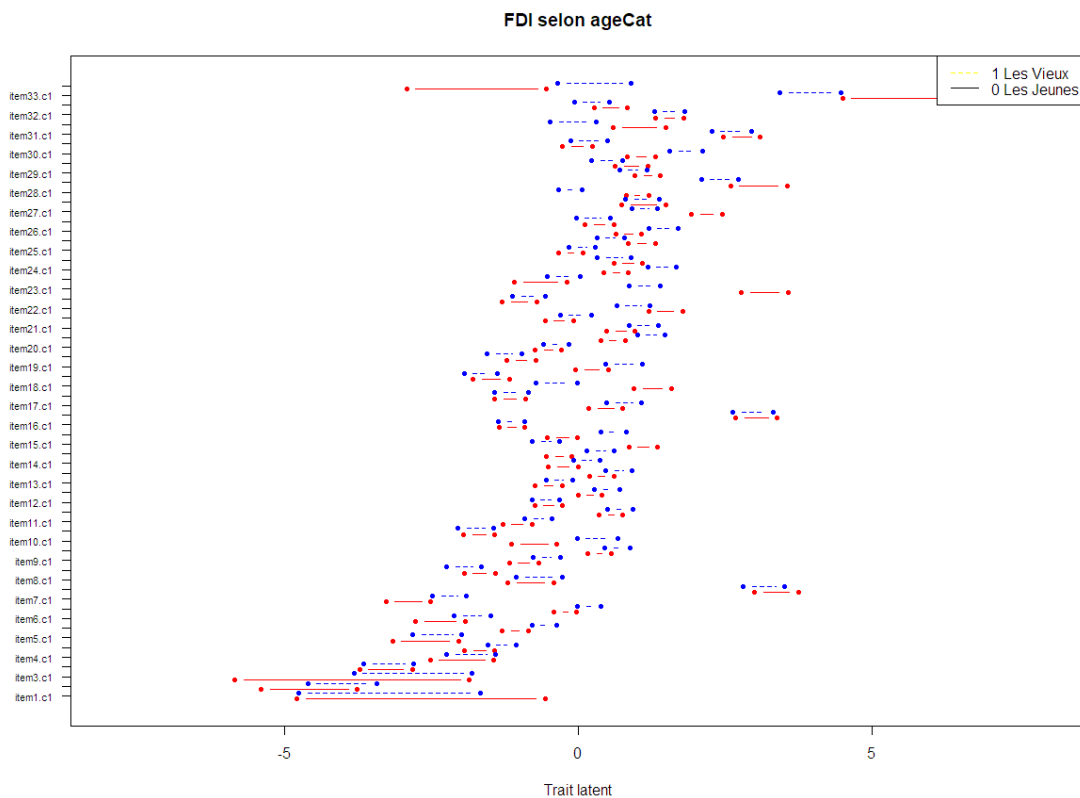


Figure 111-

En ce qui concerne le groupement par catégories, on en tire également une visualisation DIF.

```

```{r DIF Categories}

myDIFClustertest <- LRtest(myModelParameterEstimates, splitcr = myCategoriesResult$cluster)

par(mar=c(2,4,1,1))

plotDIF(myDIFClustertest,
  #Need to see how item.subset works internally as they are not all displayed.
  #item.subset = names(myModelParameterEstimates$etapar),
  main = paste("DIF Cluster"),
  xlim = c(myLTmin,myLTmax),
  xlab = myLatenTrait,
  leg = TRUE, legpos = "topright",
  distance = .2, # inter line distance
  # item.subset = c("item32","item33"),
  col = c("red", "blue", "green", "black", "cyan"),
  splitnames = c("C1", "C2", "C3", "C4", "C5"), #TODO make this automatically
  c(1,2,3,4,5)) #use the range with 1,k

```

Figure 112-

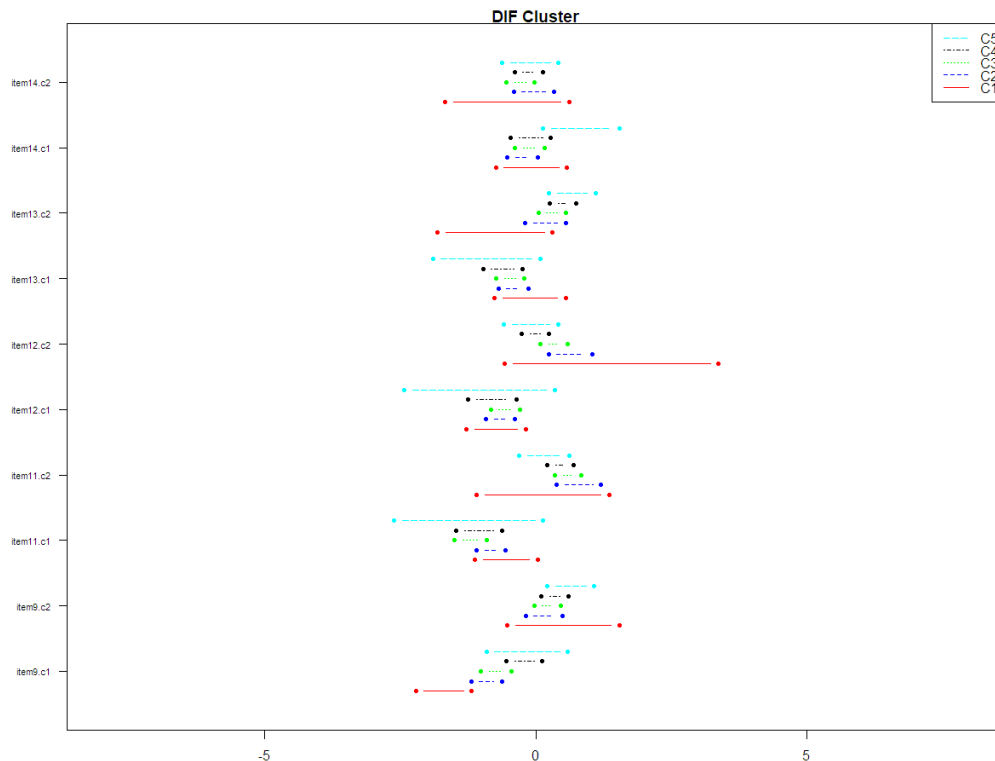


Figure 113-

10.26 Défi : Résumé des données

On souhaite ici tirer des statistiques classiques sur le jeu de données (moyenne, quartiles...). On utilise à cette fin la fonction `summary` de R sur le jeu de données et on affiche le résultat au sein d'une table knitr:kable.

10.27 Défi : Préparation et présentation des données sous forme de tableaux triés par critères

10.27.1 Préparation des données

On utilise le package **gtools** pour exploiter la fonction **mixedorder** qui permet de trier de manière assez flexible («¹⁵Order or Sort strings with embedded numbers so that the numbers are in the correct order»), surtout si les personnes et items sont nommés `person1`, `person2`, ..., `person11`, on ne veut pas avoir `person11` juste après `person1`, mais bien `person10`. Une série d'index est alors préparée permettant de présenter les données ordonnées suivant l'un ou l'autre de ces index, ces index étant appliqués pour produire de jeux de données ordonnés pour une utilisation ultérieure.

¹⁵(<http://cran.freestatistics.org/web/packages/gtools/gtools.pdf>, aide package `gtools`, p27)

```
myRealDataItByLabel<-myRealDataIt[mixedorder(myRealDataIt$Label),]
myRealDataItByLabelDesc<-myRealDataIt[mixedorder(myRealDataIt$Label, decreasing = TRUE),]

myRealDataItByLocation <- myRealDataIt[order(myRealDataIt$Location),]
myRealDataItByLocationDesc <- myRealDataIt[order(myRealDataIt$Location, decreasing = TRUE),]

myRealDataItByChisq <- myRealDataIt[order(myRealDataIt$Chisq),]
myRealDataItByChisqDesc <- myRealDataIt[order(myRealDataIt$Chisq, decreasing = TRUE),]
```

Figure 114 -

10.28 Présentation sous forme de tableaux

On utilise alors les données préparées pour les afficher sous forme de tableau avec `knitr::kable`

```
## Données de départ pour les Items
### Items Ordered by Label increasing
```{r comment='', echo=FALSE, results='asis'}
knitr::kable(myRealDataItByLabel,
 caption = "myRealDataIt ordered by Label increasing",
 floating.environment="sidewaystable")
```

### Items Ordered by Location increasing
```{r comment='', echo=FALSE, results='asis'}
knitr::kable(myRealDataItByLocation,
 caption = "myRealDataIt ordered by (computed) Location increasing",
 floating.environment="sidewaystable")
```
```

Figure 115-

Le procédé est similaire pour les personnes.

10.29 Défi : ACP des résidus

`myPersonEstimatesResiduals` contient les résidus du fit avec `person.parameters`

| | item1 | item2 | item3 | item4 | item5 | item6 | item7 | item8 | item9 | item10 | item11 | item12 |
|------|------------|-------------|-------------|-------------|------------|-------------|--------------|-------------|--------------|-------------|-------------|------------|
| 1101 | 0.09025290 | 0.031941790 | 0.14785677 | -2.39069679 | 0.4908550 | 0.72672221 | -0.064167983 | 0.32266538 | 0.961796120 | 0.36099922 | 1.01166161 | 0.9086331 |
| 1102 | 0.07247901 | 0.022591356 | 0.11877283 | -3.14861412 | 0.3947545 | 0.58621347 | -0.199963503 | 0.25220241 | -0.872018996 | 0.27458479 | 0.81469844 | 0.7193011 |
| 1103 | 0.16825847 | 0.094043009 | 0.27459128 | -0.87929177 | 0.8977037 | -0.25197750 | 0.313584225 | -0.57442585 | -1.054039171 | -1.40381440 | 1.84267442 | 0.4121054 |
| 1104 | 0.11925674 | 0.051158468 | 0.19519925 | 0.47536530 | -1.2723573 | -0.70977695 | 0.111238842 | 0.45172469 | 1.279830733 | 0.52760039 | -0.10340829 | 1.2329815 |
| 1105 | 0.08407341 | 0.028480853 | 0.13774958 | 0.33211923 | 0.4575179 | 0.67819222 | -0.108710689 | 0.29748954 | -0.628139855 | 0.32964850 | 0.94342057 | 0.8419461 |
| 1106 | 0.13787742 | 0.065993330 | 0.22547492 | 0.55326412 | 0.7432189 | -0.51053567 | 0.198938055 | 0.54449175 | 1.485102799 | -0.64069561 | 0.09589078 | -1.2171786 |
| 1107 | 0.06149896 | 0.017730195 | 0.10079018 | 0.24222135 | 0.3351274 | 0.49825451 | -0.297897377 | 0.21136131 | 0.651451107 | 0.22663186 | 0.69210598 | -3.0451985 |
| 1108 | 0.11117810 | 0.045319535 | 0.18203186 | 0.44204589 | 0.6029123 | -0.81043774 | 0.067612075 | 0.41390080 | -0.207494344 | 0.47796355 | 1.24051189 | -0.2034846 |
| 1109 | 0.13787742 | 0.065993330 | 0.22547492 | 0.55326412 | 0.7432189 | 1.08356681 | 0.198938055 | 0.54449175 | 0.099116780 | 0.65118113 | 0.09589078 | 0.1158624 |
| 1110 | 0.11925674 | 0.051158468 | 0.19519925 | 0.47536530 | 0.6457345 | -0.70977695 | 0.111238842 | -1.34542619 | 1.279830733 | -0.95960923 | 1.32783042 | -0.0985406 |
| 1111 | 0.11117810 | 0.045319535 | 0.18203186 | 0.44204589 | 0.6029123 | 0.88772244 | 0.067612075 | -1.53945601 | 1.190899153 | -1.13201133 | -0.20084935 | 1.1409316 |
| 1112 | 0.13285599 | 0.061808018 | 0.21732176 | 0.53210303 | 0.7171148 | 1.04766773 | 0.176871592 | 0.51871650 | 0.046644866 | 0.61663111 | 0.04509633 | 0.0611725 |
| 1113 | 0.08712458 | 0.030161666 | 0.14274076 | 0.34436772 | 0.4739896 | 0.70220214 | -0.086381135 | 0.30982510 | 0.921758465 | 0.34494957 | -0.55158036 | 0.8747587 |
| 1114 | 0.07530139 | 0.023952753 | 0.12339359 | 0.29704933 | 0.4100547 | -1.44676389 | -0.176754939 | 0.26300640 | -0.807446731 | -2.39174601 | 0.84610828 | 0.7488607 |
| 1115 | 0.11117810 | 0.045319535 | -5.08962755 | 0.44204589 | -1.4023352 | -0.81043774 | 0.067612075 | 0.41390080 | -0.207494344 | 0.47796355 | -0.20084935 | -1.5479010 |

Figure 116 -

On souhaite les analyser afin de détecter s'il n'existe pas une dimension qui échappe à l'analyse Rasch. On va ainsi utiliser une analyse en composantes principales sur les résidus. En R, on va utiliser la fonction `pca` du package `mdatools`. Cette analyse permet

de disposer d'un modèle en composantes principales dont on peut extraire les valeurs propres (Eigenvals).

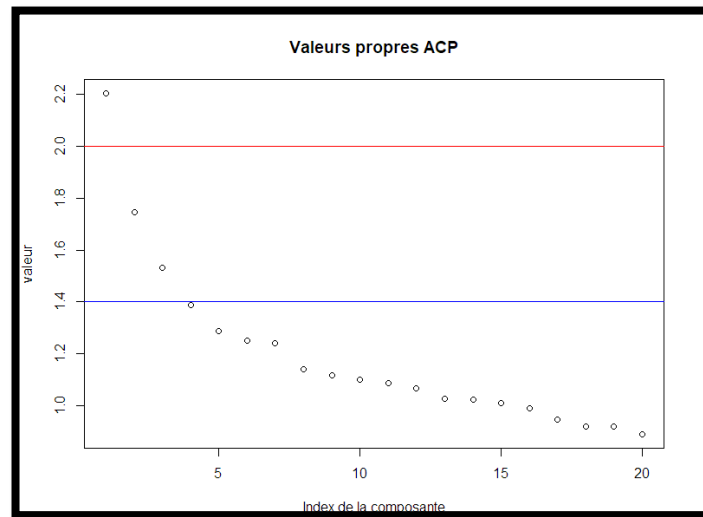


Figure 117 - Composantes Principales

Au niveau de Rasch, on considère qu'une composante principale en dessous de 1.4 correspond à de l'aléatoire. On isole alors les composantes au dessus de 1.4.

S'il en existe, celles au dessus d'un certain seuil (mettons 2.0) indiquent qu'il faut creuser plus avant car il y a une information cachée dans les résidus et donc un aspect du trait latent qui n'est pas expliqué par le modèle de Rasch actuel. C'est le cas dans la figure présentée.

10.30 Versions nécessaires

1.1.1. Packages

| | Package | Version |
|----|---------------|----------|
| 1 | askpass | 1.1 |
| 2 | assertthat | 0.2.1 |
| 3 | base64enc | 0.1-3 |
| 4 | BH | 1.72.0-3 |
| 5 | bibtex | 0.4.2.2 |
| 6 | bitops | 1.0-6 |
| 7 | CDM | 7.5-15 |
| 8 | cellranger | 1.1.0 |
| 9 | Ckmeans.1d.dp | 4.3.2 |
| 10 | cli | 2.0.2 |
| 11 | colorspace | 1.4-1 |
| 12 | corrplot | 0.84 |
| 13 | crayon | 1.3.4 |
| 14 | curl | 4.3 |
| 15 | data.table | 1.12.8 |
| 16 | digest | 0.6.25 |
| 17 | doParallel | 1.0.15 |
| 18 | dplyr | 0.8.5 |
| 19 | ellipsis | 0.3.0 |
| 20 | eRm | 1.0-0 |
| 21 | evaluate | 0.14 |
| 22 | fansi | 0.4.1 |
| 23 | foreach | 1.5.0 |
| 24 | gbRd | 0.4-11 |
| 25 | generics | 0.0.2 |
| 26 | glue | 1.3.1 |
| 27 | gtools | 3.8.2 |
| 28 | highr | 0.8 |

| | | |
|----|----------------------|-------------|
| 29 | hms | 0.5.3 |
| 30 | htmltools | 0.4.0 |
| 31 | iterators | 1.0.12 |
| 32 | jsonlite | 1.6.1 |
| 33 | knitr | 1.28 |
| 34 | lifecycle | 0.2.0 |
| 35 | log4r | 0.3.2 |
| 36 | lubridate | 1.7.8 |
| 37 | magrittr | 1.5 |
| 38 | markdown | 1.1 |
| 39 | matrixStats | 0.56.0 |
| 40 | mdatools | 0.10.3 |
| 41 | mime | 0.9 |
| 42 | mnormt | 1.5-6 |
| 43 | nlstools | 1.0-2 |
| 44 | openssl | 1.4.1 |
| 45 | packrat | 0.5.0 |
| 46 | pillar | 1.4.3 |
| 47 | pkgconfig | 2.0.3 |
| 48 | plogr | 0.2.0 |
| 49 | plyr | 1.8.6 |
| 50 | prettyunits | 1.1.1 |
| 51 | progress | 1.2.2 |
| 52 | psych | 1.9.12.31 |
| 53 | purrr | 0.3.3 |
| 54 | R6 | 2.4.1 |
| 55 | RColorBrewer | 1.1-2 |
| 56 | Rcpp | 1.0.4 |
| 57 | RcppArmadillo | 0.9.870.2.0 |
| 58 | RCurl | 1.98-1.1 |
| 59 | Rdpack | 1.0.0 |
| 60 | readxl | 1.3.1 |

| | | |
|-----|----------------------|--------|
| 61 | rematch | 1.0.1 |
| 62 | rJava | 0.9-12 |
| 63 | rlang | 0.4.5 |
| 64 | rmarkdown | 2.1 |
| 65 | rsconnect | 0.8.16 |
| 66 | rstudioapi | 0.11 |
| 67 | scatterplot3d | 0.3-41 |
| 68 | stringi | 1.4.6 |
| 69 | stringr | 1.4.0 |
| 70 | sys | 3.3 |
| 71 | tibble | 3.0.0 |
| 72 | tidyselect | 1.0.0 |
| 73 | tinytex | 0.21 |
| 74 | utf8 | 1.1.4 |
| 75 | vctrs | 0.2.4 |
| 76 | xfun | 0.12 |
| 77 | xlsx | 0.6.3 |
| 78 | xlsxjars | 0.6.1 |
| 79 | yaml | 2.2.1 |
| 108 | translations | 3.6.3 |

1.1.2. RStudio

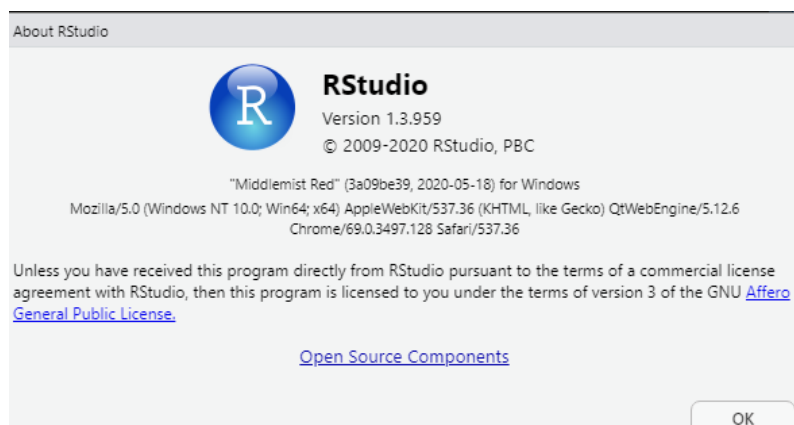


Figure 118 - Version de Rstudio utilisée pour l'élaboration du module d'analyse Rasch

1.1.3. R

```
platform      _  
arch          x86_64  
os            mingw32  
system        x86_64, mingw32  
status  
major         3  
minor         6.3  
year          2020  
month         02  
day           29  
svn rev       77875  
language      R  
version.string R version 3.6.3 (2020-02-29)  
nickname      Holding the Windsock
```

Figure 119 - Version de R utilisée pour l'élaboration du module d'analyse Rasch

11 Annexe 6 – Analyses du subtest Vocabulaire de la WAIS-III

11.1.1 Test Lof

6.5.1 MLT 4 groups

```
##  
## Martin-Loef-Test (split criterion: user-defined)  
##  
## Group 1:  
## Items: 1, 5, 10, 12, 14, 16, 19, 21, 25, 27, 28  
## Log-Likelihood: -6375.681  
##  
## Group 2:  
## Items: 2, 8, 11, 13, 17, 20, 26, 32, 33  
## Log-Likelihood: -4475.02  
##  
## Group 3:  
## Items: 3, 6, 7, 9  
## Log-Likelihood: -1112.658  
##  
## Group 4:  
## Items: 4, 15, 18, 22, 23, 24, 29, 30, 31  
## Log-Likelihood: -5033.844  
##  
## Overall Rasch-Model:  
## Log-Likelihood: -22598.94  
##  
## LR-value: 4430.31  
## Chi-square df: 74657  
## p-value: 1
```

Figure 120- Résultat au test Lof

11.1.2 Exemple d'un graphe pour analyser le DIF

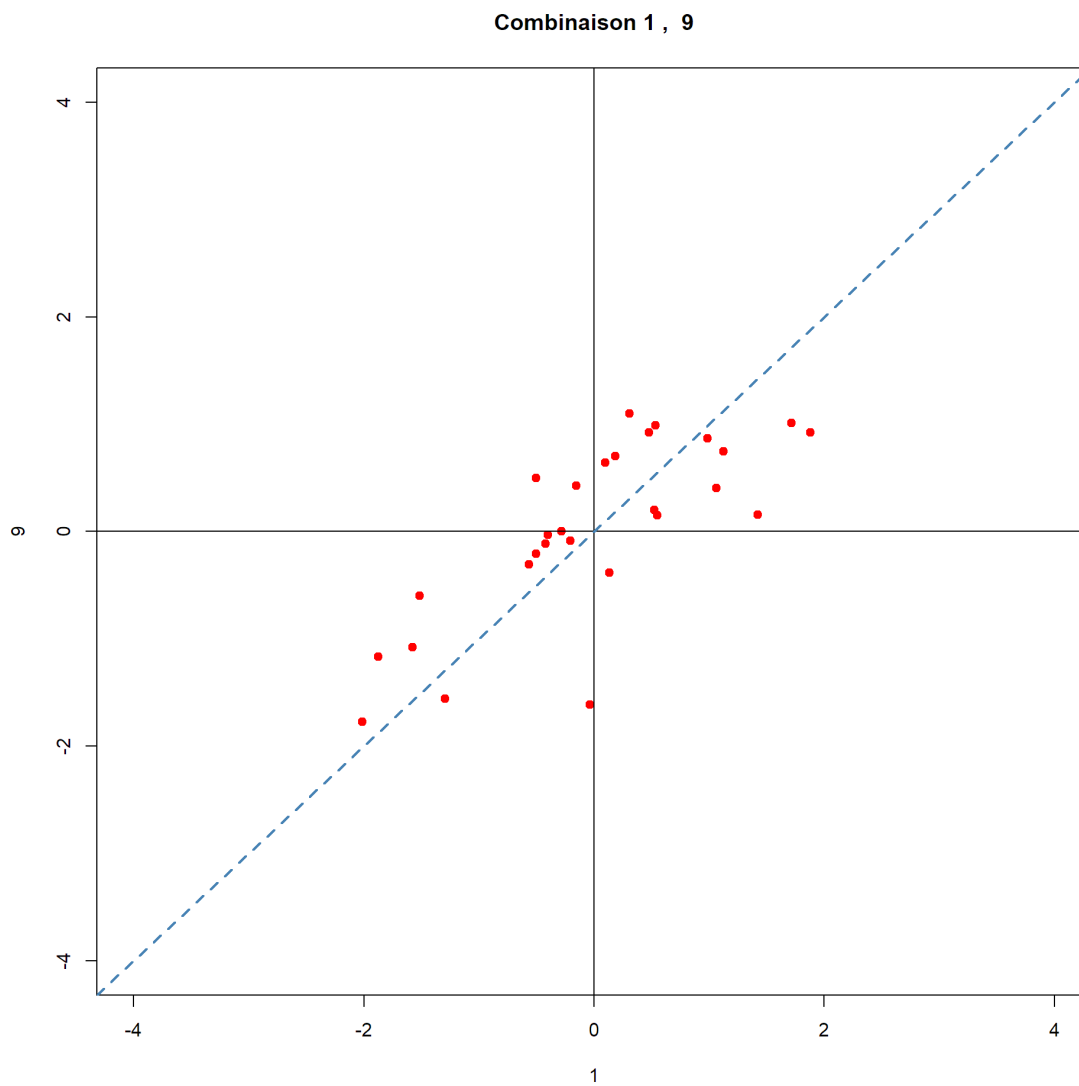


Figure 121 - FDI pour deux groupes d'âges : âgés et jeunes

Le mémoire concerne le développement d'un outil d'analyse Rasch sur base d'outils open source (R, RMarkdown). Au départ du package eRm, des programmations spécifiques pour les éléments de l'analyse Rasch non disponibles sont réalisées afin de proposer des sorties accessibles aux psychologues. Les comparaisons des résultats obtenus sont effectuées avec les résultats obtenus via le logiciel commercial RUMM2030. Il s'agit de mettre à disposition cet outil pour les psychologues confrontés au choix d'utiliser le modèle de Rasch plutôt que le modèle de la théorie classique mais qui devraient y renoncer eu égard au fait de devoir passer par des logiciels très onéreux et complexes. L'objectif est aussi de mettre en valeur l'analyse Rasch qui transforme le concept de mesure en psychologie en abordant rapidement différents contextes d'application et comparer son apport au regard de résultats en théorie classique.

The thesis concerns the development of a Rasch analysis tool based on open source tools (R, RMarkdown). At the beginning of the eRm package, specific programming for the elements of Rasch analysis not available are carried out in order to propose outputs accessible to psychologists. Comparisons of the results obtained are made with the results obtained via the commercial software RUMM2030. The aim is to make this tool available to psychologists faced with the choice of using the Rasch model rather than the classical theory model, but who would have to give it up due to the fact that they would have to use very expensive and complex software. The objective is also to highlight the Rasch analysis which transforms the concept of measurement in psychology by quickly approaching different contexts of application and comparing its contribution to results in classical theory.