Thesis

Stabilité à long terme des scores standards et CHC du WISC-IV : apports théoriques et cliniques

KIENG, Sotta

Abstract

Les indices du WISC-IV sont non seulement interprétés par les praticiens comme des indicateurs du fonctionnement cognitif actuel d'un enfant, mais sont également utilisés pour faire des prédictions à long terme. Ces prédictions reposent sur l'hypothèse que les résultats des indices sont stables dans le temps. Ainsi, la présente étude vise à examiner la stabilité à long terme des indices standard (QIT, ICV, IRP, IMT, IVT, IAG et ICC) et des indices CHC (Gc, Gf, Gv, Gwm, Gs)). Une procédure Test–Retest, avec un intervalle de un à trois ans entre les deux passations est appliquée. La stabilité des différents scores du WISC-IV est évaluée sous l'angle de la stabilité absolue (différence de moyenne), de la stabilité différentielle (coefficient de stabilité), de la stabilité intra-individuelle (différence de performance), de la stabilité catégorielle et de la stabilité des forces et faiblesses. L'échantillon comporte 277 enfants tout-venant de 7 à 12 ans.

Reference

KIENG, Sotta. *Stabilité à long terme des scores standards et CHC du WISC-IV : apports théoriques et cliniques*. Thèse de doctorat : Univ. Genève, 2017, no. Sc. 674

DOI : 10.13097/archive-ouverte/unige:96835
URN : urn:nbn:ch:unige-968351

UNIVERSITÉ
DE GENÈVE
FACULTÉ DE PSYCHOLOGIE
ET DES SCIENCES DE L'ÉDUCATION

Section de Psychologie

Sous la direction de Dr Thierry Lecerf

STABILITÉ À LONG TERME DES SCORES STANDARDS ET CHC DU WISC-IV : APPORTS THÉORIQUES ET CLINIQUES

THESE

Présentée à la
Faculté de psychologie et des sciences de l'éducation
de l'Université de Genève
pour obtenir le grade de Docteur en **Psychologie**

par

Sotta KIENG

de

Yverdon-les-Bains, VD

Thèse No 674

GENEVE

Août, 2017

No d'étudiant : 02-430-601

REMERCIEMENTS

J'aimerais ici adresser ma gratitude et mes remerciements à toutes les personnes qui m'ont aidée et soutenue durant l'ensemble de ce doctorat.

Tout d'abord, je tiens à remercier mon directeur de thèse, Dr Thierry Lecerf, pour son implication, son regard critique et sa grande disponibilité. Je lui suis tout particulièrement reconnaissante d'avoir guidé mes premiers pas en tant que jeune chercheuse et de m'avoir insufflé le goût pour les réflexions psychométriques.

Je voudrais remercier les membres de mon jury de thèse, Pr Nicolas Favez, Pr Jérôme Rossier et Pr Édouard Gentaz qui ont donné de leur précieux temps. J'ai apprécié leur analyse critique de mon travail et leurs commentaires constructifs. Je leur suis reconnaissante de leurs encouragements lors de l'aboutissement de ce manuscrit.

Un MERCI chaleureux à mes anciens collègues et amis, Isabelle, Philippe et Sophie pour les discussions éclairantes, les nombreux délires hilarants et le soutien mutuellement au cours de cette enrichissante expérience dans l'équipe EPEDI. J'ai partagé avec vous trois des moments qui rendent mes années de doctorat mythiques !

Je remercie mes anciennes collègues attachées de recherche et amies, Nathalie et Myriam ainsi que les étudiantes qui nous ont accompagnées dans les écoles. J'ai énormément apprécié leur implication, leur efficacité et notre bonne entente.

Mes remerciements vont également aux écoliers, à leurs parents, aux enseignants et aux directeurs d'école qui ont accepté de participer à cette recherche. La taille de notre échantillon et la qualité des données n'auraient pas été aussi notables sans leur précieuse contribution. Un merci particulier aux enfants pour leur enthousiasme, parfois timide, parfois débordant, mais toujours agréable durant les passations.

Je tiens à remercier tous ceux avec qui j'ai pu échanger sur ma thèse ainsi que mes amis pour les rigolades et leur soutien en toutes circonstances. Merci de m'avoir changé les idées et de m'avoir encouragée à persévérer.

Enfin, et non des moindres, j'aimerais infiniment remercier, mes parents et mes deux frères, pour toujours répondre présent lorsque j'en ai besoin et d'être fiers de mes réalisations. Dans mes moments de doute, je vous remercie de croire en moi jusqu'au bout.

RÉSUMÉ

Dès 6-7 ans, de nombreuses recherches longitudinales montrent que les performances cognitives aux tests sont relativement stables à court terme, mais également à long terme. Ces études se réfèrent essentiellement à la stabilité différentielle des scores entre au moins deux temps de mesure. Elles conduisent à l'hypothèse de la fidélité temporelle du score de Quotient Intellectuel (QI) et par extension à poser l'hypothèse de la stabilité du trait qu'il opérationnalise : l'intelligence. Partant d'un questionnement clinique sur la valeur prédictive des scores d'un test d'intelligence, notre étude se propose d'évaluer la fidélité des scores de l'une des batteries de tests cognitifs les plus utilisées pour la population enfants et adolescents : la quatrième édition de l'Échelle d'intelligence de Wechsler pour enfants et adolescents – le WISC-IV. Si de nombreuses études ont été réalisées sur les versions américaines précédentes (WISC, WISC-R, WISC-III), relativement peu d'études sur la stabilité à long terme des scores ont été conduites sur la 4e version américaine, et aucune sur l'adaptation en français. Or, les indices du WISC-IV sont non seulement interprétés comme des indicateurs du fonctionnement cognitif actuel d'un enfant, mais sont également utilisés pour faire des prédictions à long terme. Ces prédictions reposent sur l'hypothèse que les différences interindividuelles que mettent en évidence les scores du WISC-IV sont relativement stables au cours du temps. Quant à la stabilité intra-individuelle, elle n'est que peu abordée dans les études. Cela peut se comprendre. Dans la plupart des études longitudinales sur la stabilité de l'intelligence, le choix des tests cognitifs se porte sur ceux qui sont pratiques à administrer en collectif et rapides à coter. Ces tests sont à visée de recherche et non pour une évaluation clinique individuelle ; l'intérêt d'évaluer la stabilité intra-individuelle est moindre. S'agissant d'études sur la fidélité temporelle des scores du WISC, il est en revanche plus étonnant que la stabilité intra-individuelle soit si rarement développée. Notre étude pallie à ce manquement en étudiant la fidélité temporelle, c'est-à-dire la stabilité des scores du WISC-IV tant sur le plan interindividuel qu'intra-individuel. Ainsi, l'objectif principal de cette thèse est de favoriser une compréhension approfondie et une meilleure utilisation des scores du WISC-IV dans la pratique de l'évaluation psychologique. L'apport est à la fois théorique sur les connaissances relatives aux propriétés psychométriques de l'adaptation en français du WISC-IV et pratique sur des recommandations quant à l'utilisation des scores du WISC-IV. Pour cela, deux études sont réalisées : l'étude du fonctionnement différentiel des items du WISC-IV et l'étude de la stabilité des scores du WISC-IV. La première est exploratoire et contribue aux connaissances sur la validité de l'interprétation du WISC-IV, tandis que la seconde contribue à approfondir les connaissances sur la fidélité des scores du WISC-IV.

Le WISC-IV comporte dix subtests obligatoires et cinq subtests optionnels. L'interprétation courante se base sur un QI Total (QIT) et quatre indices standards : l'Indice de Compréhension Verbale (ICV), l'Indice de Raisonnement Perceptif (IRP), l'Indice de Mémoire de Travail (IMT) et l'Indice de Vitesse de Traitement (IVT). Ultérieurement à la publication de la quatrième édition en 2005, deux indices globaux se rajoutent aux indices standards : l'Indice d'Aptitude Générale (IAG) et l'Indice de Compétence Cognitive (ICC). Le WISC-IV n'est pas précisément rattaché à une théorie de l'intelligence, néanmoins, de nombreuses études montrent une correspondance entre la structure du WISC-IV et la grille de lecture du modèle des aptitudes cognitives de Cattell-Horn-Carroll (modèle CHC). Ce modèle de l'intelligence recueille un fort consensus dans la communauté scientifique et est actuellement le modèle contemporain dominant. Nous évaluons également la stabilité des scores du WISC-IV selon la classification CHC qui définit une structure en cinq indices : Compréhension-connaissances (Gc), Raisonnement fluide (Gf), Traitement visuel (Gv), Mémoire de travail (Gwm) et Vitesse de traitement (Gs).

Dans la première étude, l'objectif est d'évaluer si les items des différents subtests du WISC-IV se comportent de la même manière pour tous les individus qui ont la même habileté sur le trait latent évalué par le subtest considéré. On parle de fonctionnement différentiel d'un item, si des individus issus de différents groupes (ethnique, sexe, milieu socio-économique, etc.) ayant la même habileté sur le trait latent n'ont pas la même probabilité de réussir l'item. En revanche, si des individus issus de différents groupes ayant la même habileté sur le trait latent ont la même probabilité de réussir l'item, alors l'item ne présente pas un fonctionnement différentiel. L'évaluation du fonctionnement différentiel des items répond à la préoccupation pour l'équité dans l'évaluation psychologique.

L'échantillon de cette première étude est constitué de 483 enfants non consultants âgés de 7 à 12 ans (230 garçons et 253 filles) à qui on administre les dix subtests obligatoires et le subtest optionnel Complètement d'images. L'échantillon étant constitué d'enfants suisses francophones, la détection d'un fonctionnement différentiel des items s'est portée sur les variables âge, sexe et statut socio-économique. L'analyse des données pour cette première étude recourt aux modèles de réponse à l'item (MRI).

Pour les variables considérées (c.-à-d. âge, sexe et statut socio-économique), les résultats ne montrent aucun fonctionnement différentiel sur les items des subtests retenus – à savoir Cubes, Similitudes, Vocabulaire, Matrices et Compréhension. Ainsi,

l'adaptation en français du WISC-IV ne présente pas de biais d'items en fonction du groupe d'âge de l'enfant, de son sexe ou du statut socio-économique de ses parents. Tous les items se comportent de la même manière pour tous les individus. Pour les enfants ayant la même habileté sur le trait latent, leur probabilité de réussir aux items de Cubes, Similitudes, Vocabulaire, Matrices et Compréhension est la même quelque soit leur sexe, leur âge ou le statut socio-économique de leurs parents. Le trait latent (c.-à-d. l'habileté sur ce que le subtest évalue) explique à lui seul les performances des sujets sur ces subtests.

Notre seconde étude explore la stabilité des scores du WISC-IV au niveau inter- et intra-individuel. Les évaluations de la stabilité considérées sont : (1) la stabilité absolue, (2) la stabilité différentielle, (3) la stabilité intra-individuelle absolue, (4) la stabilité catégorielle et (5) la stabilité des forces et faiblesses. Une procédure test–retest, avec un intervalle de plus d'un an entre les deux passations est appliquée (moyenne = 1 an et 9 mois ; écart type = 6 mois). L'échantillon total se compose de 277 enfants non consultants âgés de 7 à 12 ans (132 garçons et en 145 filles).

Pour la stabilité absolue du niveau moyen du groupe, nous testons les différences de performances moyennes entre le test et le retest (*t*-test pour échantillons appariés). Les comparaisons de moyennes indiquent une stabilité du niveau moyen pour les indices ICV, IRP, IAG, Gc et Gv. Les autres indices présentent une augmentation significative des performances de la première à la seconde passation. L'augmentation peut être attribuée en partie à un effet d'apprentissage et au phénomène de régression vers la moyenne. Elle est petite à modéré pour l'IVT, l'ICC et Gs.

Pour la stabilité différentielle, nous examinons si les individus gardent le même rang au sein du groupe d'une passation à l'autre. Pour cela, nous calculons des coefficients de corrélation test–retest entre les scores à la première passation et à la seconde passation. Pour corriger la variabilité dans l'échantillon d'étude par rapport à la variabilité dans l'échantillon de standardisation, une correction de Magnusson est appliquée sur les coefficients de corrélation. Les résultats montrent des coefficients de stabilité corrigés qui varient de .66 (IMT) à .83 (QIT, IAG) pour les indices standards, et de .66 (Gwm) à .78 (Gv) pour les indices CHC. Dans une perspective de recherche ou de prise de décision à partir des performances moyennes d'un groupe, les indices ICV, IRP, QIT, IAG, ICC, Gc et Gv présentent une stabilité différentielle. Dans une perspective clinique et de prise de décision sur des scores individuels, aucun score n'atteint le seuil de .90. Le QIT et l'IAG présentent la plus haute stabilité dans les différences

interindividuelles à long terme, toutefois, la prudence reste de mise pour les prédictions sur des performances futures.

Nous intéressant également au niveau individuel si utile à la pratique clinique de l'évaluation psychologique, nous étudions la stabilité intra-individuelle à long terme sous trois angles : la stabilité intra-individuelle absolue, la stabilité catégorielle et la stabilité des forces et faiblesses personnelles.

Pour la stabilité intra-individuelle absolue, la question est de déterminer quels scores présentent des performances individuelles stables d'une passation à l'autre. Une performance est considérée comme stable si, à la seconde passation, elle reste dans l'intervalle de confiance de deux erreurs types de mesure (± 2ETM). Les résultats montrent qu'au moins 70% des enfants présentent des performances stables entre les deux passations pour les scores de l'ICV, l'IRP, l'IAG, Gf et Gv. Quant au QIT dont l'intervalle défini est de ± 7.26 points, seuls 60 % des enfants présentent des performances stables entre les deux passations.

Les résultats du WISC-IV sont fréquemment décrits de manière qualitative à l'enfant et à ses parents pour donner un sens plus parlant aux scores numériques. On peut relever trois systèmes de catégorisation qualitative des performances. La classification des performances en sept catégories : très faible (≤ 69), limite (70-79), moyen faible (80-89), moyen (90-109), moyen fort (110-119), supérieur (120-129), et très supérieur (≥ 130). La classification en trois catégories qui correspond à une lecture normative des performances de l'individu par rapport à son groupe de référence : faible (≤ 84), dans la moyenne (85-115), et élevé (≥ 116). Enfin, la classification en cinq catégories : extrémité inférieure (≤ 69), moyen faible (70-84), dans la moyenne (85-115), moyen fort (116-130), et extrémité supérieure (≥ 131).

Les résultats montrent que la classification en cinq catégories permet des descriptions nuancées et présente pour tous les scores une certaine stabilité catégorielle entre les deux passations. On observe un phénomène de régression à la moyenne qui tend à ramener les scores vers la moyenne de 100 à la seconde passation. Ainsi, les enfants, dont les performances sont dans la moyenne à la première passation, restent pour la plupart dans cette même catégorie à la seconde passation. En revanche, les enfants, dont les performances sont faibles ou élevées à la première passation, bougent d'une catégorie vers le haut ou d'une catégorie vers le bas pour revenir à des performances vers la moyenne à la seconde passation. Le phénomène de régression à la moyenne est plus fréquent chez les enfants qui, à la passation initiale, présentent des

performances faibles (< 85) que chez les enfants qui, à la passation initiale, présentent des performances élevées (> 115).

En complément à la comparaison normative, la comparaison ipsative (personnelle ou relative) des performances de l'individu par rapport à lui-même est également intéressante à réaliser. Pour cela, un indice moyen est calculé qui représente la moyenne des quatre indices (ICV, IRP, IMT et IVT) du profil d'un individu. Pour déterminer les indices déviants (c.-à-d. qui s'écartent de l'indice moyen), les performances de l'individu sur chaque indice sont comparées à son indice moyen. Un indice qui dévie significativement en étant supérieur à l'indice moyen détermine une force personnelle pour l'individu. À l'inverse, un indice qui dévie significativement en étant inférieur à l'indice moyen détermine une faiblesse personnelle pour l'individu. Si l'indice ne s'écarte pas significativement de l'indice moyen, il s'agit alors d'une moyenne personnelle. Les résultats montrent qu'à la première passation, l'ICV est l'indice le plus souvent en force personnelle, tandis que l'IMT est l'indice le plus souvent en faiblesse personnelle. À la seconde passation, l'IVT devient l'indice le plus souvent en force personnelle, tandis que l'IMT est toujours l'indice le plus souvent en faiblesse personnelle. À nouveau, le phénomène de régression à la moyenne est observé ; les forces ou les faiblesses personnelles tendent à devenir des moyennes personnelles à seconde passation.

Dans une perspective de recherche et de décisions sur la performance moyenne d'un groupe, les résultats sur un échantillon de 277 enfants suisses francophones suggèrent que l'Indice de Compréhension Verbale, l'Indice de Raisonnement Perceptif, l'Indice d'Aptitude Générale, Gc et Gv présentent une stabilité des différences interindividuelles. Dans une perspective clinique et de décision sur des scores individuels, le QIT et l'IAG classent les individus relativement au même rang sur le long terme. Néanmoins, aucun résultat sur le plan interindividuel ne permet de fonder des prédictions à long terme sur des performances futures d'un enfant sans une grande prudence et un examen de chaque cas. Sur le plan intra-individuel, il est important de tenir compte du phénomène de régression à la moyenne qui tend à ramener les scores vers la moyenne lors de mesures répétées. Si des prédictions sont possibles pour les performances autour de la moyenne, une grande prudence est en revanche de mise pour la stabilité à long terme des forces (performances très au-dessus de la moyenne) ou des faiblesses (performances très en dessous de la moyenne). L'ensemble des résultats appuie l'importance d'études psychométriques sur les tests utilisés comme aide à la prise de décision.

TABLE DES MATIÈRES

Remerciements ii
Résumé vi
Table des matières xii
Introduction 20
Cadre théorique 28
1. Évaluation de l'intelligence 29
1.1. Évaluation psychologique chez l'enfant 30
1.2. Tests en psychologie 34
1.2.1. Psychologie différentielle et psychométrie 34
1.2.2. Test mental, test psychologique et test psychométrique 36
1.2.3. Mesure vs évaluation 38
1.2.4. Intelligence, de quoi parlons-nous ? 42
1.3. Théories de l'intelligence 45
1.3.1. Approche psychométrique globale de l'intelligence 46
1.3.1.1. Binet et la première échelle métrique de l'intelligence 46
1.3.1.2. Stern, Terman et la conception du QI 49
1.3.1.3. Wechsler et l'esprit clinique de l'évaluation 49
1.3.2. Approche factorialiste de l'intelligence 52
1.3.2.1. Spearman et le modèle bi-factoriel de l'intelligence 53
1.3.2.2. Thurstone et les aptitudes mentales primaires 55
1.3.2.3. *g* psychologique vs *g* psychométrique 57
1.3.3. Approche hiérarchique de l'intelligence 59
1.3.3.1. Cattell, Horn et le modèle Gf-Gc étendu 61
1.3.3.2. Carroll et le modèle en trois strates 62
1.3.3.3. Le modèle CHC des aptitudes cognitives 64
1.4. Échelle d'Intelligence de Wechsler pour enfants et adolescents – WISC-IV 67
1.4.1. Indices standards du WISC-IV 68
1.4.2. Indices CHC du WISC-IV 72
2. Considérations psychométriques dans l'évaluation psychologique 75
2.1. Modèle de mesure en psychométrie 75
2.1.1. Théorie Classique des Tests 77
2.1.1.1. Le modèle : X = V + E 77
2.1.1.2. Postulats de la théorie classique des tests 82
2.1.1.3. Indice de difficulté et indice de discrimination 83

2.1.1.4. Limites de la théorie classique des tests 84
2.1.2. Modèles de Réponse à l'Item 85
2.1.2.1. Courbe Caractéristique d'Item 88
2.1.2.2. Paramètre de difficulté 89
2.1.2.3. Paramètre de discrimination 92
2.1.2.4. Paramètre de pseudo-chance 94
2.1.2.5. Courbe Caractéristique du Test 98
2.1.2.6. Concept d'information 99
2.1.2.7. MRI pour items polytomiques 102
2.1.2.8. Estimation des paramètres 105
2.1.2.9. Conditions d'application des MRI 106
2.2. Homogénéité 111
2.3. Sensibilité 111
2.4. Standardisation – étalonnages 113
2.5. Validité 115
2.6. Équité de l'évaluation 119
2.6.1. Notion de biais 120
2.6.2. Méthode des droites de régression 121
2.6.3. Fonctionnement différentiel des items 123
3. Fidélité des scores d'un test 127
3.1. Définition de la fidélité des scores 127
3.2. Méthodes d'estimation de la fidélité 130
3.2.1. Méthode test-retest 132
3.2.2. Méthode des formes parallèles immédiates/différées 134
3.2.3. Méthode de bissection 135
3.2.4. Méthode des covariances 136
3.2.5. Méthode interjuges 137
3.3. Interprétation de la fidélité des scores d'un test 137
3.3.1. Facteurs influençant sur l'estimation de la fidélité 138
3.3.1.1. Étendue des différences interindividuelle 138
3.3.1.2. Longueur du test 140
3.3.1.3. Difficulté d'un test 140
3.3.2. Seuils pour les coefficients de fidélité 141
3.3.3. Erreur de mesure et intervalle de confiance 146
3.4. Fidélité – stabilité des scores 153

3.4.1. Types d'évaluation de la stabilité 153
3.4.2. Stabilité des scores du WISC 154
3.4.3. Stabilité à court terme du WISC-IV 155
3.4.4. Stabilité à long terme du WISC-IV 159
4. Problématique 165
4.1. Fonctionnement différentiel des items du WISC-IV 165
4.2. Stabilité à long terme du WISC-IV 166
Méthode 172
5. Récolte de données 173
5.1. Échantillon 173
5.1.1. Échantillon étude 1 174
5.1.2. Échantillon étude 2 175
5.2. Procédure 176
5.2.1. Considérations éthiques 176
5.2.2. Procédure étude 1 177
5.2.3. Procédure étude 2 177
5.3. Instrument 178
5.3.1. Démarche de cotation des scores 179
5.3.1.1. Calcul indices standards 180
5.3.1.2. Calcul indices CHC 180
5.3.2. Épreuves 181
5.3.2.1. Cubes 181
5.3.2.2. Similitudes 182
5.3.2.3. Mémoire des chiffres 183
5.3.2.4. Identification de concepts 184
5.3.2.5. Code 185
5.3.2.6. Vocabulaire 186
5.3.2.7. Séquence Lettres-Chiffres 187
5.3.2.8. Matrices 188
5.3.2.9. Compréhension 189
5.3.2.10. Symboles 190
5.3.2.11. Complètement d'images 191
6. Analyses de données 193
6.1. Étude 1 : Fonctionnement différentiel des items 193
6.2. Étude 2 : Stabilité des scores du WISC-IV 196

6.2.1. Stabilité sur le plan interindividuel 196
6.2.1.1. Stabilité absolue 196
6.2.1.2. Stabilité différentielle 197
6.2.2. Stabilité sur le plan intra-individuel 197
6.2.2.1. Stabilité intra-individuelle absolue 197
6.2.2.1. Stabilité catégorielle 199
6.2.2.2. Stabilité des forces et des faiblesses personnelles 200
Résultats 202
7. Étude 1 : Le fonctionnement différentiel des items du WISC-IV 203
7.1. Évaluation de l'unidimensionnalité des items 203
7.2. Différences selon l'âge, le sexe et le statut socio-économique des parents 206
7.3. Fonctionnement différentiel des items du WISC-IV 211
8. Étude 2 : La stabilité du WISC-IV 214
8.1. Statistiques descriptives 214
8.2. Durée de l'intervalle test-retest, âge et différence de performances 215
8.3. Stabilité absolue 218
8.4. Durée de l'intervalle test-retest et effet d'apprentissage 221
8.5. Stabilité différentielle 223
8.6. Stabilité intra-individuelle absolue 225
8.7. Stabilité catégorielle 228
8.8. Stabilité des forces et des faiblesses 232
Discussion 236
9. Le fonctionnement différentiel des items du WISC-IV 237
10. La stabilité du WISC-IV 247
10.1. Stabilité absolue 248
10.2. Stabilité différentielle 251
10.3. Stabilité intra-individuelle 254
Conclusion et perspectives 264
Références 270
Annexes 290
Liste des abréviations et des sigles 317

To infer the existence of an ability from observed performance, the performance must exhibit some specified degree of temporal stability (be repeatable, or show consistency over time interval).

(Jensen, 1998, cité par Reeve & Bonaccio, 2011)

An individual's cognitive strengths and weaknesses . . . should also be stable over time if such patterns or characteristics are to have clinical utility.

(Canivez & Watkins, 2004, p. 113)

INTRODUCTION

Le présent travail s'inscrit dans le cadre d'un projet soutenu par le Fonds National Suisse de la Recherche Scientifique (FNS)[1] et dans la continuité d'une précédente étude FNS[2] dans laquelle la structure factorielle de la 4e édition de l'Échelle d'Intelligence pour Enfants et Adolescents de Wechsler (WISC-IV) a été étudiée. À la différence des résultats présentés dans le *Manuel d'interprétation* de la version française du WISC-IV (Wechsler, 2005b), les résultats de cette précédente étude, qui porte sur un échantillon de 249 enfants âgés de 8 à 12 ans provenant de différentes écoles du canton de Genève, ne soutiennent pas que le modèle en 4 facteurs est le meilleur pour les données du WISC-IV. En effet, dans cette première étude, le résultat phare est que la structure en 5 facteurs basée sur la théorie des aptitudes cognitives de Cattell-Horn-Carroll (CHC) montre un meilleur ajustement aux données que la structure en 4 facteurs privilégiée par les concepteurs du WISC-IV (Reverte, 2015). À l'instar d'autres études (Chen, Keith, Chen, & Chang, 2009; Keith, Fine, Taub, Reynolds, & Kranzler, 2006), cette étude recommande une lecture des scores du WISC-IV selon la théorie CHC pour une interprétation plus fine des aptitudes évaluées. Elle met également en évidence que les résultats d'analyse obtenus sur un échantillon avec certaines caractéristiques ne sont pas ipso facto généralisables à un échantillon avec d'autres caractéristiques, même la même version du test est utilisée. En effet, les propriétés psychométriques d'un test (p. ex., validité interne, fidélité) dépendent des caractéristiques du test et de l'échantillon testé. On ne peut pas transposer automatiquement les conclusions d'un test américain sur son adaptation en français. De même, on ne peut pas transposer les résultats obtenus sur une population française à une population suisse francophone ou belge. Dans la pratique, hélas, l'utilisateur d'un test tend à généraliser les résultats fournis par le manuel du test sans se questionner sur leur pertinence pour les individus qu'il est amené à évaluer. Cette situation est notamment rencontrée chez les psychologues en Suisse, où plusieurs régions linguistiques se partagent le territoire. Face à ce plurilinguisme et aux différences culturelles qui en découlent, peu d'éditeurs de tests investissent dans une adaptation spécifique à la population suisse. De ce fait, la population suisse romande se voit comparer aux étalonnages basés sur la population de France, la population suisse alémanique aux étalonnages basés sur la population d'Allemagne et la population suisse italienne aux étalonnages basés sur la population d'Italie. Or, les données psychométriques (notamment la fidélité et la validité) d'un test dépendent des

[1] Requête no. 135406, *Long-term stability of the WISC-IV: standard and CHC composite scores*. Requérant principal : Thierry Lecerf. Co-requérants : Nicolas Favez et Jérôme Rossier.

[2] Requête no. 118248, *Analysis of the French WISC-IV structure according to the Cattell-Horn-Carroll narrow ability classificatio*n. Requérant principal : Thierry Lecerf. Co-requérants : Nicolas Favez et Jérôme Rossier.

caractéristiques de l'échantillon testé. Il est donc important que des recherches sur des échantillons spécifiques soient réalisées afin de confirmer le caractère généralisable des résultats présentés dans le manuel d'un test.

Dans la lignée de la précédente étude, cette nouvelle étude poursuit l'évaluation de la version française du WISC-IV. Pour ce second volet, la validité de la structure du WISC-IV n'est plus au centre du questionnement, il s'agit cette fois d'explorer la stabilité à long terme des différents scores obtenus au WISC-IV. Le concept de fidélité est donc le cœur de ce travail. À travers la fidélité temporelle des scores d'un test qui évalue l'intelligence, nous posons une loupe spécifique sur l'intelligence : celle de la méthodologie (théorique et appliquée) de la mesure de l'intelligence. La perspective adoptée est résolument celle de la psychologie différentielle, et plus spécifiquement celle de la psychométrie.

Terme introduit par Stern (1911), la psychologie différentielle est une sous-discipline de la psychologie qui, au moyen de méthodes objectives, cherche à caractériser les différences psychologiques entre les individus. L'étude de ces différences s'intéresse principalement au domaine cognitif et au domaine conatif. Se fondant sur une approche scientifique, elle se constitue une méthodologie rigoureuse et des outils de mesure objectifs pour décrire les différences. Il s'agit là plus spécifiquement du champ de la psychométrie, qui est donc une branche de la psychologie différentielle. En effet, la psychométrie s'intéresse aux techniques d'évaluation des attributs psychologiques, ainsi qu'aux techniques d'élaboration et de validation de ces évaluations. Les tests psychologiques sont donc l'objet d'étude principal de la psychométrie, qui examine leur construction et leur utilisation. Dans le premier chapitre du présent travail, nous reviendrons sur la définition d'un test psychologique. Au sein de la psychométrie, trois grandes théories des tests sont élaborées, dont la théorie classique de tests (TCT). C'est à cette approche théorique, actuellement toujours dominante, que nous prenons comme référence pour l'évaluation de la fidélité des scores du WISC-IV. Les deux autres théories psychométriques sont : la théorie de la généralisabilité et la théorie de réponse à l'item. La première ne sera pas discutée dans ce travail. En revanche, la seconde servira de cadre à une analyse des items du WISC-IV, et plus spécifiquement, à la détection d'items biaisés. Outre l'étude principale relative à l'évaluation de la stabilité à long terme des scores du WISC-IV, une étude secondaire sera donc réalisée en préambule et porte sur le fonctionnement différentiel des items du WISC-IV. La détection d'items biaisés diminue la validité de l'interprétation qu'on peut faire sur les résultats d'un test,

car ce n'est plus le niveau d'habileté sur ce qu'évalue le test qui explique la performance (échec ou réussite) aux items, mais l'appartenance du sujet testé à un groupe particulier (homme/femme, minorités ethniques, etc.).

Objectifs de la thèse

Les échelles de Wechsler sont des instruments largement utilisés dans les services de consultation psychologique pour réaliser les bilans cognitifs des enfants et des adultes. Étant donné les répercussions des décisions et les enjeux qui y sont attachés (p. ex., dispense d'âge pour les sauts de classe, octroi d'une rente invalidité, évaluation judiciaire, etc.), le psychologue doit pouvoir fonder ses prédictions sur la base de scores suffisamment fidèles dans le temps. Or, peu de recherches ont exploré la stabilité du WISC-IV et, à notre connaissance, aucune n'en a évalué la stabilité à long terme de l'adaptation en français. Ainsi, des données empiriques sur la stabilité à long terme des scores du WISC-IV combleront des lacunes sur les données psychométriques de l'adaptation française, et conduiront également à améliorer l'utilisation de l'échelle dans le cadre clinique. Dans ce travail, nous serons guidée par la volonté de produire un apport directement utilisable pour la clinique. Aussi la stabilité à long terme du WISC-IV sera-t-elle explorée à la fois sur le plan inter- et intra-individuel.

Par une procédure test-retest, nous administrons le WISC-IV à deux reprises avec un intervalle de temps de plus d'une année entre les deux passations. Un échantillon total de 227 enfants tout-venant âgés de 7 à 12 ans est constitué auprès d'une vingtaine d'écoles du canton de Genève. D'abord, les comparaisons de moyennes entre la première et la seconde passation fournissent des informations sur d'éventuels effets d'apprentissage et leur ampleur. La question à laquelle cette première analyse répond est la question suivante : quel(s) indice(s) entre la première et la seconde passation présente(nt) des moyennes équivalentes ? Ensuite, les coefficients de stabilité calculés nous décrivent le degré de stabilité dans le rang des individus de l'échantillon. La question à laquelle cette deuxième analyse répond est aux questions suivantes : quel(s) indice(s) présente(nt) un classement similaire des individus entre la première et la seconde passation ? Est-ce que les plus forts (ou les plus faibles) à la première passation restent les plus forts (ou les plus faibles) à la seconde passation ? Nos analyses sur le plan intra-individuel renseignent sur trois aspects : la stabilité des performances individuelles à l'intérieur d'un intervalle défini de ±2 erreurs types de mesure, la stabilité catégorielle ainsi que la stabilité des forces et des faiblesses

personnelles. Elles répondent principalement à la question suivante : quelles prédictions à long terme peut-on avancer sur les performances futures d'un individu ? En clinique, les résultats du WISC-IV aident à orienter les interventions selon les difficultés et les ressources qu'ils mettent en lumière. Il est donc essentiel de pouvoir se fier à des résultats stables dans le temps. La stabilité au niveau groupal n'implique pas une stabilité au niveau individuel, et inversement. Un coefficient de stabilité, en tant que coefficient de corrélation, exprime le degré de stabilité dans le classement des individus. Dans l'ensemble, les individus peuvent garder leur position dans le groupe et néanmoins, obtenir des performances très différentes entre les deux passations. L'analyse des deux niveaux – inter- et intra-individuel – est donc importante pour une compréhension plus approfondie du WISC-IV.

Dans une étude qui ne fait pas partie des demandes du FNS[3], mais qui est motivée par un intérêt personnel, le fonctionnement des items de certains subtests du WISC-IV sera exploré grâce à l'apport de la Théorie de la Réponse à l'Item (ou en anglais : IRT pour *Item Response Theory*). L'utilité d'un instrument de mesure – tels que sont les tests psychologiques – réside dans les qualités psychométriques des items qui le composent. De ce fait, l'élaboration d'un test début par la sélection et l'analyse des items. Afin d'assurer l'équité – et par prolongement l'éthique – dans l'évaluation psychologique, les items d'un test ne doivent pas être biaisés, c'est-à-dire qu'ils ne doivent pas pénaliser (ou favoriser) un groupe d'individus par rapport à un autre (p. ex., les individus entre différents groupes d'âge, différents sexes, différentes cultures, etc.). Au moyen des modélisations selon le modèle de réponses à l'item logistique à deux paramètres (pour les subtests à cotation dichotomique des points), ainsi que selon le modèle gradué de Samejima (pour les subtests à cotation polytomique des points), le fonctionnement différentiel des items sera donc évalué sur les données de 483 enfants tout-venant âgés de 7 à 12 ans provenant de différentes écoles genevoises.

Plan de la thèse

Dans le premier chapitre consacré à l'évaluation de l'intelligence, nous commencerons par délimiter le cadre général du présent travail. L'instrument étudié tout le long de ce travail est une batterie d'intelligence fréquemment utilisée dans les évaluations psychologiques. Nous commencerons par définir la pratique de l'évaluation

[3] Requête no. 135406, *Long-term stability of the WISC-IV: standard and CHC composite scores*. Requérant principal : Thierry Lecerf. Co-requérants : Nicolas Favez et Jérôme Rossier.

psychologique dans laquelle se réalise l'évaluation cognitive. Nous nous intéresserons ensuite aux instruments qui permettent l'évaluation de l'intelligence : les tests psychologiques. Des définitions seront données pour poser les concepts, même si certains comme l'intelligence ne pourront pas être développés de manière exhaustive. L'intelligence est un vaste domaine de recherche qui touche une pluralité de champs de connaissances (médecine, biologie, etc.). Au sein d'un même champ comme la psychologie, on relève plusieurs points de vue qui apportent chacun un éclairage différent. Les questions sur son fonctionnement (psychologie cognitive) ou son développement (psychologie génétique) dépassent le cadre du présent travail qui s'inscrit comme nous l'avons dit dans la perspective de la psychologie clinique et différentielle. Les théories développementales (p. ex., Piaget, Wallon, etc.) ne seront pas mentionnées parmi les théories de l'intelligence que nous présenterons. Dans notre historique sur les théories de l'intelligence, nous retracerons les apports de différents auteurs tels que Binet, Spearman, Wechsler, Cattell, Horn et Carroll. Ce premier chapitre se terminera sur la présentation de l'instrument étudié dans ce travail (WISC-IV) et les indices sur lesquels portent nos analyses.

Le deuxième chapitre présente les considérations psychométriques dans l'évaluation psychologique. Il débutera par les modèles de mesure en psychométrie. Servant de cadre pour l'étude de l'évaluation de la fidélité des scores du WISC-IV, la théorie classique des tests sera d'abord présentée. La théorie de réponse à l'item sera également développée, car elle sert de cadre à l'étude du fonctionnement différentiel des items du WISC-IV. Ce deuxième chapitre se poursuivra sur les différentes qualités psychométriques à considérer dans l'utilisation d'un test psychologique. Le concept de validité fera l'objet d'un traitement plus approfondi pour nous permettre d'introduire la problématique de l'équité dans l'évaluation. Étant au centre de notre travail, la fidélité sera développée à part, dans le chapitre suivant qui lui sera entièrement consacré. Le deuxième chapitre se conclura sur la notion de biais d'un test qui explique l'intérêt de notre étude sur le fonctionnement différentiel.

Le troisième chapitre abordera la thématique centrale du travail avec le concept de fidélité selon la théorie classique des tests, et plus particulièrement le concept de stabilité. La première partie portera sur la fidélité des scores d'un test. Nous verrons les différentes procédures pour son estimation ainsi que les interprétations qu'on peut donner à un coefficient de fidélité. La compréhension approfondie du concept de fidélité éclairera sur la pertinence de mener des études avec des échantillons spécifiques afin de pouvoir généraliser les résultats. La seconde partie portera sur la

fidélité temporelle, ou stabilité, des scores d'un test. Une revue de la littérature fera état de la recherche sur la stabilité des scores du WISC-IV.

Le quatrième chapitre posera les objectifs et les questions de recherche sur les deux études menées : le fonctionnement différentiel des items du WISC-IV et la stabilité à long terme des scores du WISC-IV. Nous énoncerons également les hypothèses de recherche qui guident nos analyses de données.

Les chapitres cinq et six s'attacheront à décrire l'échantillon, le déroulement et les méthodes d'analyses des résultats pour les deux études réalisées. Les subtests du WISC-IV et le calcul des indices seront illustrés dans ce chapitre. Pour les analyses statistiques, les procédures seront aussi détaillées qu'il est nécessaire pour une compréhension conceptuelle.

La partie Résultats est composée de deux chapitres dans lesquels les différents résultats des études seront décrits et illustrés. La partie Discussion donnera la possibilité de mettre en lien les résultats et d'émettre des hypothèses pour les expliquer. Les limites des études réalisées seront également relevées. Le manuscrit se conclura sur les perspectives de recherches ultérieures.

CADRE THÉORIQUE

1. ÉVALUATION DE L'INTELLIGENCE

De par l'instrument étudié et les pistes qui seront proposées pour la pratique du psychologue clinicien, l'évaluation de l'intelligence est en toile de fond du présent travail. Ce premier chapitre lui est donc tout naturellement consacré. Nous commencerons par présenter en quoi consiste la pratique de l'évaluation psychologique dans lequel s'inscrit l'évaluation de l'intelligence (voir section 1.1). Après avoir présenté l'approche de l'évaluation, nous ciblerons sur un de ses principaux outils : les tests psychologiques (voir section 1.2). Nous définirons ce qui est entendu pas « tests » en psychologie ainsi que les deux champs au sein de cette dernière qui les étudient : la psychologie différentielle et plus précisément, la psychométrie. La notion de mesure en psychologie sera abordée pour rappeler les limites des tests psychologiques en tant qu'instrument de mesure. L'intelligence étant l'objet principal d'un bilan cognitif, nous nous attarderons sur ce concept polysémique et notamment sur les difficultés de lui poser une définition. Au travers des travaux de plusieurs auteurs, nous verrons le contexte historique dans lequel a émergé l'intérêt de la société pour l'évaluation de l'intelligence, ainsi que les théories de l'intelligence qui ont marqué l'histoire de la psychométrie (voir section 1.3). L'historique soulèvera également de vieux débats autour de la nature du facteur *g* qui demeurent encore et toujours d'actualité. Finalement, ce premier chapitre se terminera sur une présentation générale du test au cœur du présent travail : la 4e édition de l'Échelle d'intelligence de Wechsler pour enfants et adolescents (voir section 1.4).

Dans ce premier chapitre, nous présenterons donc le contexte général du présent travail. Nous poserons les définitions des termes récurrents afin de communiquer sur la même base. L'objet d'étude de la psychométrie que sont les tests sera défini, mais également débattu dans la discussion actuelle sur la distinction entre mesure et évaluation. Dès ce premier chapitre, nous aimerions d'emblée souligner les limites d'un test psychologique, qui, dans le cadre d'une utilisation éclairée, est toutefois un moyen précieux de recueil d'information.

1.1. ÉVALUATION PSYCHOLOGIQUE CHEZ L'ENFANT

L'activité d'évaluation n'est pas nouvelle dans le champ de la psychologie clinique. Dès les débuts de la psychologie moderne se construisent les premiers tests d'intelligence qui marquent un intérêt toujours grandissant pour l'évaluation. Si l'intelligence est un domaine fréquemment investigué, l'évaluation psychologique porte également sur d'autres domaines cognitifs (p. ex., traitement d'information, attention, langage, perception, etc.) ou sur les domaines conatifs (p. ex., personnalité, émotions, affects, etc.). Dans une enquête périodiquement menée auprès de psychologues-praticiens affiliés à l'*American Psychological Association* (APA), l'évaluation tient de manière stable la deuxième place des activités du psychologue-praticien sur une période de 1986 à 2010. Une première fois lancée par E. L. Kelly (1961), l'enquête est poursuivie dernièrement par Norcross et Karpiak (2012). En 2010, ces derniers ont transmis le questionnaire à 1'285 psychologues-praticiens sélectionnés au hasard parmi ceux membres de l'APA. Leur recueil final comporte 549 questionnaires complets et valides (42.7 % de taux de réponse). À la question sur leurs diverses activités professionnelles, 76 % des cliniciens répondent pratiquer la psychothérapie, 58 % l'évaluation psychologique, 49 % l'enseignement, 47 % la supervision clinique ainsi que la recherche et l'écriture scientifique. Se plaçant deuxième des activités les plus pratiquées par le clinicien juste après la psychothérapie, l'évaluation est identifiée comme un important marqueur professionnel dans la représentation du psychologue par le public. L'enquête de Norcross et Karpiak (2012) montre également que les cliniciens pratiquent le plus fréquemment des évaluations dans le domaine conatif (pratiqué par 54 % des cliniciens de l'étude) et dans le domaine cognitif (pratiqué par 48 % des cliniciens de l'étude). Étant un domaine plus spécifique et moins fréquent, l'évaluation neuropsychologique est pratiquée par 27 % des cliniciens répondants de l'étude.

En psychologie, la pratique de l'évaluation porte d'autres appellations telles qu'« examen » ou « bilan ». Nous allons nous attarder sur ces terminologies pour nous entendre sur leur sens et les nuances que chacune apporte. Pour commencer, le terme d'examen est souvent adopté bien qu'il soit teinté de connotations fortes telles que dans examen médical, examens dans le contexte scolaire, mise en examen par l'instance judiciaire, examen financier ou encore examen de conscience. Entendu comme examen psychologique, le terme d'examen renvoie à la fois à une investigation approfondie, un travail, d'observation, de description et de compréhension, une forme

d'introspection et surtout un processus ou une démarche en cours pour arriver à des conclusions. Utilisé par Binet, le terme d'examen psychologique est repris par ses successeurs pour le lien historique. Cependant, Binet n'administrait pas qu'un test d'intelligence aux enfants, il prenait aussi des mesures anatomiques (poids, taille, largeur d'épaules et de tête) qui apparentent sa pratique à un examen médical. Ensuite, prenons le terme de bilan qui fait d'abord penser à un état des lieux (bilan comptable, bilan d'exploitation), mais également au bilan qu'on dresse à certains moments de son existence (bilan de vie, bilan de ses réalisations, bilan d'une amitié). Entendu comme bilan psychologique, le terme bilan évoque un arrêt sur image pour faire le point, pour brosser le tableau de sa situation afin de se réorienter. L'idée de l'arrêt sur image sur la situation actuelle est pertinente, car elle rappelle la nature non figée mais évolutive de l'humain – et en particulier à la période de l'enfance. Elle souligne donc que la consultation survient à un moment donné dans l'évolution du sujet. Enfin, le terme d'évaluation renvoie à l'action d'apprécier la valeur d'une chose par une technique ou une méthode d'estimation. Une part de subjectivité est inhérente à l'acte d'évaluer qu'on ne retrouve pas dans l'acte de mesurer. En effet, « les résultats de la mesure sont purement descriptifs, c'est-à-dire qu'aucun jugement de valeur n'est porté, alors que le résultat d'une évaluation, lui, est subjectif » (Bernier & Pietrulewicz, 1997, p. 15). Certains regrettent justement que le terme d'évaluation attire l'attention sur l'idée d'« apprécier une dimension psychologique en la positionnant sur une échelle de valeurs » (Cognet & Bachelier, 2016). Mais n'est-ce pas illusoire de penser que les résultats restitués puissent échapper à une forme de jugement de valeur de la part du patient, de son entourage ou de la société ? En ce qui nous concerne, ce dernier terme rend bien compte à la fois de l'acte, qui consiste en « la mesure » d'une propriété psychologique et de la visée d'une telle démarche, qui est de situer l'individu par rapport à ses pairs. Ce terme est d'ailleurs la plus proche traduction du terme utilisé en anglais : *assessement*. Même si notre préférence se porte sur le terme d'évaluation, dans notre texte, nous employons les trois terminologies indifféremment pour nommer la pratique professionnelle du psychologue. Il n'y a pas d'arguments décisifs pour un terme plutôt qu'un autre, à chacun sa préférence selon sa sensibilité et sa filiation théorique. En revanche, nous reviendrons sur la distinction entre la mesure vs l'évaluation au moyen de tests, qui, quant à elle, ne relève pas de la nuance linguistique (voir section 1.2.3).

La pratique professionnelle qui nous intéresse plus particulièrement est celle de l'évaluation de l'intelligence chez l'enfant. Cependant, elle s'intègre souvent dans une

démarche plus complète, celle de l'évaluation psychologique, où les aspects cognitifs et les aspects psychoaffectifs sont également examinés. Décrire l'approche de l'évaluation psychologique revient finalement aussi à décrire celle de l'évaluation de l'intelligence, qui quant à elle, se distingue par une centration sur le fonctionnement intellectuel du sujet. À partir de la présentation de l'évaluation psychologique qui suit, nous relèverons le cas échéant les spécificités de l'évaluation de l'intelligence.

L'évaluation psychologique chez l'enfant se conçoit comme « une situation relationnelle au cours de laquelle un spécialiste applique des connaissances théoriques et des méthodes psychologiques à la compréhension dynamique d'un enfant présentant des difficultés à un moment donné de son évolution » (Voyazopoulos, Vannetzel, & Eynard, 2011, p. 92). Dans cette définition, l'examen psychologique s'envisage dans une perspective clinique comme un moment privilégié de rencontre entre l'enfant en difficulté et le psychologue. La démarche clinique de l'évaluation cherche à apporter un éclairage sur la nature et les origines des difficultés de l'individu (p. ex., intellectuelles, affectives, psychosomatiques, ou de dépendances toxicologiques). Pour cela, le psychologue recueille des informations auprès de sources plurielles et diversifiées (entretiens, tests, observations, échanges avec l'enseignant-e, la logopédiste, l'éducateur, l'assistante sociale, etc.), qu'il met en lien pour préciser et étayer ses hypothèses. Toute interprétation des performances à un test gagne en signification diagnostique, lorsqu'elle est corroborée par d'autres sources et quand elle est mise en lien avec les difficultés qui ont conduit à la demande d'évaluation. Le fonctionnement (intellectuel, cognitif, de la personnalité, etc.) du sujet est ainsi considéré de manière holistique et dynamique au travers des compétences et des potentialités du sujet, et non de manière fragmentée par l'addition de résultats et d'observations isolés. Les tests cognitifs s'étant considérablement développés, leurs résultats prennent de nos jours du poids dans les critères de décision (p. ex., dispense d'âge, placement en institution, admission dans une école de surdoués). Néanmoins, le psychologue pratiquant le bilan cognitif ne doit pas être réduit – ni se réduire – à un rôle de technicien. Le sens à donner aux résultats d'un test n'est pas immédiat ni mécanique, il s'élabore dans la cohérence avec d'autres sources d'information (entretiens cliniques, observations, échanges informels, etc.) et nécessite de connaître les limites de l'interprétation du test utilisé. L'utilisation d'un test psychologique est reconnue comme une spécificité du psychologue, même si parfois d'autres corps de la santé (notamment les médecins) peuvent être habilités à y recourir.

Plusieurs objectifs peuvent être poursuivis par le bilan psychologique : (a) mieux comprendre le fonctionnement et la personnalité d'un individu ; (b) contribuer à un diagnostic psychologique ; (c) contribuer à définir un projet thérapeutique ; (d) évaluer un changement après une intervention. Son déroulement suit plusieurs étapes comportant un entretien préalable pour l'anamnèse et l'étude de la demande avec les parents et l'enfant, puis des séances d'entretiens cliniques, d'observation du comportement et de passations de tests avec l'enfant, enfin, au terme du processus, un entretien de feedback et de propositions thérapeutiques. Il y a un travail de la part du psychologue de mise en lien et de construction d'un discours dans un langage compréhensible afin de permettre à l'enfant et aux parents l'appropriation des résultats du bilan psychologique.

Nous l'avons relevé, l'un des outils du psychologue pratiquant l'évaluation est le test d'intelligence. Il s'agit d'un outil particulièrement utilisé dans les bilans cognitifs, puisqu'il fournit une indication du fonctionnement intellectuel de l'individu. Les consultations chez les enfants font généralement suite à des difficultés d'apprentissage ou de comportement à l'école (troubles DYS-, TDA/H, etc.), mais il y a également de plus en plus de demandes pour l'identification d'un haut potentiel intellectuel. Quel que soit le motif de la consultation, la première inquiétude des parents se porte naturellement sur l'impact des difficultés sur les capacités intellectuelles de leur enfant et de surcroît, sur l'entrave pour ses apprentissages futurs ou pour ses relations interpersonnelles. Les bilans cognitifs, et plus particulièrement intellectuels, sont devenus routiniers au cours d'une évaluation psychologique. D'autant qu'évaluer le niveau de fonctionnement intellectuel s'inscrit de plus en plus comme une exigence administrative (p. ex., saut de classes, rente d'invalidité, placement institutionnel) ainsi que comme critère d'identification (p. ex., retard mental, haut potentiel intellectuel, trouble DYS-).

Actuellement, seules les batteries d'intelligence de Wechsler calculent un QI Total (sur la métrique de la moyenne à 100 et de l'écart type à 15). À la suite, nous allons présenter plus en détail l'échelle d'intelligence de Wechsler pour enfant et adolescent (WISC) qui est mentionnée tout au long du présent travail. Cependant, il important de présenter plus largement ce que sont les tests d'intelligence et les contributions empiriques qui font évoluer les conceptions de l'intelligence sur lesquelles ces tests reposent.

1.2. TESTS EN PSYCHOLOGIE

Les tests utilisés par les psychologues dans leur pratique sont bien différents des « psychos tests » qu'on s'amuse à remplir dans les magazines. Plusieurs critères doivent être remplis pour revendiquer avec légitimité l'appellation de test dans le sens entendu en psychologie. Avant de développer les définitions d'un test psychologique (voir section 1.2.2), nous allons d'abord rappeler les deux champs de la psychologie qui les prennent comme objet d'étude à savoir la psychologie différentielle et la psychométrie (voir section 1.2.1). L'utilisation de tests – entendu comme instrument de mesure quantitative – contribue à conférer le caractère de science à la psychologie. Or, les limites de la méthode des tests alimentent la mise en question de la mesure des propriétés mentales. De ce fait, nous ne pouvons échapper à une réflexion sur la distinction entre la mesure et l'évaluation au moyen d'un test psychologique (voir section 1.2.3). L'intelligence étant le domaine qu'évalue le test au cœur du présent travail, nous nous intéresserons également à sa définition (voir section 1.2.4).

1.2.1. PSYCHOLOGIE DIFFÉRENTIELLE ET PSYCHOMÉTRIE

On peut situer aux débuts de la psychologie moderne l'intérêt particulier pour l'évaluation de l'intelligence au moyen de dispositifs d'observation standardisés que sont les tests. Vers la fin du 19^{e} siècle, le psychologue allemand Wilhelm Wundt (1832 – 1920) étudie, dans le premier laboratoire de psychologie expérimentale, les variations des processus sensoriels élémentaires en fonction de l'intensité d'une stimulation et les seuils minimaux pour détecter un stimulus. Il y a parmi les étudiants de Wundt, un doctorant américain, James McKeen Cattell (1860 – 1944). À la différence de Wundt qui s'intéresse aux lois générales des processus sensoriels, J. M. Cattell porte son attention sur les variations interindividuelles stables qu'il relève chez les sujets soumis à l'expérience. Il approfondit cette observation dans ses propres travaux et, en 1890, est le premier à utiliser le terme de *mental test* pour désigner son matériel expérimental (J. M. Cattell, 1890). Parallèlement, en Angleterre, Francis Galton (1822 – 1911) poursuit des expériences sur l'acuité sensorielle, les seuils de discrimination et les temps de réaction dans lesquelles il recourt également à des épreuves dites standardisées. À partir de ce moment, l'intérêt pour les différences interindividuelles et les tests mentaux – et plus précisément les tests d'intelligence – vont en grandissant. Une nouvelle sous-

discipline de la psychologie se constitue alors sous le terme de psychologie différentielle introduit par le psychologue allemand William Stern en 1911.

À la différence d'autres sous-disciplines de la psychologie qui cherchent les ressemblances pour formuler des lois générales, la psychologie différentielle cherche à caractériser les différences relativement stables entre les individus au moyen de méthodes objectives. L'étude des différences stables se réalise principalement dans le domaine cognitif (l'intelligence, la mémoire, le traitement de l'information, le langage, etc.) et le domaine conatif (la personnalité, les émotions, les affects, etc.). Dans une perspective de recherche comparative, la psychologie différentielle pointe sur les différences pour établir des règles générales de fonctionnement psychologique selon les caractéristiques de l'individu ou du groupe. Elle place des individus différents dans des situations identiques pour les comparer et faire émerger des différences interindividuelles. On distingue (a) les différences entre les individus au sein d'un même groupe (variabilité interindividuelle), (b) les différences chez un même individu dans différents contextes ou à différents moments (variabilité intra-individuelle) et (c) les différences entre des individus appartenant à des groupes différents (variabilité intergroupe). La psychologie différentielle partage avec la psychologie clinique une centration sur l'individu, et s'en distingue par ses méthodes quantitatives pour l'observation des caractéristiques mentales. Se fondant sur une approche scientifique, elle se constitue une méthodologie rigoureuse qui s'appuie sur des méthodes d'analyses statistiques et des outils de mesure objectifs pour décrire les différences. Et pour répondre à son souci de rigueur méthodologique, elle développe en son sein une branche spécifique : la psychométrie.

Formée des mots grecs « psyché » (l'esprit) et « metron » (la mesure), la psychométrie signifie littéralement la mesure de l'esprit. On la définit comme « l'ensemble des théories et des méthodes de mesures en psychologie » (Dickes, Tournois, Flieller, & Kop, 1994, p. 11) ou encore comme « théorie et technologie des instruments de mesure en psychologie » (Reuchlin, 1992, p. 617). Nous aurons l'occasion de revenir sur la notion de mesure en psychologie qui est entendue dans un sens moins strict que dans les sciences naturelles. « Fille de la psychologie différentielle » (Pichot, 1999, p. 6), la psychométrie développe la méthode des tests pour mettre en évidence l'existence et l'importance de différences individuelles sur des traits de personnalité ou des composantes de la cognition (intelligence, langage, mémoire). Ses champs d'études portent sur les techniques d'évaluation des propriétés mentales (c.-à-d. les tests), ainsi que sur les techniques d'élaboration et de validation

des mesures recueillies (p. ex., les analyses factorielles, les corrélations, les équations structurales). Branche de la psychologie qui étudie la construction, les qualités métrologiques et l'utilisation des tests, ces derniers sont donc au cœur de la psychométrie. À la suite, il convient d'expliciter et de définir le terme de « test » dans son usage en psychologie.

1.2.2. TEST MENTAL, TEST PSYCHOLOGIQUE ET TEST PSYCHOMÉTRIQUE

Plusieurs définitions d'un test sont proposées dans les ouvrages de psychologie, on peut citer quelques-unes des plus courantes, dont celle de Pichot (1999) :

> On appelle test mental une situation expérimentale standardisée servant de stimulus à un comportement. Celui-ci est comparé statistiquement à celui d'autres individus placés dans la même situation, de manière à classer le sujet examiné par rapport à ceux constituant le groupe de référence. (p. 5)

Un test mental est donc une situation strictement définie qui permet de déclencher et d'enregistrer un comportement chez l'individu. Le comportement observé acquiert une signification dans la comparaison statistique avec un groupe de référence qui a été placé dans la même situation. Dans cette définition, la filiation des tests dans une perspective différentielle est clairement inscrite, puisque les tests servent à mettre en évidence des différences interindividuelles.

Huteau et Lautrey (2006) articulent plus explicitement les caractéristiques psychométriques d'un test et formulent leur définition ainsi :

> Un test est un dispositif d'observation des individus qui présente quatre propriétés :
>
> – il est standardisé ;
>
> – il permet de situer la conduite de chaque sujet dans un groupe de référence ;
>
> – le degré de précision des mesures qu'il permet est évalué (fidélité) ;
>
> – la signification théorique ou pratique de ces mesures est précisée (validité). (p. 20)

À nouveau, le test apparaît à la fois comme un instrument de mesure et une méthode d'observation, qui cherche à quantifier de manière objective et précise une

conduite. Un test permet de capter un échantillon de comportements par le biais d'une série d'items sur lesquels le sujet doit travailler. L'item d'un test est une question (simple ou complexe) qui se présente sous divers formats selon ce que le test essaie d'évaluer, la population à laquelle il s'adresse ou les modalités de passation. L'objectivité du test est conférée par une standardisation de la passation et des règles de cotation. Nous reviendrons sur les notions psychométriques de standardisation, de normes, de fidélité et de validité dans les prochains chapitres, poursuivons ici sur la distinction entre test mental, test psychologique et test psychométrique.

Le qualificatif de mental a été largement associé au mot test dans le passé, mais de nos jours, beaucoup lui préfèrent le terme de psychologique, moins connoté négativement (maladie mentale, malade mental, débile mental, etc.). On voit aussi apparaître le terme de test psychométrique, souvent utilisé dans la distinction avec test projectif. Ces derniers fournissent généralement peu de documentation sur leurs qualités métrologiques. Les tests psychométriques se réfèrent alors à des instruments qui fournissent des étalonnages pour comparer les performances et qui disposent de données sur leurs qualités métrologiques (preuves de fidélité, preuves de validité, etc.). Cette distinction n'est pas d'usage dans les classifications des tests psychologiques où la séparation se réalise habituellement entre d'un côté les tests cognitifs ou d'efficience (échelles composites d'intelligence, tests d'aptitudes, tests de connaissance, tests d'évaluation de processus) et de l'autre côté, les tests de personnalité (questionnaires et inventaires de personnalité, tests projectifs, tests d'intérêts ou de valeurs).

Dans les définitions présentées, un test psychologique est considéré comme un instrument de mesure. Mesurer revient fondamentalement à évaluer une grandeur. À l'inverse de grandeurs physiques (p. ex., hauteur, poids, distance, durée), les grandeurs qui intéressent la psychologie ne sont pas des phénomènes mesurables directement (à des exceptions près comme par exemple les temps de réaction). De plus, les propriétés mentales qu'on cherche à « mesurer » au moyen de tests ne sont pas des grandeurs additives. Peut-on alors dire qu'un test psychologique donne des mesures ? Nous allons maintenant discuter de la question de la mesure, qui agite toujours les psychométriciens.

1.2.3. MESURE VS ÉVALUATION

Pour Kant, philosophe allemand du 18e siècle, la psychologie ne peut être considérée comme une science au même titre que les sciences naturelles, étant donné qu'aucune analyse mathématique n'est possible sur les grandeurs psychologiques. Un siècle plus tard, l'homme de science britannique Galton énonce la même idée : « jusqu'à ce que les phénomènes d'une branche quelconque de connaissances aient été soumis à la mesure et au nombre, cette branche ne peut assumer le statut et la dignité d'une science » (cité par Pichot, 1999, p. 6). Un philosophe français du 20e siècle, Henri Bergson affirme qu'« il faut tracer une ligne de démarcation bien nette entre la physique et la physiologie d'une part, la psychologie d'autre part. Avec la psychologie, nous abordons l'étude d'objets réfractaires à la mesure » (cité par Martin, 1997, p. 320). Les déclarations de ces auteurs montrent que, dans l'histoire de la psychologie, la question de la mesure est étroitement liée à la revendication du statut de science. Néanmoins, notre réflexion ne va pas se porter sur la légitimité de la psychologie en tant que science. De par la démarche et les méthodes scientifiques qu'elle adopte, la psychologie fait pour nous partie des sciences humaines. On peut évidemment discuter sur ce que nous englobons dans la psychologie et ce que sont les sciences humaines. Mais pour ne pas digresser dans des discussions idéologiques, nous allons cibler notre réflexion sur la notion de mesure s'agissant des tests psychologiques. Ces derniers permettent-ils d'obtenir des mesures ?

Au cours du 20e siècle, l'essor du développement de tests mentaux montre une volonté claire pour la mesure des propriétés mentales (notamment l'intelligence). Cependant, la mesure telle que permise par un test psychologique n'est pas compatible avec la conception classique et dominante de la mesure en science. En effet, celle-ci « repose sur la définition d'une unité de mesure et sur le "comptage du nombre de fois que cette unité est présente dans la grandeur à mesurer" ; . . . la mesure produite, c'est-à-dire le nombre, est la marque d'une caractéristique réaliste de la grandeur mesurée » (Martin, 1997, p. 279). Or, avec les grandeurs psychologiques, on n'arrive pas à définir une unité/quantité élémentaire qu'on pourrait additionner. Par exemple, si l'on pense aux échelles d'intelligence, que représenterait une unité d'intelligence ? On n'arrive pas non plus à définir une valeur nulle. Comment concevoir un zéro absolu d'activités et de productions intellectuelles chez un être humain vivant ? De fait, « le zéro de ces échelles correspond simplement à l'échec de l'item le plus facile, lequel représente une borne toujours arbitraire » (Grégoire, 2007a, p. 44). En outre, la définition classique de

la mesure repose aussi sur une « conception réaliste des caractéristiques mesurables des objets : les objets mesurés sont supposés posséder de "vraies mesures", de "vraies magnitudes" (Martin, 1997). Or, les grandeurs psychologiques n'ont pas des propriétés directement accessibles. Les « mesures » que donne un test psychologique reflètent donc des quantités non divisibles et non observables. Elles sont également relatives, puisqu'elles rendent compte du positionnement de l'individu comparativement à un groupe d'autres individus placés dans la même situation.

Dans la conception paradigmatique de la mesure qui a cours au début du 20e siècle, la plupart des scientifiques (y compris des psychologues) refusent de considérer les grandeurs psychologiques comme mesurables. En 1920, le physicien expérimentaliste Norman Robert Campbell (1880–1949) cherche à apporter une clarification entre les entités mesurables et les entités non mesurables. Il introduit la notion de représentation et énonce la définition de la mesure comme suit : « *measurement is the assignment of numerals to represent properties* » (Campbell, 1920, cité par Martin, 1997, p. 290). Fondamentalement, Campbell reste néanmoins proche de la conception traditionnelle de la mesure. Il pose aussi la condition de relation d'ordre et la nécessité de l'opération mathématique d'addition sur les objets qu'on mesure. Pour cela, trois propriétés de l'addition doivent caractériser les choses mesurées : l'associativité[4], la commutativité[5] et la propriété affirmant que $a + 1$ est plus grand que a.

L'essor des tests mentaux contribue à poursuivre le débat sur la mesure en science. En 1946, le psychologue américain Stanley Smith Stevens (1906 – 1973) cherche à défendre une conception plus générale de la mesure afin d'englober les formes de mesures en psychologie. Il énonce : « au sens large la mesure est définie comme l'attribution de nombres aux choses ou aux événements suivant des règles » (Stevens, 1946, cité par Martin, 1997, p. 303). De son aveu, sa définition paraphrase celle de Campbell qu'il généralise aux choses et aux événements. Si en surface, les deux définitions sont proches, dans la définition de Stevens, la propriété additive des mesures est volontairement passée à la trappe. En fait, Stevens modifie l'esprit de la conception de Campbell ; « les propriétés ne sont plus posées *a priori* puis testées empiriquement mais sont établies par l'expérience . . . ce sont les opérations de mesure qui permettent d'identifier les propriétés [de la mesure] » (Martin, 1997, p. 304). Pour attribuer des nombres aux choses, il n'est plus besoin que lesdites choses possèdent les

[4] C'est-à-dire que : $a + (b + c) = (a + b) + c$

[5] C'est-à-dire que : $a + b = b + a$

propriétés d'additivité. La conception stevenienne de la mesure porte sur : quelles sont les règles d'attribution des nombres aux choses ? Et « de la réponse à cette question dépendent non pas la possibilité de mesurer mais les propriétés de la mesure effectuée » (Martin, 1997, p. 305). Pour Stevens, il existe donc différentes règles de conversion et également différents types de mesure. Il propose alors une classification des types de mesure en fonction de leurs propriétés empiriques. Sa classification définit quatre types d'échelle : (1) échelle nominale, (2) échelle ordinale, (3) échelle d'intervalle et (4) échelle de rapport. Une échelle nominale comporte des catégories différentes les unes des autres que l'on labellise par un nom (p. ex., fille et garçon pour le sexe, nationalité, métier). Chaque observation doit entrer dans une seule des catégories définies par l'échelle. Il y a une relation d'équivalence (d'égalité) entre les observations incluses dans une même catégorie. En plus des propriétés de l'échelle nominale, une échelle ordinale comporte des catégories ordonnées (p. ex., échelle d'appréciation de Likert). En plus de la propriété d'ordre, une échelle d'intervalle rajoute la notion de distance entre les catégories (p. ex., les échelles de température Celsius, date). La différence entre 20°C et 21°C correspond au même intervalle qu'entre 35°C et 36°C. Dans l'échelle d'intervalle, le zéro est situé de manière arbitraire. Par exemple, le zéro degré Celsius ne correspond pas à une absence de température. Finalement, une échelle de rapport comprend toutes les propriétés des précédents types d'échelle et comporte en plus la notion de proportionnalité (p. ex., distance, durée, âge). Le zéro n'est pas arbitraire et correspond, par exemple, à une absence de distance pour une échelle métrique.

Ainsi, d'un rapport entre une grandeur et une unité de cette grandeur, la définition de la mesure a intégré la notion de représentation avant de s'élargir grâce à Stevens vers une redéfinition de la mesure qui accepte l'inaccessibilité directe aux grandeurs (comme pour les grandeurs psychologiques). La publication de l'article « *Mathematics, Measurement and Psychophysics* » de 1951 dans lequel Stevens consolide théoriquement et empiriquement sa conception de la mesure, connaît une large et rapide diffusion dans la communauté scientifique. La définition usuelle en sciences physiques est alors perçue comme excessivement restrictive. Au sein de la psychologie, la conception stevenienne de la mesure et sa typologie des niveaux de mesure sont inscrites dans la méthodologie. Depuis Stevens, on trouve des définitions de la mesure plus « souples ». Par exemple, cette définition est proposée par l'*International Encyclopedia of Social Science* :

> La mesure est généralement considérée comme étant toute procédure par laquelle des nombres sont attribués à des individus (au sens de personnes, d'objets ou d'évènements) selon certaines règles. Ces règles précisent les caractéristiques d'un attribut ou d'un aspect quantitatif d'une observation, et définissent dès lors une échelle. (1968, cité par Martin, 1997, pp. 324–325)

Après avoir resitué la contribution majeure de Stevens dans l'acceptation de la psychologie en tant que science, puisqu'elle permet des mesures quantitatives, nous allons soulever un débat qui divise actuellement la communauté des psychométriciens. Un front minoritaire, porté notamment par l'épistémologue de la psychologie Joel Michell, conteste le caractère quantitatif d'un attribut psychologique. L'une des hypothèses qui sous-tendent les théories de mesure des tests et leur utilisation, est que les différences interindividuelles dans les scores d'un test traduisent des différences quantitatives sur la propriété mentale. Or selon Michell (1999, 2000, 2004), aucune donnée ne soutient cette hypothèse et il reproche aux psychométriciens de la postuler sans chercher à la vérifier. Illustrons l'hypothèse quantitativiste des tests en prenant l'exemple d'un test qui évalue la vitesse de traitement. D'après cette hypothèse, les différences de scores sur le test révèlent des différences interindividuelles dans la quantité d'information traitée. Les individus avec les plus hauts scores traitent quantitativement plus d'information que les individus avec des scores plus faibles. Cette lecture omet l'hypothèse que les individus puissent adopter différentes stratégies pour répondre à un item d'un test et que, de plus, un même individu puisse adopter plusieurs stratégies au cours de la passation du test. Si comme Michell, on suppose que les individus ne s'y prennent pas de la même manière pour répondre à un item. Les différences de scores au test peuvent traduire des différences de nature qualitative. Dans l'exemple précédent du test de traitement d'information, ce n'est pas la quantité d'information que peuvent traiter les individus qui conduisent à des différences, mais la manière dont les individus traitent l'information. Dans cette seconde interprétation, les tests ne fournissent pas des mesures, mais une évaluation ou une estimation de la propriété mentale. Le score obtenu à un test ne traduit pas seulement de la quantité de la propriété mentale évaluée chez l'individu, mais tienne également compte de la qualité avec laquelle le sujet a fait appel à la propriété mentale.

Comme un test n'est pas une mesure pure de la propriété mentale qu'il infère à partir des réponses aux items et que les individus ne répondent pas forcément aux items avec la même stratégie, nous partageons la position qui consiste à préférer l'emploi des termes « évaluer » ou « évaluation » au lieu de « mesurer » ou « mesure ». La critique virulente de Michell sur la conception quantitativiste des tests adoptée par

un pan des psychométriciens – qu'il assimile pour leur pratique imprudente à des astrologues – est une réflexion épistémologiste à garder en tête dans l'utilisation et l'interprétation des scores d'un test.

Un dernier terme reste à définir dans ce chapitre : l'intelligence. L'objectif peut paraître périlleux à atteindre tant il s'agit d'une entité tentaculaire. Cependant, gardons à l'esprit que, dans le présent travail, nous examinons l'intelligence dans la perspective méthodologique (théorique et appliquée) de sa mesure.

1.2.4. INTELLIGENCE, DE QUOI PARLONS-NOUS ?

Malgré la multitude d'études et le vif intérêt pour le domaine, aucune définition de l'intelligence n'appelle un consensus universel que cela soit en psychologie ou en philosophie (Lanz, 2000). Selon les auteurs, on observe des différences – parfois importantes – sur les délimitations des champs de compétences et de comportements qu'englobe l'intelligence. Au fil des écrits sur l'intelligence et son évaluation, les contours d'une définition de l'intelligence se dessinent. Nous allons en présenter quelques-uns pour illustrer les composantes de l'intelligence que cherchent à évaluer les tests d'intelligence.

Commençons en 1921 où se tient un symposium intitulé *Intelligence and its Measurement* regroupant les experts de l'époque. Quatorze d'entre eux se prononcent sur leur définition de l'intelligence. Thorndike (1921) donne une définition de l'intelligence comme *the power of good responses from the point of view of truth or fact* (p. 124). D'autres auteurs relèvent des composantes qui vont devenir récurrentes dans les définitions de l'intelligence : « *the ability to carry on abstract thinking* » (Terman, 1921, p. 128), « *the ability to adapt oneself adequately to relatively new situations in life* » (Pintner, 1921, p. 139), « *learn to adjust himself to his environment* » (Colvin, 1921, p. 136). À l'issue du symposium, il y a à la fois un certain accord sur ce qu'est l'intelligence en général et des écarts sur ce qui la compose en particulier.

Dans une étude américaine, Snyderman et Rothman (1987) sollicitent par questionnaire 1020 psychologues et spécialistes de l'éducation ayant une expertise dans le domaine de l'intelligence. Leur questionnaire est composé de 48 questions sur différentes controverses autour de l'intelligence et des tests d'intelligence, qui sont sous le feu de la critique pour leur biais en défaveur des minorités ethniques ou des groupes socio-économiques bas, ou encore pour leur stigmatisation de ceux ayant de

basses performances. Snyderman et Rothman recueillent 661 questionnaires remplis (65 % de taux de réponse). À la question sur l'énumération des composantes importantes de l'intelligence, les dix caractéristiques suivantes se révèlent récurrentes parmi les répondants : pensée ou raisonnement abstrait (99.3 %), aptitude à résoudre des problèmes (97.7 %), capacité à acquérir des connaissances (96 %), mémoire (80.5 %), adaptation à l'environnement (77.2 %), vitesse mentale (71.7 %), compétence linguistique (71 %), compétence en mathématique (67.9 %), culture générale (62.4 %), créativité (59.6 %). À la question d'estimer si les composantes de l'intelligence sont adéquatement évaluées par les tests d'intelligence existants, les répondent déplorent les lacunes des tests d'intelligence pour la créativité (88.3 %), pour l'adaptation à l'environnement (75.3 %) et pour la capacité à acquérir de nouvelles connaissances (42.2 %). À travers les composantes relevées par les répondants de l'étude se dessine une définition de l'intelligence qui rend compte d'une vision principalement académique de l'intelligence, axée sur des compétences valorisées dans la réussite académique.

À la suite de la publication très débattue de « *The Bell Curve* » écrite par le psychologue Herrnstein et le politologue Murray (1994), l'*American Psychological Association* (APA) constitue un groupe de travail pour rafraîchir l'état des connaissances empiriques liées à l'intelligence. Nous ne développerons pas davantage sur les controverses de l'ouvrage de Herrnstein et Murray, mais pour donner un exemple, l'une d'elles est leur conclusion selon laquelle les différences de QI dans la population américaine relèvent de la génétique et des différences raciales. Dans leur réflexion, la *task force* de l'APA définit ainsi l'intelligence :

> *Ability to understand complex ideas, to adapt effectively to the environment, to learn from experience, to engage in various forms of reasoning, to overcome obstacles by taking thought*. (Neisser et al., 1996, p. 77)

Dans une vision de l'intelligence proche des répondants de l'étude de Snyderman et Rothman (1987), cette définition ne met cependant pas explicitement en évidence certaines composantes telles que la créativité ou les compétences linguistiques.

Également en réaction à la publication de « *The Bell Curve* », 52 chercheurs du domaine de l'intelligence ont approuvé un texte qui définit l'intelligence comme suit :

> *Intelligence is a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience. It is not merely book learning, a*

> *narrow academic skill or test taking smarts. Rather it reflects a broader and deeper capability for comprehending our surroundings, catching on, making sense of things, or figuring out what to do.* (L. S. Gottfredson, 1997, p. 13)

Par cette définition de l'intelligence, un élargissement est apporté, qui souligne une capacité mentale transversale à comprendre, à donner du sens et à savoir quoi faire dans différentes situations.

Ces différentes définitions proviennent d'experts. Une question qu'on peut se poser est de savoir si la vision de l'intelligence des experts sur laquelle les tests d'intelligence sont construits, coïncide avec celle de l'homme de la rue. Dans une recherche auprès de 140 psychologues ayant une expertise dans domaine de l'intelligence et de 122 sujets « profanes » (p. ex., des étudiants dans une bibliothèque, des gens attendant un train ou entrant dans un supermarché), Sternberg, Conway, Ketron, et Bernstein (1981) montrent que la conception de l'intelligence entre les deux groupes ne diffère pas radicalement. Le groupe des profanes évoque dans l'ordre (1) la capacité à résoudre des problèmes (p. ex., raisonnement logique, faire des connexions entre plusieurs idées), (2) les compétences verbales (p. ex., élocution claire, savoir s'exprimer avec aisance) et enfin (3) les compétences sociales (p. ex., acceptation de l'autre tel qu'il est, savoir reconnaître ses propres erreurs). De leur côté, les experts évoquent d'abord (1) l'intelligence verbale (p. ex., utilisation d'un vocabulaire adéquate, compréhension verbale), puis (2) la capacité à résoudre des problèmes (p. ex., capacité d'appliquer ses connaissances pour résoudre un problème, prendre de bonnes décisions) et enfin (3) l'intelligence pratique (p. ex., adaptation à une situation, capacité de planifier pour atteindre un but). D'après les résultats de cette recherche, la conception de l'intelligence formulée par les experts du domaine reçoit un écho dans la population qui semble également privilégier une intelligence basée sur des compétences cognitives. Une légère différence est que les profanes semblent mettre en avant l'intelligence d'un individu quand elle se manifeste dans les interactions sociales.

Le manque de consensus sur une seule et unique définition s'explique, d'une part, parce que les chercheurs peuvent appréhender l'intelligence sous de nombreux angles et, d'autre part, parce qu'il y a un jugement de valeur inhérente à toute définition de l'intelligence. Concevoir l'intelligence – et donc la définir – est porteur d'enjeux importants pour les individus qui la survalorisent dans leur image de la réussite. Malgré l'absence d'une définition partagée par tous, cela n'est pas un frein aux nombreuses recherches menées dans ce domaine. Dans les études longitudinales, les résultats des corrélations montrent que les différences interindividuelles sur

l'intelligence tendent à être stables dans le temps dès 6-7 ans (p. ex., Deary, Whiteman, Starr, Whalley, & Fox, 2004; Hertzog & Schaie, 1986; R. A. Hoekstra, Bartels, & Boomsma, 2007; W. Schneider, Niklas, & Schmiedeler, 2014). L'intelligence est supposé un trait stable dans le temps, même si certaines fluctuations dans les performances peuvent survenir à cause de la fatigue, de la motivation et d'autres facteurs. À l'inverse d'un état, un trait représente une caractéristique durable et une (pré-)disposition. Le trait est sous-jacent à une configuration de conduites qui montrent une relative stabilité temporelle et, une relative cohérence intra-individuelle et trans-situationnelle. Le trait résume les conduites d'un individu, les explique et permet de les prévoir dans une situation similaire.

Pour résumer la situation actuelle, on peut citer le psychologue Roger Lécuyer : « il est courant chez les pessimistes de dire qu'il y a autant de définitions de l'intelligence qu'il y a de spécialistes, les optimistes pensant, eux, qu'il y a seulement autant de définitions que de théories » (2009, p. 11). Le bref historique sur l'évolution de l'interprétation des tests d'intelligence sera l'occasion pour nous de présenter quelques-unes des grandes théories de l'intelligence. Les tests d'intelligence sont construits pour opérationnaliser les théories de l'intelligence. Ils reposent donc sur une définition de l'intelligence qui provient des experts, et moins sur des définitions qu'on peut retrouver chez les profanes. Cela explique que les batteries d'intelligence actuelles n'intègrent pas la créativité par exemple.

1.3. THÉORIES DE L'INTELLIGENCE

Au travers des contributions théoriques et scientifiques de grands noms de la psychométrie, nous allons voir comment le concept d'intelligence est opérationnalisé dans les tests cognitifs. Notre tour d'horizon des théories de l'intelligence couvre la période de Binet au modèle contemporain et dominant de Cattell-Horn-Carroll, toutefois, il ne prétend pas à l'exhaustivité. Le regroupement des théories s'effectue selon la conception de l'intelligence qu'elles décrivent ; le fil des contributions ne suit pas un ordre strictement chronologique et certaines périodes se chevauchent. En outre, les théories plus développementales de l'intelligence (p. ex., la théorie piagetienne) dépassent le cadre de ce travail et ne seront donc pas présentées.

1.3.1. APPROCHE PSYCHOMÉTRIQUE GLOBALE DE L'INTELLIGENCE

À la suite des travaux de l'économiste sociologue allemand Weber, du psychologue allemand Fechner, de Galton en Angleterre, de Wundt dans le laboratoire de psychologie expérimentale à Leipzig et de J. M. Cattell en Amérique émerge un vif intérêt pour les différences entre les individus sur leurs traits psychologiques. Nous sommes à l'aube du 20e siècle, la psychologie qui se veut scientifique émerge avec la volonté « d'établir un savoir positif, c'est-à-dire explicité et validé, sur le psychisme humain, savoir qui se distinguerait radicalement des discours de la tradition philosophique, de la littérature et de la clinique médicale » (Huteau, 2006, p. 24). Pour cela, la psychologie se constitue une méthodologie qui décrit les procédures permettant des observations objectives, contrôlées et précises, dont font partie les tests mentaux.

Les concepteurs des premiers tests d'intelligence construisent des instruments qui permettent principalement d'identifier chez l'individu un niveau global d'efficience intellectuelle. Il y a durant cette époque une forte demande sociale pour classer les individus dans des groupes distincts (p. ex., élèves adaptés à l'école vs élèves en difficultés scolaires ou l'attribution du poste le plus adapté à un soldat). Ainsi, l'approche psychométrique globale de l'intelligence considère le fonctionnement cognitif de l'individu dans son ensemble. Il s'agit néanmoins d'une vision plurielle – et non unitaire – de l'intelligence, puisque différentes composantes du fonctionnement cognitif sont mises en perspective dans le calcul de la note globale. Les auteurs que nous rattachons à l'approche psychométrique globale de l'intelligence sont Binet, Stern, Terman et Wechsler.

1.3.1.1. Binet et la première échelle métrique de l'intelligence

Ce début de 20e siècle est marqué par de profondes transformations économiques, sociétales et culturelles dans les pays développés (notamment en France, en Angleterre, en Allemagne et aux États-Unis). L'enseignement public se démocratise avec l'arrivée sur les bancs d'école d'enfants de différents milieux socio-économiques. Les interrogations et les préoccupations liées aux enfants en difficultés scolaires conduisent à rechercher des outils permettant la caractérisation des individus selon leur efficience intellectuelle. En France, le Ministère de l'Instruction Publique mandate le psychologue Alfred Binet (1847 – 1911) pour identifier les enfants inaptes à profiter du

système d'enseignement public, devant ainsi être redirigés dans des classes de perfectionnement. En collaboration avec un jeune médecin aliéniste Théodore Simon (1873 – 1961), Binet cherche une manière objective pour dépister les enfants déficients mentaux parmi l'ensemble des élèves en échec scolaire. Dans leurs travaux, ils développent une batterie d'épreuves – connue depuis sous le nom de « test Binet-Simon » ou « Échelle métrique de l'intelligence » – dont la première version est publiée en 1905. Quelques années plus tard, la publication d'une nouvelle version en 1908 élargira le champ d'application du Binet-Simon à des contextes non scolaires (p. ex., repérage des « anormaux militaires »). Une troisième version est publiée avant la mort prématurée de Binet en 1911. À la suite des travaux du psychologue français René Zazzo (1910 – 1995), le Binet-Simon est à nouveau révisé et devient la Nouvelle Échelle Métrique de l'Intelligence (NEMI), dont la version la plus récente est la NEMI-2 (Cognet, 2006).

Si l'intérêt pour la quantification des différences individuelles est déjà présent avant les travaux de Binet, on lui attribue toutefois la naissance des tests d'intelligence dans leur forme moderne (ensemble d'épreuves sur des processus généraux). Jugeant naïve et réductrice toute tentative de définir un concept aussi complexe que l'intelligence, il s'abstient de proposer une véritable définition. Toutefois, il pose certaines propriétés à l'intelligence. (a) Il existe plusieurs formes d'intelligence ; ainsi l'intelligence entre l'enfant et l'adulte ne se distingue pas seulement en degré et en quantité. (b) Les différences d'intelligence au sein d'un groupe sont surtout qualitatives plutôt que quantitatives. (c) L'intelligence est constituée d'un faisceau d'aptitudes, éventuellement indépendantes les unes des autres. Selon Martin (1997), Binet conçoit l'intelligence comme une entité qui « englobe, elle unit, toutes les capacités mentales des individus. Elle ne se réduit pas, en tout cas, à une grandeur unique dont les variations de niveau expliqueraient les différences entre individus » (p. 33). L'évaluation de l'intelligence nécessite donc un ensemble d'épreuves cognitives relativement complexes et variées qui chacune fait appel à des aptitudes spécifiques. « Un test particulier, isolé de tout le reste, ne vaut pas grand-chose . . . ce qui donne une force démonstrative, c'est un faisceau de tests, un ensemble dont on conserve la physionomie moyenne » (Binet, 1910, p. 200).

Dans le débat sur la mesure des grandeurs psychologiques, Binet et Simon ne soutiennent pas que leur test permette une mesure de l'intelligence au sens mathématique ou physique.

> En quoi consiste au juste la mesure de l'intelligence ? . . . le mot mesure n'est pas pris ici au sens mathématique : il n'indique pas le nombre de fois qu'une quantité est contenue dans une autre. L'idée de mesure se ramène pour nous à celle de classement hiérarchique ; de deux enfants, est le plus intelligent celui qui réussit le mieux un certain ordre d'épreuves. (Binet, 1911, cité par Martin, 1997, p. 34)

Le terme mesure est utilisé par commodité et souvent Binet et Simon le remplacent par « appréciation » de l'intelligence. En ramenant la notion de mesure à un classement hiérarchique, le test du Binet-Simon fournit une échelle ordinale. Binet accorde peu d'intérêt aux outils statistiques et élabore une démarche clinique dans laquelle l'articulation des signes recueillis exige un travail d'interprétation du psychologue.

Réputé pour son efficacité pratique, le test Binet-Simon est exporté outre-Atlantique et traduit par Henri Goddard (1866 – 1957), psychologue américain travaillant dans une institution pour déficients mentaux. Le succès que va connaître le Binet-Simon (puis le test Stanford-Binet) aux États-Unis relance le développement d'autres tests dits « mentaux », qui, à la différence des tests physiologiques ou anthropométriques, mesurent des phénomènes purement psychiques et mentaux. De même qu'en France, la société américaine a un contexte qui favorise l'application et la diffusion des tests d'intelligence. D'abord, il y a une demande de l'administration de l'immigration pour identifier les déficients mentaux parmi les nouveaux arrivants sur le Nouveau-Continent. Puis vers 1917, l'Amérique est à la veille de son entrée dans la Première Guerre mondiale. Les recrues militaires étant un bassin hétéroclite d'hommes, l'armée américaine s'ouvre aux apports des travaux en psychologie et donne la légitimité à des psychologues pour mener leurs recherches auprès de millions de recrues. Le test *Army alpha* (épreuves avec du verbal) et sa version *Army beta* (épreuves non verbales pour les non-anglophones), entre autres, sont administrés aux nouveaux militaires pour déterminer leur affectation. Après la Première Guerre mondiale, le *testing* va également intéresser les universités (pour la sélection des admissions) et les recruteurs du monde du travail. Les psychologues se voient aussi conviés à faire état d'expertise au moyen de tests dans les cours de justice.

1.3.1.2. Stern, Terman et la conception du QI

Composé d'une série d'épreuves adaptées à chaque âge de 3 à 15 ans, le Binet-Simon calcule un niveau intellectuel qui permet de déterminer le nombre d'années de retard ou d'avance de l'enfant.

> En clair, le niveau intellectuel d'un enfant réussissant toutes les épreuves de 3 ans jusqu'à 9 ans est de 9 ans. En calculant la différence entre le niveau intellectuel (ici 9 ans) et l'âge réel (par exemple 10 ans), il est possible de classer les enfants en différents groupes suivant le nombre d'années de retard ou d'avance (ici un an de retard). (Martin, 1997, p. 29)

La notion de « niveau intellectuel » introduite par cette échelle, est plus tard modifiée en « âge mental », puisqu'il s'agit bien d'un âge calculé. Avec les contributions successives du psychologue allemand Stern en 1912 et du psychologue américain Terman en 1916, la notion d'« âge mental » devient « le quotient mental », qui est, quant à lui, ensuite remplacé par le « Quotient Intellectuel » (QI). En effet, William Stern (1871 – 1938) propose de définir un quotient mental qui est le ratio entre l'âge mental et l'âge chronologique $\left(\frac{Age\ mental}{Age\ chronologique}\right)$. Quant à Lewis Terman (1877 – 1956), professeur à l'université de Stanford, il réalise des études sur les épreuves du Binet-Simon qu'il cherche à améliorer. En 1916, il publie une adaptation sous le nom de test « Stanford-Binet » qui, à la différence du Binet-Simon, est destiné à la mesure de l'intelligence aussi bien des enfants que des adultes. Son test calcule désormais le rapport entre l'âge mental et l'âge chronologique multiplié par 100, soit un Quotient Intellectuel Développemental $\left(QID = \frac{Age\ mental}{Age\ chronologique} \times 100\right)$. Avec la contribution de la psychologue Maud Merrill (1888 – 1978), Terman propose une deuxième révision du Stanford-Binet en 1937. Test d'intelligence de référence jusqu'à l'apparition de la première échelle de Wechsler – le Wechsler-Bellevue – en 1939, le Stanford-Binet « servira de critère de validation pour les nouveaux tests » (Huteau, 2006, p. 28). Toujours largement utilisée, la dernière version en date est la 5^{e} édition du Stanford-Binet (SB-5, Roid, 2003).

1.3.1.3. Wechsler et l'esprit clinique de l'évaluation

Un autre psychologue américain que nous pouvons affilier à l'approche psychométrique globale est David Wechsler (1896 – 1981). Pour concevoir ses échelles

d'intelligence, ce dernier s'inspire de ses expériences cliniques au *Bellevue Psychiatric Hospital* de New York et dans l'armée américaine où psychologue fraîchement diplômé, il administre les tests *Army alpha* et *Army beta*. En 1939, il publie la *Wechsler-Bellevue Intelligence Scale* (WBIS) destinée à une population de 7 à 69 ans. Puis en 1946 paraît la Wechsler-Bellevue Forme II (WBIS-II) destinée à des individus de 10 à 79 ans. S'ensuivent le *Wechsler Intelligence Scale for Children* (WISC ; Wechsler, 1949) et la *Wechsler Intelligence Scale for Adults* (WAIS ; Wechsler, 1955), deux échelles qui remplacent les Wechsler-Bellevue. Composé d'une dizaine de subtests, les premières échelles de Wechsler permettent d'évaluer une intelligence globale opérationnalisée par le QI Total et deux indices (QI Verbal, QI Performance). Les subtests sont regroupés sous l'un des deux indices selon qu'ils font appel à un mode de communication verbale (QIV) ou non verbale (QIP).

L'étendue des âges à qui s'adresse les échelles soulève un problème dans la manière de calculer un QI en fonction du rapport entre l'âge mental et de l'âge chronologique. Selon la formule de Terman, si l'âge mental correspond à l'âge chronologique alors le QI calculé est de 100, correspondant à une performance dans la moyenne du groupe de référence $\left(QID = \frac{Age\ mental}{Age\ chronologique} \times 100\right)$. Si nous prenons un sujet âgé de 8 ans et que l'âge mental évalué par le test est de 10 ans, il aura un QID de 125, correspondant à une performance supérieure à la moyenne des enfants de son âge. Or, cette manière de calculer conduit à un contresens chez les adultes. Par exemple, si un sujet est âgé de 35 ans et que son âge mental évalué par le test est de 20 ans, il aura un QID de 57, correspondant à une performance très faible. Pourtant, il est absurde d'interpréter comme un retard mental le fait d'avoir des capacités intellectuelles dans la moyenne des adultes de 20 ans. En outre, il apparaît que le rythme de développement chez l'enfant est plus rapide (de l'ordre de quelques mois entre chaque phase) que chez l'adulte. Devenues inadaptées pour l'évaluation de l'intelligence chez les enfants et chez les adultes, les notions d'âge mental et de QI Développemental sont abandonnées. Wechsler innove alors une nouvelle manière de procéder. Le score total du sujet, obtenu en sommant les items réussis, est situé par rapport à la distribution des scores obtenus par d'autres sujets du même groupe d'âge. Désormais, le QI devient une note relative qui caractérise chaque individu « par le rang auquel sa performance permet de le classer dans son groupe d'âge » (Huteau & Lautrey, 2006, p. 123). La délimitation des groupes d'âge pour les enfants est plus rapprochée (de trois mois en trois mois).

À l'instar de Binet, Wechsler s'inscrit également dans une approche globale de l'intelligence qui n'exclut pas une vision plurielle. En effet, différents éléments du fonctionnement intellectuel contribuent au calcul d'une note globale de QI. Dès les premières éditions de ses tests, des indices sont proposés pour différentes composantes de l'intelligence (QI Verbal et QI Performance), et cela bien que l'interprétation repose principalement sur le QI Total (QIT). Pour Wechsler, la dichotomie Verbal-Performance est une lecture parmi d'autres pour regrouper les subtests. Il précise à ce sujet que cette distinction n'implique pas que les domaines du verbal et de la performance soient « *the only abilities involved in the tests. Nor does it presume that there are different kind of intelligence, e.g., verbal, manipulative, etc. It merely implies that these are different ways in which intelligence may manifest itself* » (Wechsler, 1958, p. 64). La conception de Wechsler repose sur l'hypothèse d'une intelligence qui est non seulement une entité globale – car elle caractérise le comportement d'un individu dans sa totalité –, mais également une entité spécifique – car elle se compose d'aptitudes ou d'éléments complexes qualitativement distincts les uns des autres (Wechsler, 2005b). Avec ses échelles, dont la popularité ne se dément pas, Wechsler a marqué l'histoire de l'évaluation de l'intelligence. Il a à cœur de souligner que la pratique de l'évaluation par un psychologue ne se réduit pas au testing. Le testing (ou la passation de tests) n'est qu'une partie dans le processus plus large et plus complexe d'une évaluation respectueuse de la personne. Pour Wechsler, la finalité des tests d'intelligence se révèle avant tout dans leur utilité clinique comme aide à la prise d'une décision au bénéfice du sujet, et non comme un moyen de mesure des aptitudes cognitives spécifiques.

> *What we measure with tests is not what tests measure – not information, not spatial perception, not reasoning ability. These are only means to an end. What intelligence tests measure, what we hope they measure, is something much more important: the capacity of an individual to understand the world about him and his resourcefulness to cope with its challenges.* (Wechsler, 1975, p. 139)

C'est cet accent sur l'utilité d'un test comme aide à la prise de décision pour la clinique qui guide les révisions successives des échelles d'intelligence de Wechsler et qui fait perdurer leur popularité.

L'approche psychométrique globale propose une pratique de l'évaluation de l'intelligence qui répond aux besoins de la société de catégoriser les individus. Les systèmes de classification tout d'abord proposés font référence à des catégories nosographiques de type médical, tandis que, de nos jours, les classifications cherchent

à décrire une performance en fonction de la distribution normale. Le Tableau 1 (p. 52) montre une comparaison historique de trois systèmes de classification. Wechsler apporte une contribution pérenne en segmentant des catégories en fonction de la déviation à la moyenne des scores de QI. Rappelons que la distribution des QI dans la population est supposée suivre une loi normale. Selon les fréquences statistiques d'individus sous la courbe normale connues, environs 68 % des individus ont des performances situées à+/- 1 écart type de la moyenne (soit entre 85 et 115 pour les notes de type QI de moyenne 100 et d'écart type 15).

Tableau 1

Comparaison historique entre trois systèmes de classification

Levine & Marks (1928)		Échelles de Wechsler (1944)			Flanagan, Ortiz, & Alfonson (2013)	
Étendue des QI	Classification	Étendue des QI	% de la population	Classification	Étendue des QI	Classification
0-24	Idiot (*idiot*)					
25-49	Imbécile (*imbecile*)					
50-74	Débile léger (*moron*)	≤ 69	2.2%	Très faible	< 70	Extrémité inférieure
75-84	État limite (*borderline*)	70-79	6.7%	Limite	70-84	En dessous de la moyenne (faiblesse normative)
85-94	Faible (*dull*)	80-89	16.1%	Moyen faible	85-89	Moyen faible
95-104	Moyen (*average*)	90-109	50.0%	Moyen	90-110	Moyen
105-114	Doué (*bright*)	110-119	16.1%	Moyen fort	111-115	Moyen fort
115-124	Très doué (*very bright*)	120-129	6.7%	Supérieur	116-129	En dessus de la moyenne (force normative)
125-149	Supérieur (*superior*)	≥ 130	2.2%	Très supérieur	> 130	Extrémité supérieure
150-174	Très supérieur (*very superior*)					
>175	Précoce (*precocious*)					

1.3.2. APPROCHE FACTORIALISTE DE L'INTELLIGENCE

En même temps que l'usage des tests d'intelligence se multiplie en Angleterre et aux États-Unis, la psychométrie se dote de deux méthodes statistiques – la corrélation et l'analyse factorielle – grâce aux contributions de Spearman et Thurstone. L'approche factorialiste de l'intelligence dont ces derniers peuvent être affiliés repose

sur des analyses statistiques qui permettent de mettre en évidence la structure factorielle des différences interindividuelles.

1.3.2.1. Spearman et le modèle bi-factoriel de l'intelligence

En 1904, le psychologue anglais Charles Edward Spearman (1863 – 1945) publie deux articles scientifiques majeurs à partir de données sur des épreuves cognitives administrées à un échantillon d'enfants. Le premier *The proof and measurement of association between two things* (Spearman, 1904b) présente des outils statistiques d'évaluation et de correction des degrés d'association entre deux grandeurs, tels que le *Pearson's product-moment correlation* et le *Spearman's rank correlation*. Dans le second article *"General Intelligence", objectively determined and measured* (Spearman, 1904a), il recourt à la méthode de l'analyse factorielle sur ses données dont les résultats montrent l'existence d'un facteur commun sous-tendant à toutes les activités mentales (voir Figure 1). En effet, Spearman constate des corrélations positives et non nulles entre différents tests mentaux (*positive manifold*). Pour Spearman, la raison en est l'influence d'une cause commune qui intervient dans les performances de tout test d'intelligence. Cette cause commune représente un facteur général d'intelligence, qu'il nomme le facteur *g*. Il existe également des facteurs spécifiques – facteurs *s* – qui sont propres à une tâche intellectuelle et qui, par conséquent, ont une influence plus limitée dans le fonctionnement global. Selon sa conception unifactorielle basée sur le facteur *g*, l'influence des facteurs spécifiques est marginale sur la performance cognitive. Sa théorie est d'abord perçue comme « monarchique » ou « unifocale », puisque c'est principalement le facteur *g* qui explique les différences de performances entre les individus (Martin, 1997).

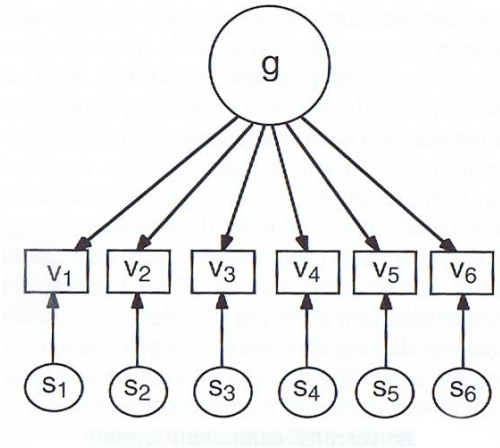


Figure 1. Modèle bi-factoriel de Spearman. Avec g = facteur *g* ; v# = variable/test ; s# = facteur spécifique. *Source* : Grégoire (2009, p. 56).

Cette approche unifactorielle de l'intelligence souffre de la critique et fait débats parmi les contemporains de Spearman. On peut relever les critiques des partisans des théories « anarchiques » ou « afocales » (*non-focal*) et ceux des théories « oligarchiques » ou « multifocales » (*multifocal*). Les premiers considèrent que les activités mentales ne suivent pas un principe structurant et organisé. Des auteurs, comme le psychologue américain Edward Lee Thorndike (1874 – 1949), réfutent en effet l'idée d'un facteur général. Pour lui, les faibles corrélations entre les tests mentaux (autour de .30) suggèrent, au contraire, l'absence d'unité dans les fonctions mentales. Les seconds défendent non pas l'existence d'un seul facteur général, mais « l'existence de plusieurs facultés générales indépendantes dont chacune gouverne un type d'activité mentale » (Martin, 1997, p. 153).

En réponse aux critiques, Spearman et ses alliés mènent de nouveaux travaux qui vont réaffirmer l'existence d'un facteur commun. Cependant, Spearman diminue un peu de l'importance de *g* pour la redonner aux facteurs spécifiques. Dans sa première conception théorique, l'accent ayant été trop mis sur la primauté de *g*, il rectifie le tir et fait évoluer sa théorie de l'intelligence dans un article intitulé *The theory of two factors* (Spearman, 1914). Il distingue toujours un facteur général (facteur *g*) et autant de facteurs spécifiques qu'il y a de tests cognitifs (facteurs *s*). En effet, dès le départ, il s'agit d'une théorie qui comprend l'existence de deux espèces de facteurs, mais la primauté de *g* a marginalisé la contribution des facteurs spécifiques. De monarchique, l'approche de Spearman devient « éclectique », puisqu'il accorde désormais autant d'importance au facteur d'intelligence générale qu'aux facteurs spécifiques. Son

modèle est alors défini de modèle bi-factorielle. Il décrit le facteur *g* comme « une sorte d'énergie mentale » dont disposent tous les individus, mais à des quantités différentes d'un individu à l'autre. Hart et Spearman (1912) illustrent la contribution des deux facteurs comme suit :

> Tout acte intellectuel mobilise à la fois un système précis de neurones (facteur spécifique) et l'énergie totale de l'ensemble du cortex (facteur général). (cité par Martin, 1977, p. 155)

D'autres travaux vont alimenter et continuer à faire évoluer les outils statistiques ainsi que la pensée de Spearman jusqu'à la publication de *The Abilities of Man* (1927). Dans ce livre, le psychologue anglais synthétise les connaissances sur les structures mentales et y expose la forme la plus aboutie de son modèle bi-factoriel de l'intelligence. Chapeautant le tout, il y a le facteur *g* qui influence le plus les différences interindividuelles dans les performances aux tests cognitifs. En dehors du facteur *g*, il reconnaît d'autres facteurs généraux de moindre importance et de nature parfois non cognitive : facteur *c* d'inertie générale introduit par Jones, facteur *w* de maîtrise de soi introduit par Webb et facteur *o* d'oscillation générale introduit par Spearman lui-même. D'une influence plus limitée que *g* mais tout aussi important dans l'observation des différences inter- et intra-individuelles, il y a les facteurs spécifiques. À côté des facteurs généraux et des facteurs spécifiques, les performances cognitives dépendent aussi de cinq facteurs de groupe (l'aptitude mécanique, l'aptitude logique, l'aptitude psychologique, l'aptitude mécanique et l'aptitude à apprécier la musique). Les facteurs de groupe sont conçus comme des résidus qui expliquent ce qui ne peut être expliqué par les facteurs généraux et les facteurs spécifiques.

1.3.2.2. Thurstone et les aptitudes mentales primaires

Dans le prolongement de la méthode d'analyse des facteurs de Spearman, Louis Leon Thurstone (1887 – 1955), une des figures emblématiques de la psychométrie américaine, développe une méthode d'analyse pluridimensionnelle (*Multiple Factor Analysis*; 1931). Bien que s'inspirant au départ des travaux de Spearman, Thurstone va largement s'en démarquer tant sur le plan théorique que méthodologique. Spearman vise à élaborer une théorie psychologique sur le fonctionnement de l'activité mentale (démarche inductive de vérification de l'existence d'un facteur commun), tandis que les interrogations de Thurstone sont davantage orientées vers le développement d'une méthode qui permet de trouver le nombre de facteurs décrivant les résultats aux tests

(estimation inductive sur les caractéristiques des données). Thurstone relègue au débat épistémologique le désaccord sur la réalité ou non des facteurs. La méthode est ainsi faite qu'à partir des données, l'analyse factorielle met en évidence des facteurs. Pour Thurstone, seule reste donc la question du nombre de facteurs.

Dans le développement de sa méthode, Thurstone teste des données provenant de 56 épreuves cognitives administrées à un échantillon d'étudiants. Les résultats des analyses factorielles l'amènent à élaborer une théorie multifactorielle de l'intelligence qui remet en question l'idée d'une faculté générale et globale telle que le facteur *g* (voir Figure 2, p. 56). Il met en lumière plusieurs facteurs de premier ordre, ou « Aptitudes Mentales Primaires » (PMA ; p. ex., aptitude numérique, compréhension verbale, visualisation spatiale), qui en tant que facteurs généraux indépendants, expliquent les différences de niveaux dans des domaines distincts de l'intelligence. Pour Thurstone, il n'y a plus « une grandeur prédominante, dont le statut est radicalement différent des autres . . . tous les facteurs ont le même statut *a priori* » (Martin, 1997, p. 215). Ainsi s'effectue le passage d'une conception unifactorielle à une conception multifactorielle de l'intelligence où « le nombre et la nature des facteurs ne sont pas postulés à l'avance, mais déterminés et interprétés *a posteriori* en fonction des caractéristiques des corrélations » (Martin, 1997, p. 216). Pour Thurstone, « les facteurs sont simplement les moyens de la description des données et n'ont pas de contenu psychologique *a priori* » (Martin, 1997, p. 215).

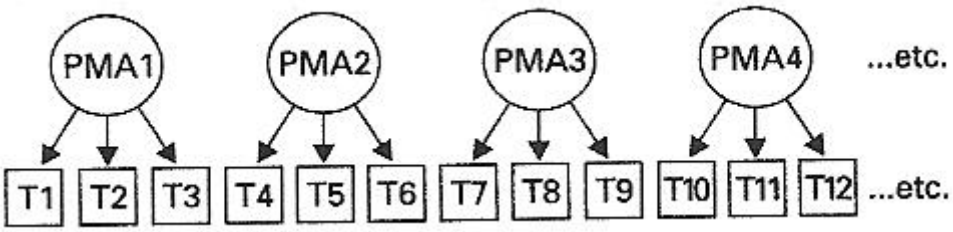


Figure 2. Modèle multifactorielle de Thurstone. Avec PMA# = aptitude mentale primaire ; T# = tâche/test. *Source* : Lautrey (2006).

En élargissant la méthode d'analyses multifactorielles à divers champs d'application et à diverses données, Thurstone en a fait une méthode générale ; on lui attribue souvent la paternité de la méthode d'analyse factorielle.

1.3.2.3. *g* psychologique vs *g* psychométrique

L'observation d'une corrélation positive modérée (mais bien réelle) entre différentes épreuves cognitives (*positive manifold*) est un résultat robuste en psychologie. De nombreuses recherches ont empiriquement établi ce phénomène avec des tâches cognitives de nature différente et auprès de populations variées. Concrètement, le *positive manifold* implique que les individus qui obtiennent un score élevé (ou faible) à une épreuve cognitive tendent à obtenir des scores élevés (ou faibles) sur d'autres épreuves cognitives. Si le phénomène est communément admis, en revanche, le débat sur la nature de ce facteur commun *g* découvert par Spearman n'est pas résolu. S'agit-il d'un *g* psychologique qui rend compte d'une intelligence générale ou d'un *g* psychométrique qui rend compte d'un artefact statistique ? Nous l'avons vu, pour Spearman, il s'agit d'un *g* psychologique. Pour ce dernier, il y a une cause commune (une intelligence générale) qui sous-tend toutes les performances cognitives et qui explique les différences interindividuelles sur les scores aux tests ainsi que les corrélations entre les tests cognitives. En fait, ce débat confronte deux conceptions : le modèle réflectif vs le modèle formatif (voir Figure 3, p. 57).

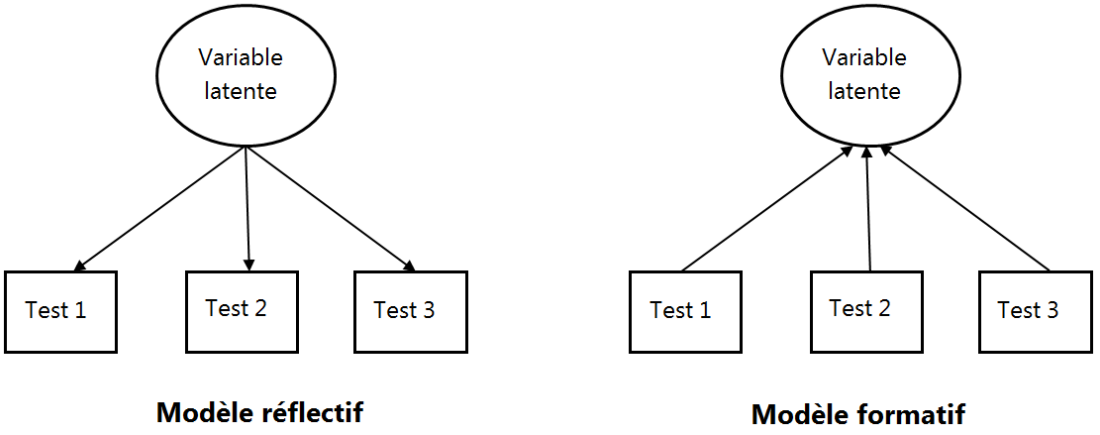


Figure 3. Modèle réflectif vs modèle formatif. À noter que par simplification, les intercorrélations entre les variables manifestes sont omises pour le modèle formatif.

Dans le modèle réflectif, la variable latente (p. ex., facteur *g*) explique (est à l'origine) des performances sur les variables manifestes qui l'évaluent. Ce sont les différences dans le niveau de *g* (variable latente) qui conduisent à des différences interindividuelles dans les performances aux tests cognitifs (variables manifestes). Dans cette conception, la variable latente est un construit psychologique qui existe, même s'il ne peut être observé qu'indirectement (p. ex., un score au test). Dans le modèle réflectif,

on essaie donc d'estimer un construit psychologique à l'aide d'indicateurs qui l'évaluent. Les analyses factorielles modélisent la conception de la variable latente réflective, « *in which the factor is a hypothesized entity that is posited to provide a putative explanation for the positive manifold* » (van der Maas, Kan, & Borsboom, 2014, pp. 12–13).

Dans le modèle formatif, les variables manifestes forment (définissent) la variable latente, qui n'est donc pas un facteur causal. Ce sont les changements dans les variables manifestes qui déterminent des variations dans la variable latente, qui s'apparente à un index. Dans cette conception, la variable latente n'existe pas. Elle est composée par les indicateurs. Ces derniers n'évaluent pas exactement la même chose, mais des composantes plus ou moins proches. Dans ce modèle, on essaie de résumer plusieurs attributs au moyen d'un index (variable latente). Un exemple parlant est une variable latente qui représente le milieu socio-économique. Cette variable latente peut être créée à partir de la sélection de plusieurs indicateurs tels que le revenu, le niveau d'études, le type d'habitation, le budget de vacances, etc. Les analyses en composantes principales modélisent la conception de la variable latente formative, qui *« does not say anything about the nature of the correlations in the positive manifold. . . . Importantly, in formative models, the nature of the constructed components is fixed by the subtests used to determine them: a different choice of subtests yield conceptually different components »* (van der Maas et al., 2014, p. 13). En effet, à l'inverse du modèle réflectif où les variables manifestes sont interchangeables sans changer la nature de la variable latente, dans le modèle formatif, les variables manifestes partagent non seulement une variance commune, mais chaque paire partage également quelque chose en plus et de différent avec les autres pairs. La modification ou la suppression d'un indicateur change la nature de la variable latente. Les variables manifestes ne sont donc pas interchangeables, car elles ne sont pas équivalentes dans le modèle formatif.

Dernièrement, le modèle du mutualisme (van der Maas et al., 2006) propose une explication au *positive manifold* alternative qui n'inclut pas le facteur *g* (voir Figure 4). Dans le modèle du mutualisme, « the correlations *between test scores are not explained through the dependence on a common latent variable, but as a result of reciprocal positive interactions between abilities and processes that play key roles in cognitive development, like memory, spatial ability, and language skills »* (van der Maas et al., 2014, p. 13).

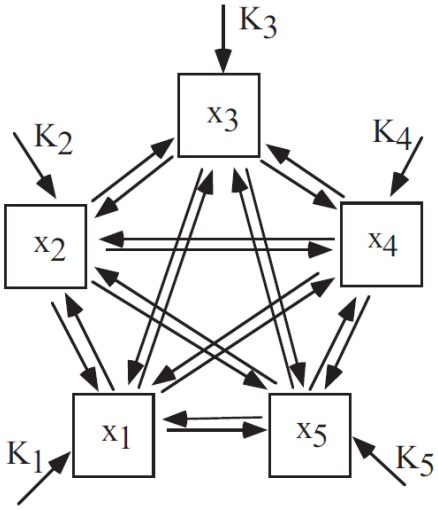


Figure 4. Modèle du mutualisme. Les symboles x représentent des processus cognitifs et les K représentent des ressources cognitives. *Source* : van der Maas et al., (2006, p. 843).

Dans ce modèle, les processus cognitifs élémentaires ne corrèlent pas ensemble dès le début. C'est durant le développement de l'individu que le *positive manifold* émerge comme résultante des interactions complexes et dynamiques entre les aptitudes et les processus en jeu dans la réalisation des tâches cognitives. Au cours de ces interactions, les aptitudes et les processus cognitifs s'influencent réciproquement de manière bénéfique et positive, c'est ce qui est entendu par mutualisme ou coopération. Selon nous, le modèle du mutualisme apporte une explication heuristique au phénomène du *positive manifold*.

1.3.3. APPROCHE HIÉRARCHIQUE DE L'INTELLIGENCE

Après la découverte du facteur *g* par Spearman, les visions se confrontent dans un bras de fer entre les tenants d'une théorie unifactorielle et les tenants d'une théorie multifactorielle de l'intelligence (voir Figure 5). Les contemporains de Spearman et de Thurstone s'interrogent sur la structure des différences individuelles dans les scores aux tests : « les individus se différencient-ils par une capacité générale, qui est à l'œuvre dans toutes les tâches intellectuelles, ou se différencient-ils relativement à des capacités spécialisées et indépendantes les unes des autres ? » (Huteau & Lautrey, 2003, p. 143). À partir des méthodes d'analyses multifactorielles de Thurstone, la mise en

évidence d'une multiplicité de facteurs mentaux indépendants de *g* met à mal l'édifice théorique de l'unicité du *g* de Spearman.

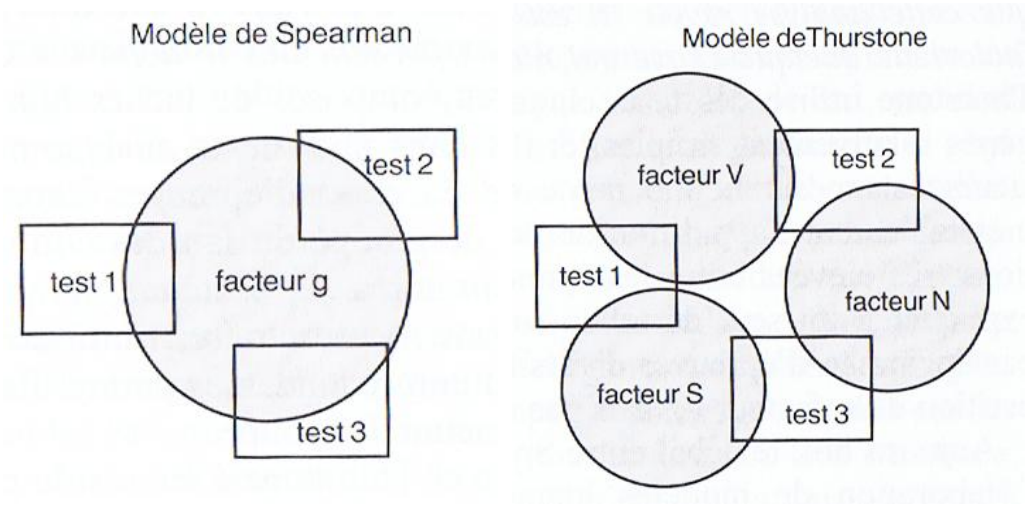


Figure 5. Modèle de l'évaluation du *g* de Spearman vs modèle de l'évaluation des facteurs de Thurstone. Avec facteur V = verbal ; facteur N = numérique ; facteur S = spatial. *Source* : Grégoire (2009, p. 59).

Comme les données empiriques tendent à montrer que le facteur *g* n'est pas suffisant pour rendre compte des intercorrélations entre les variables (différents tests d'intelligence ou subtests d'un test), peu à peu les théories de l'intelligence basculent vers une approche hiérarchique de l'intelligence qui, à la fois, intègre un facteur général similaire au facteur *g* de Spearman et des facteurs de groupe comme dans la théorie multifactorielle de Thurstone. Parmi les modèles hiérarchiques qui se sont développés à partir de la seconde moitié du 20^{e} siècle (p. ex., Johnson & Bouchard Jr., 2005a, 2005b; Vernon, 1950), notre propos se focalisera sur la présentation du modèle de Cattell-Horn-Carroll (modèle CHC). Depuis plusieurs années, le modèle CHC des aptitudes cognitives remporte un certain consensus dans la communauté scientifique en tant que modèle psychométrique de l'intelligence. Construit sur les apports de plus de 60 ans de recherche sur les analyses factorielles de tests d'intelligence, il est le modèle contemporain le plus validé empiriquement (Ackerman & Lohman, 2006; W. J. Schneider & McGrew, 2012). Ce modèle émerge de la réunification des travaux des psychologues Cattell et Horn d'une part, et Carroll d'autre part.

1.3.3.1. Cattell, Horn et le modèle Gf-Gc étendu

Vers 1940, un ancien élève de Spearman, Raymond Bernard Cattell (1905 – 1998) distingue deux facteurs de l'intelligence : l'Intelligence Fluide (Gf) et l'Intelligence Cristallisée (Gc). Le premier, influencé par le biologique et le neurologique, manifeste la capacité de raisonner de manière inductive et déductive pour résoudre des tâches nouvelles ou non familières, tandis que le second concerne la résolution de tâches où interviennent les connaissances engrangées dans l'expérience, l'éducation et la culture environnante. À partir des aptitudes de premier ordre, R. B. Cattell extrait deux facteurs de second ordre et propose une théorie Gf-Gc qui conceptualise les aptitudes cognitives de manière dichotomique (R. B. Cattell, 1941, 1943).

En 1965, dans la suite directe des travaux de R. B. Cattell, John Horn (1928 – 2006), élève de ce dernier, identifie d'autres facteurs qui rendent compte des activités intellectuelles humaines : le Traitement Visuel (Gv), la Mémoire de Travail (Gwm, anciennement Mémoire à court terme Gsm), la Récupération à Long Terme (Glr) et la Vitesse de Traitement (Gs). Plus tard, il rajoute le Traitement Auditif (Ga), la Connaissance Quantitative (Gq), et encore d'autres qui élargissent et clarifient le modèle initial de Cattell pour donner le modèle Gf-Gc étendu de Cattell-Horn (voir Figure 6, p. 61).

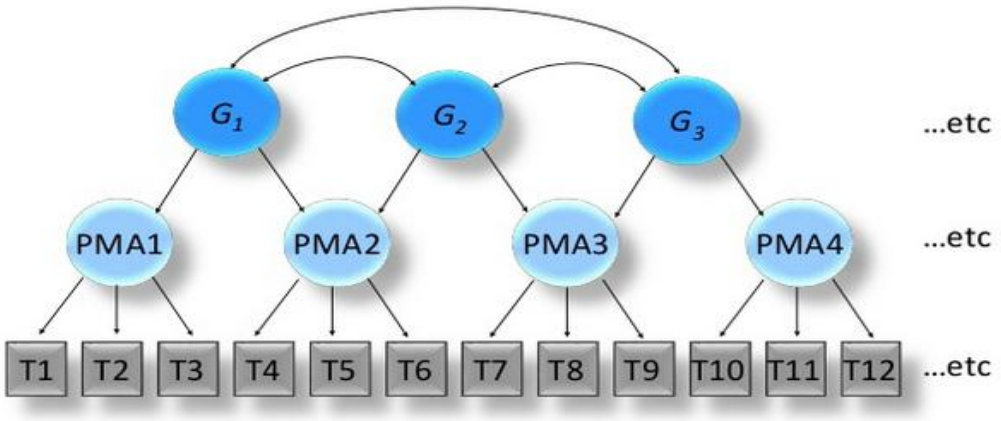


Figure 6. Modèle hiérarchique de Cattell-Horn. Avec T# = test ; PMA# = aptitude mentale primaire (facteur de 1er ordre) ; $G_{\#}$ = aptitude cognitive (facteur de 2e ordre). *Source* : présentation de McGrew (2014) pour *Institute for Applied Psychometrics* (IAP).

1.3.3.2. Carroll et le modèle en trois strates

Dans un ouvrage devenu un classique (Carroll, 1993), le psychologue américain, John Bissell Carroll (1916 – 2003) présente un modèle des aptitudes cognitives organisées sur trois niveaux hiérarchisés allant des aptitudes les plus restreintes à l'aptitude la plus étendue (voir Figure 7, p. 63). Le premier niveau (*stratum I*) comprend une septantaine de facteurs de premier ordre extraits de la matrice de corrélations entre les variables observées, classifiés sous l'appellation : aptitudes restreintes (*narrow abilities*). De faible étendue, les aptitudes restreintes représentent des « *greater specializations of abilities, often in quite specific ways that reflect the effects of experience and learning, or the adoption of particular strategies of performance* » (Carroll, 1993, p. 634). Le deuxième niveau (*stratum II*) inclut de nombreux facteurs de deuxième ordre extraits de la matrice de corrélations entre les facteurs de premier ordre, classifiés sous l'appellation : aptitudes étendues (*broad abilities*). D'un plus large spectre, les aptitudes étendues correspondent à des « *basic constitutional and long standing characteristics of individuals that can govern or influence a great variety of behaviors in a given domain* » (Carroll, 1993, p. 634). Au troisième niveau (*stratum III*), on trouve l'aptitude la plus étendue et générale, le facteur général *g,* qui est un facteur de troisième ordre extrait de la matrice de corrélations entre les facteurs de deuxième ordre. Ce dernier chapeaute les aptitudes étendues et restreintes. Dans la Figure 7 (p. 63), les intercorrélations entre les différentes aptitudes restreintes (*narrow abilities*) et entre les différentes aptitudes étendues ($G_{\#}$) ne sont pas représentées, cependant, elles sont non nulles et positives. Cela indique que les aptitudes restreintes et étendues sont, à un certain degré, interdépendantes les unes des autres.

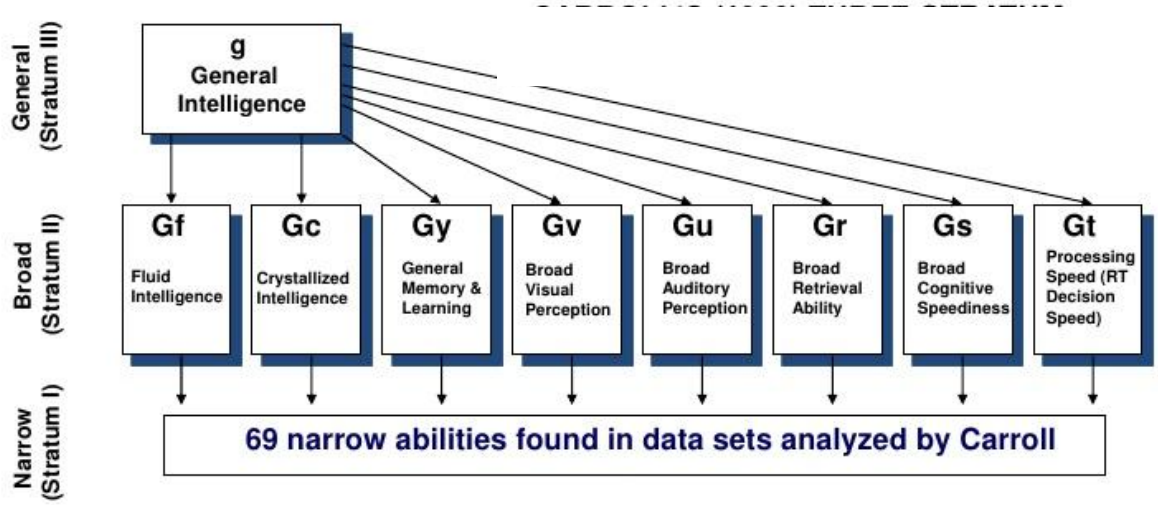


Figure 7. Modèle « Three-Stratum » de Carroll. *Source* : présentation de McGrew (2010) pour *The 2010 CNN conference in Fremantle*.

Les similarités sont nombreuses entre le modèle Gf-Gc étendu de Cattell-Horn et le modèle hiérarchique en trois strates de Carroll, mais on peut aussi relever des différences (voir Figure 8, p. 64). D'abord, le modèle de Carroll inclut le facteur *g*, tandis que *g* n'est pas présent dans le modèle de Cattell-Horn. Certaines aptitudes étendues dans un modèle sont, en revanche, des aptitudes restreintes dans l'autre modèle. Par exemple, la connaissance quantitative forme une aptitude étendue (Gq) dans le modèle Cattell-Horn, tandis que, dans le modèle de Carroll, elle est incluse dans les aptitudes restreintes de Gf. Néanmoins, dans l'ensemble, il y a une bonne correspondance entre les facteurs de groupe de ces deux modèles, conduisant McGrew (1997) à proposer leur fusion et, ainsi, faire émerger le modèle Cattel-Horn-Carroll des aptitudes cognitives (voir Figure 9, p. 65).

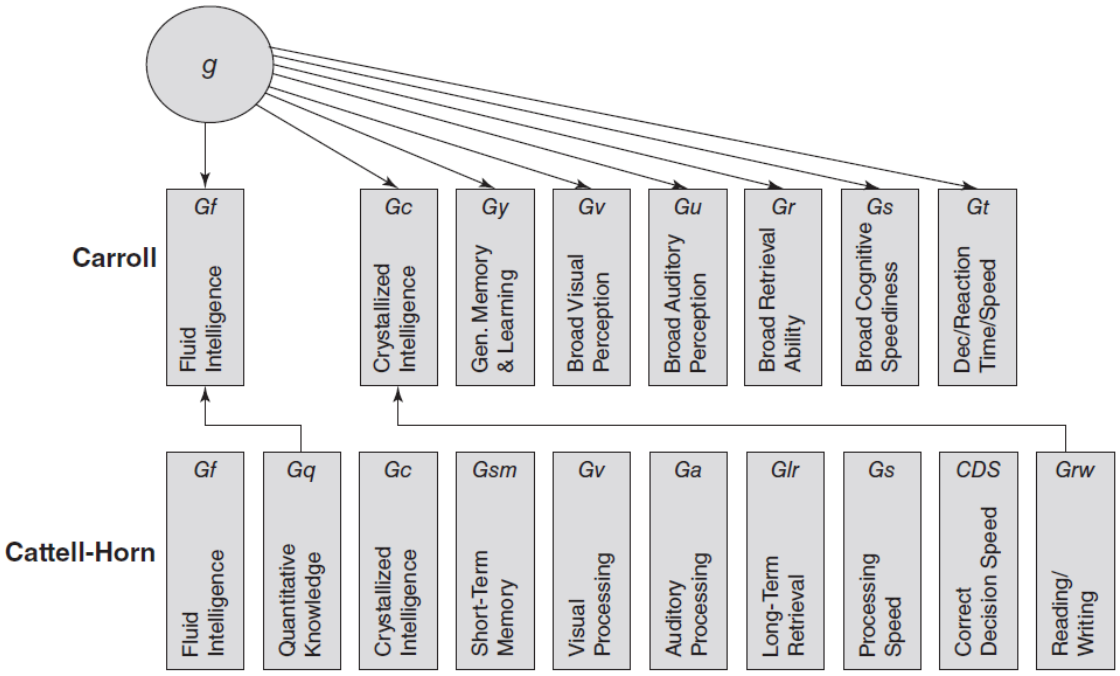


Figure 8. Comparaison des modèles de Cattell-Horn et de Carroll. *Source* : Flanagan et Dixon (2013).

1.3.3.3. Le modèle CHC des aptitudes cognitives

Le modèle de Cattell-Horn-Carroll des aptitudes cognitives (modèle CHC) ainsi formé par la mise en commun des modèles de Cattell-Horn et de Carroll est un modèle hiérarchique en trois strates (voir Figure 9, p. 65). Au sommet, nous retrouvons un facteur global et général qui chapeaute un certain nombre d'aptitudes cognitives étendues, qui, quant à elles, regroupent plusieurs aptitudes restreintes. La littérature sur le modèle CHC est abondante depuis plus de dix ans et le modèle s'enrichit de facteurs (*broad* ou *narrow*) nouvellement identifiés au fil des contributions (voir Keith & Reynolds, 2010; McGrew & Wendling, 2010; W. J. Schneider & McGrew, 2012, pour un développement approfondi). D'après les derniers travaux, on dénombre 16 aptitudes cognitives étendues et plus de 80 aptitudes cognitives restreintes, cependant, les tests cognitifs actuels ne sont pas en mesure de tous les évaluer (voir Figure 10, p. 66). Actuellement, 9 facteurs étendus et environ 40 facteurs restreints (rectangle rouge dans la Figure 10) peuvent être évalués à l'aide des tests psychologiques en vente (p. ex., les échelles de Wechsler, les échelles de Kaufman, le Woodcock-Johnson). Nous allons terminer le chapitre par la présentation de la 4ème édition de l'échelle d'intelligence de Wechsler pour enfants et adolescents sur lequel de nombreux travaux ont montré

l'adéquation d'une lecture selon le modèle CHC. De plus, cet instrument est au cœur de ce travail et sa mention revient régulièrement dans notre propos.

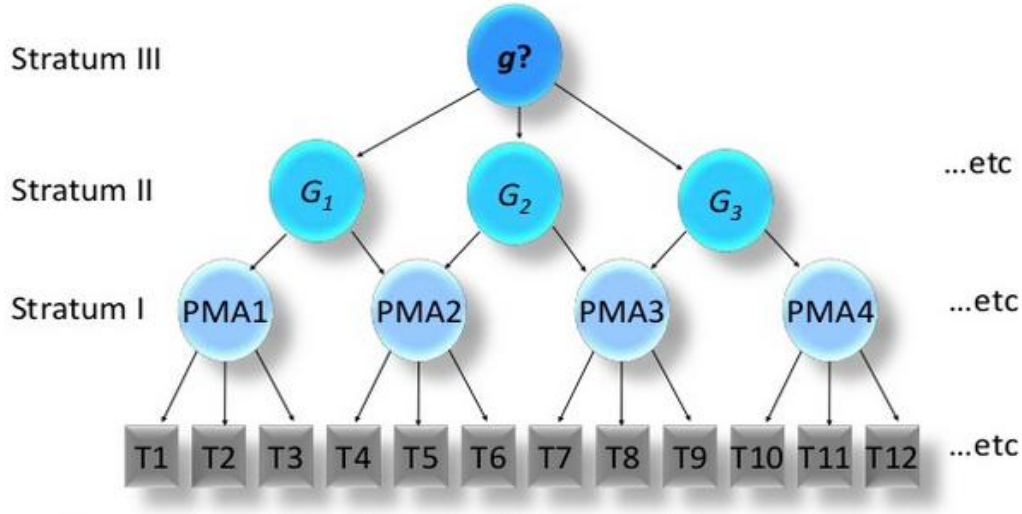


Figure 9. Modèle Cattell-Horn-Carroll (CHC). Avec T# = test ; PMA# = aptitude mentale primaire ; $G_{\#}$ = aptitude cognitive. *Source* : présentation de McGrew (2014) pour *Institute for Applied Psychometrics* (IAP).

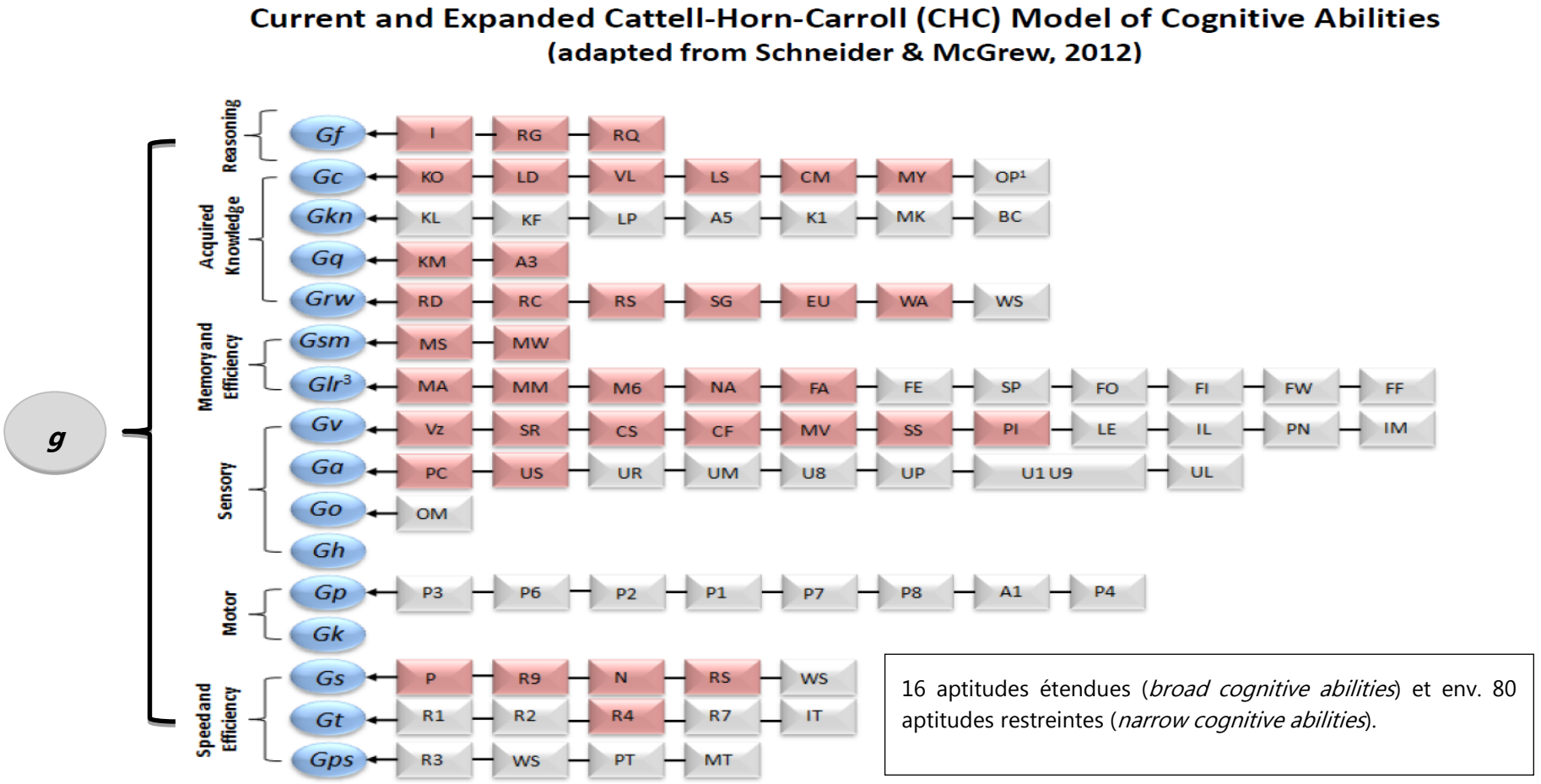


Figure 10. Modèle Cattell-Horn-Carroll (CHC). Avec des ronds bleus pour les aptitudes étendues et des rectangles pour les aptitudes restreintes.

1.4. ÉCHELLE D'INTELLIGENCE DE WECHSLER POUR ENFANTS ET ADOLESCENTS – WISC-IV

Dans le présent travail, notre intérêt s'est porté sur l'étude d'aspects psychométriques liés à une batterie d'évaluation de l'intelligence : la 4e édition de l'Échelle d'Intelligence de Wechsler pour Enfants et Adolescents (WISC-IV). Cette batterie de tests est fréquemment utilisée pour l'évaluation de l'intelligence chez les enfants et les adolescents. Pour certaines procédures administratives, les résultats à ce test peuvent être un critère pour un saut de classe ou une admission dans une école spécialisée ou dédiée aux enfants précoces.

Sur l'échantillon d'étalonnage de 1'103 enfants âgés de 6 à 16 ans 11 mois qui ont passé le WISC-IV courant 2004, des analyses factorielles confirmatoires (AFC) font ressortir la structure factorielle attendue en 4 facteurs (compréhension verbale, raisonnement perceptif, mémoire de travail et vitesse de traitement). Postulé à partir de la structure trouvée dans la version américaine, ce modèle en 4 domaines cognitifs est établi par les concepteurs de l'adaptation française comme étant celui qui montre le meilleur ajustement aux données de standardisation (Wechsler, 2005b). Relevons que leurs analyses factorielles confirmatoires se sont limitées à la comparaison de quelques modèles et qu'aucune analyse factorielle exploratoire (AFE) n'a été réalisée. En effet, les AFCs ont été réalisées pour comparer les modèles suivants : modèle à 1 facteur[6] vs modèle à 0 facteur commun, modèle à 0 facteur vs modèle à 2 facteurs[7], modèle à 1 facteur vs modèle à 2 facteurs, modèle à 0 facteur vs modèle à 3 facteurs[8], modèle à 1 facteur vs modèle à 3 facteurs, modèle à 0 facteur vs modèle à 4 facteurs et modèle à 1 facteur vs modèle à 4 facteurs.

Sur la base des résultats sur la structure factorielle du manuel du WISC-IV, l'interprétation courante du WISC-IV repose sur une note qui rend compte du niveau de fonctionnement intellectuel général, appelé le Quotient Intellectuel Total (QIT), ainsi que de quatre indices qui évaluent des domaines cognitifs plus spécifiques : (1) l'Indice

[6] 10 subtests sous-tendus par un facteur général.

[7] 3 subtests de compréhension verbale et 2 subtests de mémoire de travail sur un facteur, et 3 subtests de raisonnement perceptif et 2 subtests de vitesse de traitement sur un autre facteur.

[8] 3 subtests de compréhension verbale sur un premier facteur, 3 subtests de raisonnement perceptif sur un deuxième facteur, et 2 subtests de mémoire de travail et 2 subtests de vitesse de traitement sur un troisième facteur.

de Compréhension Verbale (ICV) ; (2) l'Indice de Raisonnement Perceptif (IRP) ; (3) l'Indice de Mémoire de Travail (IMT) ; et (4) L'Indice de Vitesse de Traitement (IVT).

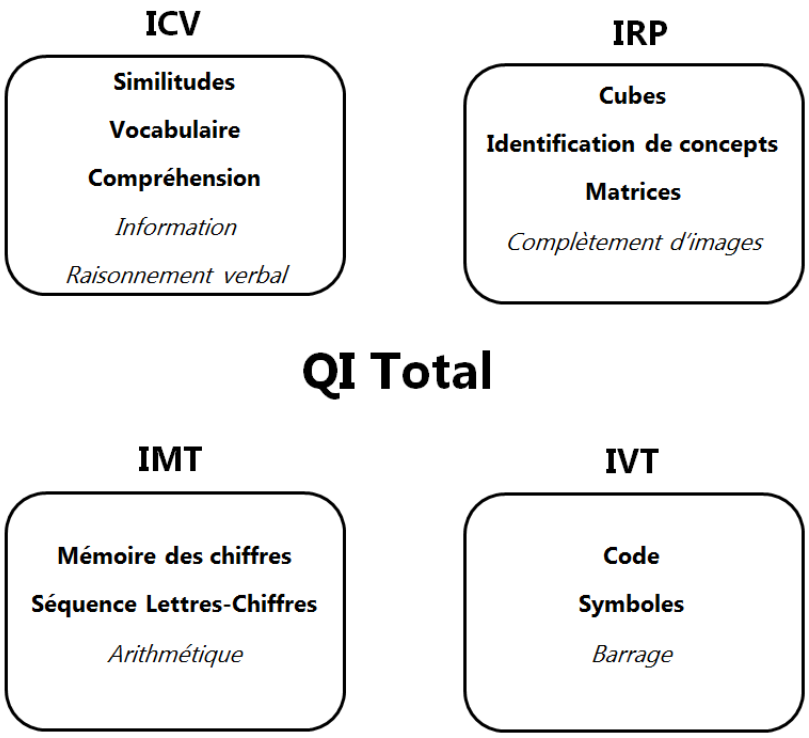


Figure 11. Structure du WISC-IV. Avec les subtests obligatoires en gras et les subtests optionnels en italique. Adapté de Wechsler (2005b, p. 6).

Quinze subtests – dont dix obligatoires et cinq optionnels – constituent la batterie (voir la section Méthode pour la description détaillée des subtests utilisés). Dans la Figure 11, les subtests sont regroupés sous l'indice dont ils permettent le calcul. Nous regroupons sous l'appellation « indices standards », les indices initialement proposés, ou repris ultérieurement par les concepteurs du WISC-IV (c.-à-d. l'Indice d'Aptitude Générale et l'Indice de Compétence Cognitive). Nous allons commencer par décrire les indices standards, puis les indices du modèle de Cattell-Horn-Carroll que peut calculer le WISC-IV.

1.4.1. INDICES STANDARDS DU WISC-IV

Le QI Total (QIT) est une note composite calculée à partir des performances aux quatre indices du WISC-IV. Il représente l'efficience intellectuelle générale d'un individu,

et peut être considéré comme un excellent indicateur du facteur *g*. Les notes de QIT sur le WISC-IV varient de 40 à 160 (soit une étendue de -4 à +4 écarts types de la moyenne de 100). Le QI Total est largement utilisé comme aide à la décision (p. ex., dispense d'âge, critère pour le retard mental ou le haut potentiel intellectuel). Les travaux sur le WISC-IV et ses précédentes éditions montrent en général que le QIT est un bon prédicteur de la réussite scolaire, académique et professionnelle (Deary et al., 2004; L. Gottfredson & Saklofske, 2009; Mayes & Calhoun, 2007; Sternberg, Grigorenko, & Bundy, 2001; Watkins, Lei, & Canivez, 2007).

Dans le prolongement de la discussion sur la nature du facteur *g*[9], on peut se questionner sur ce que traduit le score de QIT. Est-ce une variable réflective ou formative ? La question peut paraître purement théorique, or, la nature qu'on attribue au QIT a une implication pour l'interprétation de ce score. Si l'on postule l'existence d'un *g* psychologique, le QIT sera perçu comme une variable réflective qui traduit une intelligence générale. L'interprétation adoptera une lecture causale ; le niveau de QIT du sujet cause (explique) ses performances aux subtests du WISC-IV. Si l'on considère le QIT comme un index qui est construit par l'addition d'indicateurs, le QIT sera perçu comme une variable formative. Le QIT est interprété comme un index qui résume les indicateurs qui le forment. Notre position théorique est qu'il s'agit d'une variable formative. En effet, la modification ou la suppression d'un indicateur (c.-à-d. un des dix subtests obligatoires) changent la nature du QIT qui ne tient plus tout à fait compte des mêmes aptitudes cognitives. Les subtests du WISC-IV intercorrèlent entre eux, car ce sont toutes des épreuves cognitives. Cependant, chaque subtest contribue aussi à un des quatre domaines plus spécifiques du WISC-IV (compréhension verbale, mémoire de travail, etc.). Il n'y a donc pas une seule et même propriété mentale derrière chaque subtest ; ils ne sont pas interchangeables. Le QIT du WISC-IV n'est qu'une opérationnalisation de l'intelligence définie selon quatre domaines cognitifs. D'autres tests d'intelligence fournissent une évaluation du fonctionnement intellectuel selon d'autres index (p. ex., l'Indice Fluide Cristallisée pour la batterie du KABC-II de Kaufman et Kaufman).

L'Indice de Compréhension Verbale (ICV) se réfère à la profondeur et à l'étendue des connaissances verbales. Selon le manuel d'interprétation, l'ICV est une mesure de la formation de concepts verbaux, du raisonnement verbal et des connaissances acquises dans l'environnement (Wechsler, 2005b). Les épreuves de l'ICV mettent en jeu les capacités verbales de l'individu. Il est aussi un bon prédicteur de la

[9] Voir section 1.3.2.3, p. 53.

réussite scolaire et des apprentissages (Grégoire, 2009). Comme il est très sensible aux opportunités d'apprentissage, on observe des différences dans le niveau du score moyen à l'ICV en faveur des sujets provenant de milieux socioéconomiques élevés (Grégoire, 2009).

L'Indice de Raisonnement Perceptif (IRP) s'entend comme la capacité de manipuler du matériel visuel, et de raisonner dessus. Selon le manuel d'interprétation, l'IRP est une mesure du raisonnement fluide, du traitement spatial et de l'intégration visuomotrice (Wechsler, 2005b). Les épreuves de l'IRP sont de nature moins homogène en comparaison des épreuves qui contribuent aux autres indices. En effet, les subtests de l'IRP saturent modérément sur le facteur commun qui les sous-tend (autour de .50, Grégoire, 2009, p. 165, Tableau 31). De plus, la part de variance spécifique de chaque subtest est importante (autour de 50 % ; Grégoire, 2009, p. 170, Tableau 32), tandis que la part de variance partagée (la propriété mentale qu'évalue le test et d'autres tests) est autour de 30 %. Cela suggère que la moitié des différences interindividuelles dans les scores aux subtests de l'IRP est expliquée par les différences sur ce qui est spécifique à chaque subtest, et non sur ce qui est commun. On peut relever que l'appellation de cet indice ne se réfère pas à une aptitude cognitive identifiable. Il s'agit plutôt d'un indice contruit avec des subtests qui évaluent à la fois une composante visuo-spatiale et une composante d'intelligence fluide.

L'Indice de Mémoire de Travail (IMT) renvoie à la capacité d'encoder, de maintenir temporairement des informations en mémoire et de manipuler du matériel en mémoire immédiate. Selon le manuel d'interprétation, les tâches de mémoire de travail impliquent l'attention, la concentration, le contrôle mental et le raisonnement (Wechsler, 2005b). Les épreuves de l'IMT évaluent la mémoire à court terme sur sa forme verbale ; l'enfant doit énoncer verbalement une séquence de chiffres ou/et lettres comme modalité de réponse aux items. La mémorisation sur des stimuli auditifs non verbaux, visuels ou multisensoriels n'est pas évaluée. Les performances aux subtests de l'IMT sont sensibles aux connaissances des symboles numériques et de l'alphabet ainsi qu'aux troubles de l'attention (Grégoire, 2007b).

L'Indice de Vitesse de Traitement (IVT) est une note composite calculée à partir des performances aux subtests Code et Symboles. La vitesse de traitement dans le WISC-IV s'entend comme la capacité à inspecter rapidement et correctement des informations visuelles simples, et à réaliser rapidement des tâches cognitives simples. L'IVT évalue donc la rapidité cognitive de « bas niveau ». Selon le manuel d'interprétation, l'IVT fournit également une mesure de la vitesse de discrimination

visuelle, de mémoire visuelle à court terme, d'attention et de coordination visuomotrice (Wechsler, 2005b). Les épreuves de l'IVT présentent des stimuli visuels asémantiques et demandent au sujet de fournir une réponse motrice manuelle. Les performances aux subtests sont sensibles à l'habileté graphomotrice, la maîtrise de l'écriture et des troubles de l'attention (Grégoire, 2007b). Contrairement à l'ICV, l'IVT est peu influencé par l'origine socioéconomique des individus (Grégoire, 2009).

L'Indice d'Aptitude Générale (IAG) est une note composite calculée à partir des performances aux subtests obligatoires de l'ICV et de l'IRP. Il est proposé par des auteurs (Prifitera, Saklofske, & Weiss, 1998; Raiford, Weiss, Rolfhus, & Coalson, 2005) comme alternative au QI Total. Composé de deux indices – l'ICV et l'IRP – fortement intercorrélés et qui saturent de manière importante sur le facteur *g* (Lecerf et al., 2011), l'IAG présente en outre une très forte corrélation avec le QI Total : $r = .96$ pour la version américaine (Saklofske, Prifitera, Weiss, Rolfhus, & Zhu, 2005, p. 43) et $r = .92$ pour l'échantillon d'étalonnage français (Grégoire, 2009, p. 185). À la différence du QI Total, l'IAG écarte l'influence des processus cognitifs de base liés à la mémoire de travail et à la vitesse de traitement. Comme l'IAG et le QIT ne reflètent pas exactement les mêmes facettes de l'intelligence, ils ne doivent pas être confondus. Dans certaines populations cliniques, l'IAG fournit une meilleure estimation du niveau cognitif général, par exemple dans les profils d'enfants à Haut Potentiel Intellectuel (HPI) ou les enfants ayant un diagnostic de troubles de l'attention (Grégoire, 2009; Saklofske, Weiss, Raiford, & Prifitera, 2006). En effet, dans les épreuves de l'IMT (Mémoire des chiffres, Séquence Lettres-chiffres, Arithmétique) et surtout dans celles de l'IVT (Code, Symboles, Barrage), les performances des enfants HPI sont généralement dans la moyenne comparativement à celles d'enfants tout-venant du même groupe d'âge (Flanagan & Kaufman, 2009; Sparrow, Pfeiffer, & Newman, 2005; Wechsler, 2005b). Cette observation a déjà été relevée dans des études réalisées sur le WISC-III (Watkins, Greenawalt, & Marcell, 2002; Wechsler, 1991). Une des raisons peut être liée à la validité des tâches ; les subtests de l'IVT (surtout, Code) évaluent moins la vitesse de traitement « mentale » d'un individu que sa vitesse graphomotrice et sa maîtrise de l'écriture (Grégoire, 2007b). Comme ces deux dernières compétences sont davantage liées à l'âge de développement de l'individu qu'à ses aptitudes cognitives, les enfants HPI ne se montrent pas forcément supérieurs aux enfants tout-venant du même âge. En effet, les habiletés graphomotrices se développent avec l'expérience de la pratique de l'écrit à l'école notamment. Comme dans le WISC-IV, chaque indice contribue pour un quart dans la note du QIT, ce dernier peut se voir affaibli chez les enfants HPI à

cause de leurs performances moyennes à l'IMT ou/et à l'IVT. Dans la situation d'un important écart de performances entre d'un côté l'ICV et l'IRP et de l'autre, l'IMT et l'IVT, l'interprétation de l'IAG est à privilégier à la place de celle du QIT. Néanmoins pour une représentation du fonctionnement général, l'IAG est utilisé avec son pendant l'Indice de Compétence Cognitive.

L'Indice de Compétence Cognitive (ICC) est une note composite calculée à partir des performances aux subtests obligatoires de l'IMT et de l'IVT. Il s'agit d'un indice optionnel utilisé en complément à l'IAG pour avoir une image globale du fonctionnement cognitif de l'individu. Tandis que l'IAG rend compte des connaissances acquises et des habiletés de raisonnement, l'ICC reflète, quant à lui, la compétence et l'efficacité dans le traitement cognitif (Bremner, McTaggart, Saklofske, & Janzen, 2011).

1.4.2. INDICES CHC DU WISC-IV

En abandonnant sous l'influence de récentes théories de l'intelligence la dichotomie QI verbal et QI performance au profit d'une structure hiérarchique et multifactorielle, le WISC-IV actualise les fondements théoriques et compte de nouveaux subtests pour améliorer l'évaluation de l'intelligence. À la différence des batteries du KABC-II ou de la Woodcock–Johnson-IV qui cherchent à opérationnaliser un modèle théorique (le modèle CHC), le WISC-IV ne se réfère pas à une théorie de l'intelligence particulière à partir de laquelle les concepteurs auraient construit des subtests pour en évaluer les aptitudes cognitives. Pour le WISC-IV, les concepteurs sont partis d'une définition de l'intelligence qui comprend des composantes verbales, de raisonnement, de mémoire de travail et de vitesse de traitement et ont rassemblé des tests qui évaluent ces composantes. Les résultats des analyses factorielles confirmatoires sur les subtests sélectionnés ont étayé une structure en quatre facteurs. Cependant, d'autres études testent d'autres modèles et leurs résultats montrent l'adéquation du modèle CHC avec les données du WISC-IV (Keith et al., 2006; Lecerf, Rossier, Favez, Reverte, & Coleaux, 2010). Le modèle des aptitudes cognitives de Cattell-Horn-Carroll est empiriquement soutenu par les recherches menées au cours des 60 dernières années. Le modèle décrit une taxonomie exhaustive des facettes de l'intelligence. Les batteries de tests cognitifs actuelles ne permettent pas d'évaluer toutes les aptitudes décrites dans le modèle. Étant un modèle contemporain qui rallie un large consensus, il est particulièrement intéressant de pouvoir appliquer la lecture CHC pour l'interprétation des scores du WISC-IV. Sur les données françaises de l'échantillon de standardisation, il ressort un modèle en 6 facteurs, tandis que sur les données suisses-romandes du

précédent FNS *« Analysis of the French WISC-IV structure according to Cattell-Horn-Carroll (CHC) narrow ability classification »,* 5 facteurs CHC sont mis en évidence. Nous allons succinctement les présenter à la suite (voir Flanagan & Dixon, 2013; McGrew, 2005; J. H. Newton & McGrew, 2010, pour une description détaillée).

L'Intelligence cristallisée (Gc) évalue la capacité à faire appel à des connaissances acquises au travers de l'expérience, de l'éducation, de la culture et des interactions avec l'environnement. Comme Gc repose sur le langage de base déclaratif (savoir quoi) et procédural (savoir comment) acquis pendant les expériences éducatives – formelles ou informelles – de la vie quotidienne, les tâches verbales saturent fortement sur ce facteur. Dans le WISC-IV, on associe les aptitudes spécifiques de *connaissances lexicales* (VL) aux subtests Similitudes, Vocabulaire et Raisonnement verbal et de *connaissances verbales générales* (KO) aux subtests Compréhension et Information. L'aptitude restreinte VL évalue l'étendue du vocabulaire et de la compréhension du sens des mots. L'aptitude restreinte KO évalue l'étendue des connaissances générales.

L'Intelligence fluide (Gf) est déterminée par l'utilisation d'opérations mentales délibérées et contrôlées pour résoudre de nouveaux problèmes. Gf évalue les capacités à raisonner de manière inductive et déductive, à former des concepts, à s'adapter et à résoudre des problèmes nouveaux. Dans le WISC-IV, on associe l'aptitude restreinte d'*induction* (I) aux subtests Identification de concepts et Matrices. L'aptitude restreinte I évalue la capacité à découvrir des règles sous-jacentes.

La Mémoire de travail (Gwm) évalue la capacité à maintenir une information en conscience immédiate et à rappeler celle-ci dans un court délai (quelques secondes). Les informations qu'on maintient et qu'on traite pour les utiliser après un court laps de temps sont sensibles à des interférences au moment de l'encodage (Keppel & Underwood, 1962). Dans le WISC-IV, on associe les aptitudes restreintes d'*empan mnésique* (MS) au subtest Mémoire des chiffres et de *mémoire de travail* (MW) aux subtests Séquence Lettres-Chiffres et Arithmétique. L'aptitude restreinte MS évalue l'habileté à être alerte, à emmagasiner et à se rappeler une série d'éléments aléatoirement liés (lettres, chiffres) après quelques secondes (dans un certain ordre). L'aptitude restreinte MW évalue l'habileté à effectuer des opérations cognitives sur de l'information en mémoire à court terme.

La Vitesse de traitement (Gs) mesure la capacité à comparer un simple stimulus visuel et à le scanner rapidement, à réaliser des tâches simples rapidement et efficacement, à maintenir son attention et sa concentration. Dans le WISC-IV, on

associe les aptitudes restreintes de *vitesse perceptive* (P) aux subtests Symboles et Barrages et de *rapidité de réponse au test* (R9) au subtest Code. L'aptitude restreinte P évalue sous la pression du temps l'habileté à distinguer efficacement des symboles visuels similaires (ou différents) ou à repérer des cibles placées côte à côté ou dans des champs visuels séparés. L'aptitude restreinte R9 évalue l'habileté à exécuter rapidement des tâches relativement simples ou qui impliquent une prise de décision rapide.

Le Traitement visuel (Gv) évalue la capacité à percevoir, analyser, synthétiser, transformer, manipuler mentalement des stimuli visuels, à les tourner et à les déplacer dans des états variés pour résoudre un problème. Dans le WISC-IV, on associe les aptitudes restreintes de *visualisation* (VZ) au subtest Cubes et de *flexibilité de fermeture* (CF) au subtest Complètement d'images. L'aptitude restreinte VZ évalue l'habileté à manipuler mentalement des objets ou des stimuli visuels en faisant des rotations en deux ou trois dimensions. L'aptitude restreinte CF évalue la capacité à identifier une image visuelle imbriquée dans une figure visuelle complexe.

2. CONSIDÉRATIONS PSYCHOMÉTRIQUES DANS L'ÉVALUATION PSYCHOLOGIQUE

Dans le précédent chapitre, nous avons posé un cadre général et élargi à notre travail. L'évaluation de l'intelligence est une pratique qui s'inscrit dans le contexte historique et culturel d'une société. Il est important de contextualiser la pratique des tests psychologiques avant de développer et approfondir des concepts en lien plus direct avec nos résultats de recherche dans ce chapitre et le suivant. Ce deuxième chapitre s'ouvre sur la présentation des deux modèles de mesure auxquels nous nous référons pour l'analyse de nos données de recherche : la Théorie Classique des Tests (voir section 2.1.1) et les Modèles de Réponse à l'Item (voir section 2.1.2). La troisième grande théorie psychométrique – la Théorie de la Généralisabilité – ne sera pas développée dans ce manuscrit. À l'heure actuelle, la majorité des tests sont conçus dans le cadre de la théorie classique des tests (TCT). Ce modèle de mesure a l'avantage de s'appuyer sur des calculs statistiques de base. Plus accessibles grâce au développement des logiciels de modélisation, les modèles de réponse à l'item (MRI) fournissent une approche intéressante dans la détection de biais liés aux items d'un test. Dans les sections consacrées à ces deux théories, nous ne prétendons pas à l'exhaustivité. Nous visons à poser une vue d'ensemble des principaux fondements théoriques ; les équations ou formules mathématiques ne seront pas démontrées. Pour cela, veuillez vous référer aux ouvrages de Allen et Yen (1979) et, Lord et Novick (1968) pour la TCT et à l'ouvrage de Bertrand et Blais (2004) pour les MRI. Il s'ensuit plusieurs sections sur les considérations psychométriques pour l'utilisation des tests (homogénéité, sensibilité, etc.). Nous rappelons des concepts qui peuvent être familiers afin de s'accorder sur leur définition. La validité (voir section 2.5) peut faire l'objet d'une thèse entière, toutefois, nous l'introduisons pour la mettre en lien avec la préoccupation d'équité dans l'évaluation psychologique (voir section 2.6). À travers plusieurs considérations psychométriques, ce chapitre a pour objectif de mieux comprendre ce que traduit un score de test, comment on peut l'interpréter et dans quelles limites.

2.1. MODÈLE DE MESURE EN PSYCHOMÉTRIE

Les instruments d'évaluation – tels les tests psychologiques – sont construits dans le cadre conceptuel d'un modèle de mesure qui définit les caractéristiques de la

mesure recueillie par l'instrument. Le terme modèle a une double origine sur le plan étymologique. Une origine issue d'un mot latin *modulus* qui signifie mesure et, une autre issue d'un mot italien *modello*, qui signifie modèle. Historiquement, le terme est d'abord employé dans le sens d'une maquette ou d'un prototype qui sert de reconstitution miniature d'une entité plus grande (p. ex., un globe terrestre) avant de signifier également « ce qui est donné pour servir d'exemple, d'objet de référence ou d'imitation ». Dans un sens plus mathématique, le modèle renvoie à une représentation simplifiée d'un phénomène complexe pour mieux en rendre compte. Un modèle cherche donc à s'ajuster à la réalité, qu'il imite pour la décrire et l'expliquer. Tout en se voulant fidèle à la réalité, le modèle demeure néanmoins une représentation simplifiée de celle-ci. Il s'agit donc d'avoir conscience des limites de son application et des limites de la portée des interprétations qu'on peut faire à partir d'un modèle.

Trois grands modèles de mesure sont développés en psychométrie : la théorie classique des tests, la théorie de réponse à l'item et la théorie de la généralisabilité. La théorie classique des tests repose sur une équation de base qui stipule que le score observé à un test résulte de l'addition du score vrai de l'individu et de l'erreur de mesure. La théorie de réponse à l'item repose sur deux postulats de base :

> *(a) The performance of an examinee on a test item can be predicted (or explained) by a set of factors called traits, latent traits, or abilities; and (b) the relation between examinees' item performance and the set of traits underlying item performance can be described by a monotonically increasing function called an item characteristic function or item characteristic curve (ICC).* (Hambleton, Swaminathan, & Rogers, 1991, p. 7)

Enfin, la théorie de la généralisabilité est considérée comme une extension de la théorie classique des tests. Au moyen d'un plan expérimental établi, cette théorie psychométrique cherche à différencier les sources d'erreur de mesure (sont-elles dues à l'individu ? aux cotateurs ? aux items ? au moment de la passation ? etc.). En théorie classique des tests, on ne peut pas différencier les sources d'erreur qui composent l'erreur de mesure. Les objectifs de la théorie de la généralisabilité sont donc de déterminer quelles sources d'erreur sont associées à une situation de mesure donnée et quelle source d'erreur est la plus importante afin de pouvoir éventuellement la contrôler. Cette théorie ne manque pas d'intérêt, néanmoins ses contraintes (p. ex., nécessité d'un nombre important de sujets pour constituer les sous-échantillons des diverses conditions du plan expérimental) limitent sa popularité. Nous ne la

développerons pas davantage dans notre travail. En revanche, la part belle sera faite à la théorie classique des tests et à la théorie de réponse à l'item.

2.1.1. THÉORIE CLASSIQUE DES TESTS

La Théorie Classique des Tests (TCT) trouve ses origines dans les travaux de Spearman sur les coefficients de corrélation (Spearman, 1907). Dans les prémices de sa formalisation, elle porte le nom de « modèle de score vrai avec des postulats faibles » (*weak true-score model*) par opposition aux « modèles de score vrai avec des postulats forts » (*strong true-score model*) tels que le modèle binomial ou le modèle de Poisson. Avec des postulats peu contraignants comme son nom l'indique, la théorie du score vrai avec des postulats faibles offre un cadre théorique simple et applicable dans des situations variées, dont notamment le développement des tests psychologiques. De 1950 à 1980, plusieurs théoriciens des tests publient des ouvrages de référence en psychométrie dans lesquels ils posent les contours de la forme actuelle de la théorie du score vrai avec des postulats faibles (Allen & Yen, 1979; Gulliksen, 2013; Lord & Novick, 1968; Magnusson, 1967). Avec l'émergence des modèles de réponse à l'item, un nouveau paradigme de mesure pour les tests cohabite avec celui du score vrai, qui est alors renommé la théorie classique des scores/des tests par Lord et Novick (1968). À partir de maintenant, nous pourrons l'appeler simplement la théorie classique ou par son abréviation TCT.

2.1.1.1. Le modèle : X = V + E

La théorie classique s'est constituée à partir de l'équation suivante :

$$X = V + E \qquad (1)$$

Où X est le score observé d'un individu à un test, V est le score vrai de l'individu au test et E est l'erreur de mesure. Cette équation de base signifie que le score observé d'un individu à un test donné est déterminé par deux composantes additives : son score vrai sur ce qu'évalue le test ainsi que l'erreur qui entache toute mesure. Ces deux composantes ne sont pas directement observables, ce sont des entités théoriques qu'on cherche à inférer. Le score observé de l'individu peut varier d'une répétition à l'autre du même test. Il est la réalisation d'une variable aléatoire. Dans le sens utilisé

par la théorie des probabilités, une variable aléatoire s'entend comme l'ensemble des résultats possibles d'une expérience aléatoire. Par exemple, avant l'expérience aléatoire, le QI d'une personne prise au hasard est compris entre 40 et 160 et X est une variable aléatoire qui résume cette étendue de valeurs possibles. Après l'expérience, c'est-à-dire une fois le résultat observé au test, une valeur unique est obtenue qui n'a plus rien d'aléatoire. Le score observé X désigne alors une réalisation de la variable aléatoire.

Le score vrai représente la moyenne des scores obtenus lorsque le même test est administré un très grand nombre de fois au même individu. Il s'agit d'une constante pour un individu et pour un test particulier. Toutefois, sa valeur n'est jamais connue. La différence entre le score observé et le score vrai donne l'erreur de mesure ($X - V = E$).

La valeur de l'erreur de mesure est également inconnue. Il s'agit d'une variable aléatoire[10] qui fluctue d'une répétition à l'autre du test. Elle est constituée de deux composantes additives : l'erreur de mesure systématique (es) et l'erreur de mesure aléatoire (ea). En faisant apparaître les deux types d'erreur, l'équation (1) peut être développée comme suit :

$$X = V + ea + es \qquad (2)$$

La composante systémique de l'erreur (es) affectera les mesures répétées de façon constante et prévisible sur tous les individus d'un échantillon, tandis que la composante aléatoire de l'erreur (ea) varie de façon imprévisible d'une mesure à l'autre et d'un individu à l'autre. Dans la TCT, on suppose que l'erreur aléatoire n'est corrélée ni avec le score observé ni avec le score vrai. Ainsi, sur une infinité de mesures répétées pour un même individu, l'erreur aléatoire tend vers zéro. À la différence de l'erreur aléatoire, l'erreur systématique n'est pas variable d'un individu à l'autre et sa résultante est non nulle. L'erreur systématique est corrélée avec le score observé et avec le score vrai. Plus il y a d'erreurs de mesure systématiques, plus le score observé s'éloigne de l'estimation du score vrai. Comme son nom l'indique, l'erreur systématique se présente comme l'influence d'une contamination qui augmente ou diminue systématiquement le score au test pour tous les individus d'un échantillon spécifique (Haladyna & Downing, 2005).

On peut distinguer deux types d'erreur systématique. Le premier type est une erreur constante qui touche tous les individus d'un échantillon spécifique en surévaluant (ou sous-évaluant) systématiquement leur score vrai. La source de l'erreur

[10] Dans le sens utilisé dans la théorie des probabilités.

est externe aux sujets. Par exemple, si un examinateur est plus sévère que ses confrères, les sujets dont la performance est évaluée par cet examinateur auront des scores plus faibles qu'ils ne le méritent. Un autre exemple serait d'administrer deux versions d'un même test à chaque moitié d'un échantillon et que l'une des versions ait des items globalement plus difficiles que l'autre version. Le second type est une erreur qui surévalue (ou sous-évalue) systématiquement les scores des individus d'un échantillon à cause d'une différence sur un attribut psychologique qui est mobilisé par le test, mais sur lequel l'interprétation au test ne porte pas à proprement parlé (variance non pertinente). La source d'erreur est interne aux sujets. Messick (1989) donne l'exemple du format d'items en lien avec l'habileté en lecture. On évalue des sujets sur leur acquis en science avec un travail écrit où il y a un article scientifique à analyser. Des différences de notes entre sujets peuvent s'expliquer par une meilleure compréhension du texte plus le sujet a une bonne habileté en lecture. Dans les tests de rendement (*achievement test*) en particulier, les aptitudes verbales (c.-à-d. à la lecture, à l'écrit, au parlé ou à l'écoute) sont souvent sollicitées même si les tests ne portent pas sur ces habiletés. Dans leur étude avec une population de 946 écoliers américains présentant différents niveaux de fluence en anglais, Abedi, Lord, Hofstetter et Baker (2005) font passer des versions d'un test de mathématiques avec plus ou moins de demandes en compréhension verbale. Les résultats montrent notamment que les compétences en vocabulaire influencent grandement les performances au test passé. En effet, les performances sur le test de mathématiques s'améliorent pour les écoliers moins fluents en anglais pour qui on simplifie le vocabulaire utilisé dans l'épreuve, on permet d'utiliser un dictionnaire ou on donne plus de temps. Parmi les nombreuses sources d'erreur systématique, on identifie aussi la motivation, l'anxiété, la fatigue ou la tricherie. Par exemple, il y aurait une erreur systématique si le nombre d'items omis au test augmente de façon constante et prévisible avec la motivation ou la fatigue des sujets. Ayant un effet constant sur chaque mesure répétée, l'erreur de mesure systématique affecte la validité des résultats, mais aucunement la fidélité des scores. Notons que comme la théorie classique des tests s'est développée en premier lieu autour du concept de la fidélité, elle ne prend en compte que l'erreur aléatoire. Ainsi comme le souligne Laveault et Grégoire, les « sources d'erreur dont l'effet est constant et dont la résultante est non nulle : les erreurs systématiques. Ces sources d'erreur ne sont pas prises en ligne de compte par la théorie classique et doivent faire l'objet d'une étude particulière : la validité des résultats » (2014, p. 109).

Pour une première approche des notions de fidélité et de validité qui seront développées plus en détail ultérieurement, la décomposition de la variance du score observé à un test permet de se les représenter schématiquement. La répartition des variances est représentée dans le schéma de la Figure 12. Le score observé à un test peut se décomposer en variance pertinente (variance partagée, variance commune), en variance non pertinente (variance spécifique, variance d'erreur systématique) et en variance d'erreur (aléatoire). La variance pertinente (*construct-relevant*) représente la part de variance qui est déterminée par la propriété mentale évaluée par le test, et d'autres tests qui évaluent la même chose. La performance à un test s'interprète principalement en lien avec la propriété mentale pour lequel le test a été conçu. Or, un test ne permet pas une mesure pure. D'autres compétences psychologiques spécifiques à l'épreuve interviennent également dans la performance et constitue la part de la variance dite non pertinente (*construct-irrelevant*). En effet, un test « *at best, reflect not only the psychological constructs of knowlegde and skills that intended to be measured, but invariably a number of contaminants* » (Messick, 1984 cité par Haladyna & Downing, 2005, p. 18). Ainsi, la part de variance non pertinente englobe tout ce qu'évalue le test, mais qui ne relève pas de son interprétation de base. Dans la part de variance non pertinente réside l'erreur systématique. La dénomination pertinente et non pertinente est à mettre en lien par rapport au construit que cherche à évaluer le test. La variance d'erreur, quant à elle, comprend les inévitables erreurs qui entachent toute mesure de manière aléatoire d'une mesure à l'autre.

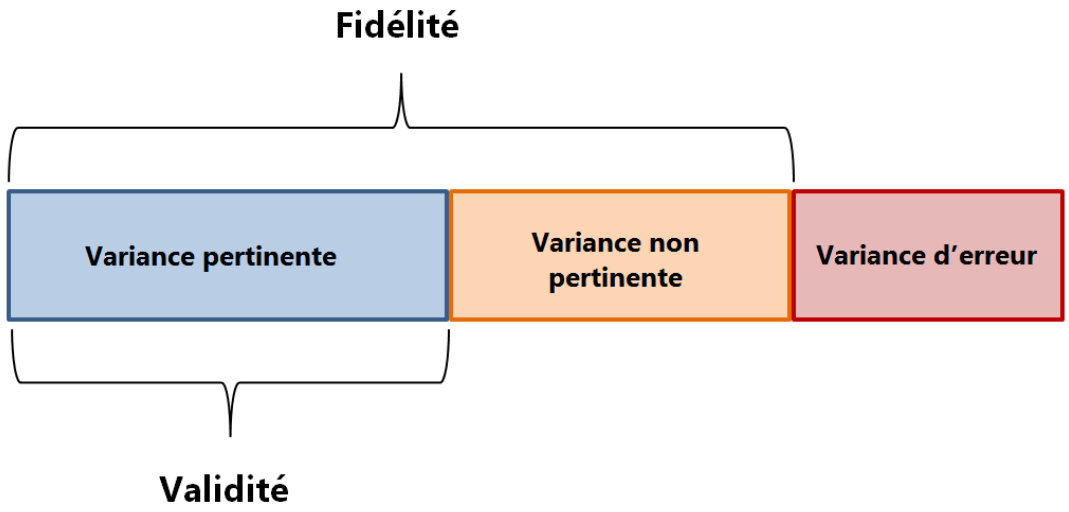


Figure 12. Illustration de la répartition des différentes variances pour un test.

Prenons l'exemple du test des matrices de Raven (Raven, 1998). Il s'agit d'un test qui évalue le raisonnement inductif (variance pertinente). L'interprétation des résultats aux matrices de Raven fomule des hypothèses sur l'habileté de raisonnement inductif du sujet. D'autres tests évaluent du raisonnement inductif tels que le subtest Matrices du WISC-IV ou le test du D2000 (Rennes, Pichot, Anstey, & Kourovsky, 2000). On s'attend à trouver une cohérence dans les résultats de tests évaluant la même propriété mentale. Toutefois, des différences dans les scores peuvent s'observer, car des aptitudes plus spécifiques à l'un ou l'autre test déterminent également la performance du sujet (variance non pertinente).

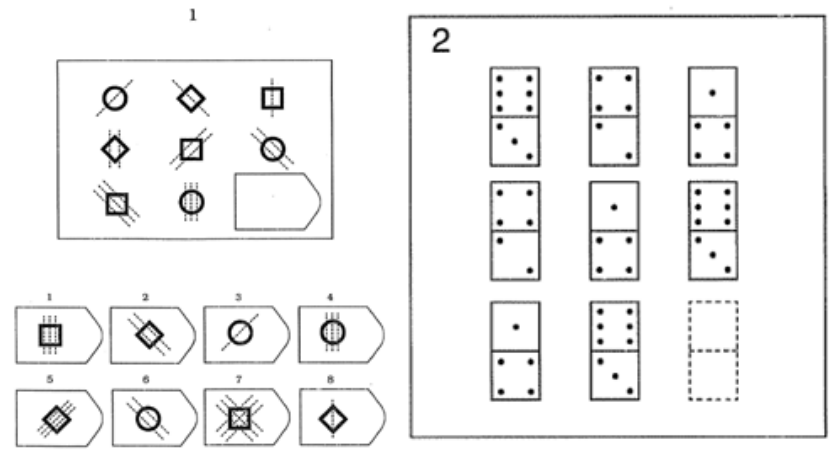


Figure 13. Exemple d'un item des matrices de Raven et du D2000.

Dans les matrices de Raven, le sujet doit trouver la logique dans une matrice de stimuli abstraits, tandis que, dans le D2000, il s'agit de trouver la logique d'une suite de dominos (voir Figure 13). Les deux tests évaluent du raisonnement inductif non verbal, toutefois pas exactement de la même manière. Chacun fait intervenir un ensemble d'habiletés plus spécifiques. Par exemple, les matrices de Raven font appel à des aptitudes plus visuospatiales que le D2000 qui, quant à lui, fait appel à des aptitudes numériques. Dans l'un, le sujet sélectionne une réponse parmi des propositions, dans l'autre, le sujet doit produire lui-même une réponse. La différence pour un sujet dans ses performances à deux tests évaluant la même chose peut donc relever de différences dans les habiletés spécifiques de chaque test. De plus, des erreurs de mesure aléatoires (non prévisibles) liées au sujet, à l'expérimentateur, au contexte de passation, etc. déterminent également la performance du sujet lors de sa passation du test.

2.1.1.2. Postulats de la théorie classique des tests

Nous allons maintenant définir les principaux postulats de la théorie classique. Nous privilégions une compréhension de la théorie classique sur le plan conceptuel, il n'y a donc pas de démonstrations pour les relations mathématiques mentionnées.

(a) La TCT postule une distribution normale du score observé ainsi que de l'erreur aléatoire de mesure. Le premier se distribue autour du score vrai et le second se distribue autour de zéro. On l'a mentionné, le score observé et l'erreur de mesure varient d'une mesure répétée à l'autre. Néanmoins, la probabilité d'obtenir un score observé très éloigné du score vrai du sujet est moins probable que d'obtenir un score observé proche de son score vrai. De même, l'erreur de mesure se distribue normalement autour de la moyenne de zéro. Sur un très grand nombre de répétitions d'un test par un même individu, la moyenne de ses scores observés est par définition son score vrai. Par conséquent, sur un très grand nombre de répétitions, la moyenne des erreurs de mesure aléatoires devient nulle.

(b) La TCT postule une corrélation nulle entre les scores vrais et les erreurs de mesure aléatoire pour un test donné. Théoriquement, il n'y a pas de relations entre les scores vrais et les erreurs de mesure d'une population d'individus à qui l'on administre un même test. Cela signifie par exemple qu'on n'observerait pas d'erreurs de mesure plus importantes chez les individus ayant de faibles habiletés sur une propriété mentale que chez les individus ayant des habiletés élevées. La valeur de l'erreur type de mesure – qui est la moyenne des écarts types des erreurs de mesure pour l'échantillon – est donc considérée constante quel que soit le score observé de l'individu.

(c) La TCT postule l'absence de relation entre les erreurs de mesure associées à deux tests différents. Cela signifie qu'« on ne peut donc pas prédire directement les erreurs de mesure d'individus à un test à partir des erreurs des mêmes individus à un autre test » (Bertrand & Blais, 2004, p. 46). Ce postulat peut ne pas se vérifier par exemple, si un même groupe d'individus passent deux tests différents en fin de journée, le facteur fatigue peut affecter leur performance dans les deux tests (Laveault & Grégoire, 2014).

En somme dans la théorie classique, l'interprétation des scores observés repose sur la condition que les erreurs aléatoires de mesures sont indépendantes de toute autre variable. En effet, « si les différences sources d'erreur sont indépendantes les unes des autres, alors celles-ci pourront s'annuler de sorte que sur un grand nombre de mesures répétées, l'espérance mathématique des scores observés soit le score vrai de

l'individu » (Laveault & Grégoire, 2014, p. 108). Pour que le score observé soit interprétable comme une estimation du score vrai, il faut pouvoir exclure que le score observé soit lié à l'influence de la relation entre l'erreur de mesure et une variable (p. ex., la motivation qui amènerait les sujets à répondre d'autant plus au hasard qu'ils sont peu motivés).

2.1.1.3. Indice de difficulté et indice de discrimination

Dans la TCT, deux indices permettent d'évaluer les caractéristiques des items : le p-indice et D-indice. Il nous semble intéressant de les évoquer pour la comparaison avec les indices des modèles de réponse à l'item. Dans la théorie classique, on détermine la difficulté d'un item par le calcul d'un indice de difficulté – nommé p-indice – selon la formule suivante pour les items dichotomiques.

$$p - indice = \frac{nombre\ d'individus\ qui\ réussissent\ l'item}{nombre\ total\ d'individus} \qquad (3)$$

Il s'agit d'un simple calcul de proportion. La valeur du p-indice d'un item varie entre 0 (tous les individus ont échoué l'item) et 1 (tous les individus ont réussi l'item). Plus le p-indice est grand, plus l'item est facile ; il s'interprète en fait comme un indice de facilité. D'après le calcul, on voit que le p-indice va dépendre du niveau d'habileté des individus de l'échantillon qui sert à son estimation. Si les individu de l'échantillon présentent en moyenne de faibles habiletés sur ce que le test évalue, un item du test peut être échoué par la plupart des individus de l'échantillon (p-indice proche de 0) et donc considéré comme difficile, alors qu'il peut être considéré comme facile avec un autre échantillon d'individus possédant des habiletés plus élevées.

L'évaluation de la discrimination d'un item renseigne sur la finesse avec laquelle l'item peut différencier les individus les uns par rapport aux autres. Un item discriminant permet de situer dans quel groupe appartient un individu (p. ex., le groupe des « forts », groupe des « moyens », groupe des « faibles »). Dans la théorie classique, il y a trois méthodes principales : les corrélations inter-items, les corrélations item-score total et l'indice de discrimination D-indice. Pour le calcul du D-indice, on constitue deux groupes extrêmes parmi les individus d'un échantillon. On calcule ensuite le p-indice de chaque item pour le groupe des ~30 % les plus forts et pour le groupe des ~30 % les plus faibles. La différence des p-indices entre le groupe « fort » et le groupe « faible » donne le D-indice pour chaque item. La valeur du D-indice varie entre -1 et +1. Une

valeur négative indique que les individus avec un faible score total réussissent mieux l'item que les individus avec un score total élevé. Un D-indice nul indique qu'autant les individus avec un faible score que les individus avec un score élevé réussissent l'item. Les items avec des D-indice à partir de .40 discriminent très bien.

2.1.1.4. Limites de la théorie classique des tests

Plusieurs limites sont connues pour la théorie classique. Tout d'abord, la TCT s'inscrit dans les statistiques fréquentistes, lesquelles se réfèrent à des expériences aléatoires répétées un grand nombre de fois (*long-run repetition of the same experiment*). Or, dans la pratique, nous ne pouvons pas répéter la passation d'un même test à un même individu une infinité de fois. Nous avons donc qu'une observation de l'expérience (*single case*). De plus, les inférences fréquentistes reposent sur le postulat des mesures répétées indépendantes. En effet, pour se conformer au cadre fréquentiste, il faudrait non seulement répéter l'expérience aléatoire un très grand nombre de fois, mais aussi que les résultats de chaque expérience répétée soient indépendants les uns des autres. Dans le cas d'une passation de test, cela suppose que le score obtenu d'un sujet à chacune des innombrables passations successives du même test doit théoriquement être indépendant de l'expérience de la précédente passation. Ce postulat n'est pas tenable, à moins de faire subir au sujet un lavage de cerveau qui le restaure « *to his original state – not only with respect to memory, learning, and fatiguing effects, but with respect to time itself* » (Borsboom, 2005, p. 259).

Une autre limite de la TCT est qu'il n'est pas stipulé de restriction sur l'erreur de mesure. Rappelons que la valeur des deux entités théoriques que sont le score vrai et l'erreur de mesure nous est inconnue. On pourrait inférer leur valeur sur un grand nombre de répétitions de l'expérience aléatoire, mais pas à partir d'un petit nombre de cas et encore moins à partir d'une seule mesure. Ainsi, si un sujet passe à deux reprises un même test, et qu'on observe une diminution du score observé à la seconde passation, cela n'implique pas que la diminution soit expliquée par une diminution de son score vrai. Comme il n'y pas de restriction sur l'erreur de mesure, la diminution du score peut aussi bien résulter d'une augmentation de l'erreur de mesure que d'une diminution effective sur le niveau de la propriété mentale évaluée. Nous ne pouvons pas le savoir. De même, si deux sujets obtiennent des scores observés différents (dans une certaine mesure) à un même test, cela n'implique pas que le sujet avec le score observé le plus élevé a effectivement le score vrai le plus élevé. De plus, l'erreur de

mesure est supposée de même ampleur pour tous les individus quel que soit leur habileté sur la propriété mentale évaluée. Ce postulat d'homoscédasticité peut ne pas se vérifier en toutes circonstances.

La dernière grande limite de la TCT est que les estimations reposent sur une dépendance circulaire. En effet, « le paramètre d'habileté (c.-à-d. le résultat observé) est dépendant de l'échantillon d'items et les paramètres d'items (c.-à-d. la difficulté et la discrimination de l'item) sont dépendants de l'échantillon de sujets » (Frenette, Bertrand, Valois, Dussault, & Hébert, 2007). Sans mener une étude, on ne peut pas généraliser les résultats d'un échantillon sur d'autres échantillons (différents en termes de sexe, d'appartenance à un groupe clinique, de culture, etc.) à qui on passe le même test. Comme on l'a vu avec le p-indice, un item facile (p-indice élevé) au sein d'un échantillon peut ne pas l'être (ou ne pas l'être autant) dans un autre échantillon. Dans cette dépendance circulaire, les propriétés des items sont dépendantes de l'échantillon de sujets, et les caractéristiques des sujets sont dépendantes des items. Dit autrement, on évalue les sujets par rapport à leur performance sur les items d'un test, dont les propriétés dépendantes des caractéristiques de l'échantillon qui a servi à les estimer. Nous aurons l'occasion de revenir sur certaines limites de la théorie classique dans la comparaison avec les modèles des réponses à l'item que nous allons présenter à la suite.

2.1.2. MODÈLES DE RÉPONSE À L'ITEM

Se développant à partir de la seconde moitié du 20^{e} siècle, la Théorie de la Réponse à l'Item (TRI, ou IRT pour Item Reponse Theory dans la littérature anglophone) apporte un nouveau cadre de mesure dans le champ de la psychométrie. La TRI est constituée de plusieurs modèles de réponses à l'item selon les paramètres estimés (p. ex., modèle logistique à un, deux ou trois paramètres), les contraintes sur les paramètres (p. ex., modèle de Rasch) ou selon qu'il s'agit d'items dichotomiques ou polytomiques (p. ex., modèle gradué de Samejima). Il est également possible de modéliser des modèles multidimensionnels, mais nous ne traiterons que des modèles unidimensionnels, à savoir ceux qui modélisent un seul trait latent, une seule habileté. Pour rendre compte des différents modèles que regroupe cette théorie de mesure, certains préfèrent parler de Modèles de Réponse à l'Item (MRI). Le développement des modèles de réponse à l'item est lié au projet de créer une banque informatisée qui recense les items des tests psychologiques. L'idée étant qu'on puisse piocher dans un

large ensemble d'items ceux qu'on utilisera pour construire un test, il est donc nécessaire que les items intégrés dans la banque d'items possèdent une métrique commune en termes de propriétés métrologiques et de niveau d'habileté des sujets. Or, les caractéristiques des items dans le cadre de la théorie classique des tests sont dépendantes de l'échantillon testé qui sert à leur estimation. En revanche, dans le cadre de la théorie de réponse à l'item, les propriétés de l'item sont estimées indépendamment des caractéristiques de l'échantillon particulier utilisé pour l'estimation. L'estimation de la difficulté ou de la discrimination d'un item n'est donc plus dépendante du niveau moyen des individus de l'échantillon à qui on a administré le test. Au premier abord, les modèles de réponses à l'items nous plongent dans des concepts théoriques ainsi que des procédures mathématiques et techniques plus complexes que la théorique classique des tests. Le coût d'entrée est élevé pour appréhender ce cadre théorique, freinant son expansion. Néanmoins durant ces dernières années, les avancées des logiciels statistiques toujours plus performants relancent l'intérêt pour les MRI. À la suite, nous allons retracer les grandes étapes qui conduisent à la construction de ce cadre de mesure.

Tout commence au début des années 50. Louis Guttman cherche à construire une échelle absolue. On parle d'échelle absolue, lorsque les mesures sont indépendantes d'un groupe de référence. Dans sa tentative, il élabore une échelle sur un modèle déterministe selon lequel, pour chaque item, il existe un point particulier du continuum au-delà duquel l'individu ayant une habileté supérieure à la difficulté de l'item doit le réussir, tandis qu'un individu ayant une habileté inférieure à la difficulté de l'item doit l'échouer. Toute autre combinaison est considérée comme une incohérence ou une dissimulation du sujet. À chaque score total d'un test construit selon le modèle de Guttman correspond donc un seul et unique pattern de réponses. Dans la Figure 14 (p. 87), le seuil de coupure est graphiquement représenté par une droite perpendiculaire à l'abscisse (trait en pointillé). Pour satisfaire le modèle de Guttman, les items d'un test doivent être correctement ordonnés de difficulté croissante afin d'évaluer une seule et même propriété mentale sur un continuum. La position du sujet sur le continuum est entièrement déterminée par son niveau sur la variable évaluée. Une limite du modèle déterministe de Guttman est que pour un item donné, on ne peut pas discriminer entre deux individus qui ont une habileté supérieure à la difficulté de l'item, mais différentes. Dès lors, il faut un très grand nombre d'items pour établir une hiérarchie entre les individus. Une autre limite vient des inévitables erreurs de mesure qui introduisent des variations dans les réponses du sujet qui ne sont pas liées

à la variable évaluée. Le tout ou rien du modèle d'échelle de Guttman est laissé de côté au profit d'un modèle probabiliste sur lequel repose les modèles de réponse à l'item.

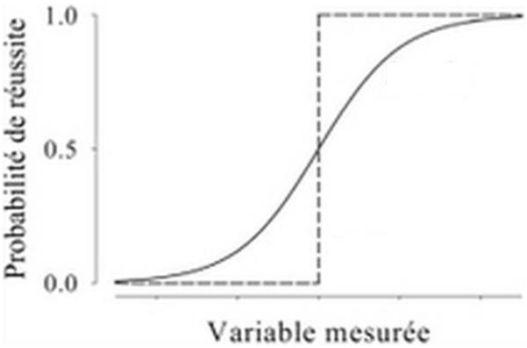


Figure 14. Comparaison entre le modèle déterministe de Guttman (trait en pointillé) et un modèle probabiliste (trait plein).

Dans le modèle probabiliste, la probabilité de réussir l'item augmente en continu (mais pas linéairement) avec le niveau d'habileté du sujet. Contrairement au modèle de Guttman, le modèle probabiliste permet de distinguer entre deux personnes ayant une habileté supérieure à la difficulté de l'item, mais différente, puisque leur probabilité de réussir l'item est graduellement différente. La probabilité de réussir chaque item est fonction de l'habileté de l'individu sur la propriété mentale évaluée (caractéristique psychologique du sujet) et des propriétés métrologiques de l'item (son degré de difficulté, son pouvoir discriminant). Selon Wim van der Linden, c'est en réaction à la théorie classique qu'émerge la théorie de réponse à l'item dans les années 1950 et 1960 :

> [À la différence de la théorie classique des tests], la théorie de réponse à un item n'est pas attachée aux scores obtenus à des tests par des échantillons aléatoires, mais aux réponses individuelles à des items particuliers. Ces réponses sont modélisées comme le résultat d'un processus stochastique dans lequel la probabilité de donner une réponse d'un certain type dépend de plusieurs paramètres. Ces paramètres peuvent soit être liés aux personnes [c.-à-d. habileté, compétence], soit aux items [p. ex., difficulté, discrimination]. [traduction libre] (1986, p. 329)

Comme l'habileté de l'individu n'est pas directement observable, il s'agit d'un trait latent, désigné par la lettre grecque θ (thêta). Pour les modèles de réponse à l'item les plus couramment utilisés, la probabilité de réussite selon l'habileté du sujet et les caractéristiques de l'item est traduite mathématiquement par une fonction logistique à un, deux ou trois paramètres appelée fonction caractéristique de l'item. Par exemple, ci-

dessous l'équation de la fonction caractéristique de l'item pour un modèle à un paramètre :

$$P_i(\theta) = \frac{1}{1 - e^{-D(\theta - b_i)}} \quad (4)$$

où $P_i(\theta)$ est la probabilité de réussir l'item i pour un individu possédant un certain degré d'habileté θ (ou niveau sur le trait latent θ), e est la constante de Neper (valant env. 2.71828), D est un facteur d'échelonnement (généralement fixé à la valeur de 1.7 pour une allure d'ogive normale) et b_i est le paramètre de difficulté pour l'item i. Les fonctions caractéristiques de l'item ont une représentation graphique non seulement aisé à lire, mais surtout leur forme d'ogive décrit de façon appropriée la relation non linéaire entre l'habileté et la probabilité de répondre correctemement à un item.

2.1.2.1. Courbe Caractéristique d'Item

On appelle Courbe Caractéristique d'Item (CCI) la représentation graphique de la relation entre le trait latent θ et la probabilité de réussir l'item (voir Figure 15). Modélisant une fonction logistique, la CCI prend une forme sigmoïde (c.-à-d. un S plus ou moins étiré).

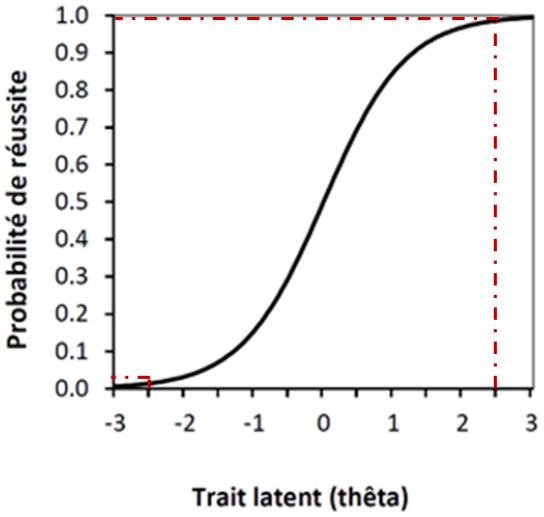


Figure 15. Exemple d'une courbe caractéristique de l'item

En ordonnée, nous avons la probabilité de réussir l'item (P_i) allant d'une probabilité nulle (soit 0 % de réussite) à une probabilité de 1 (soit 100 % de réussite). En abscisse, nous avons le trait latent θ exprimé généralement sur une échelle analogue aux scores z, qui, pour rappel, ont une distribution centrée et réduite de moyenne 0 et d'écart type 1. Théoriquement, les valeurs de thêta varient de $-\infty$ à $+\infty$, en pratique, on présente une échelle allant de -3 (extrêmement faible habileté) à 3 (extrêmement forte habileté). Une valeur de thêta élevée et positive signifie un niveau élevé d'habileté sur la propriété mentale évaluée par l'item. Par exemple, dans la Figure 15 (p. 88), la CCI montre que les individus ayant un niveau très faible d'habileté sur la propriété mentale évaluée ($\theta = -2.5$) auront une probabilité de réussir l'item quasi nulle à l'item (~3 %), tandis que les individus ayant un niveau d'habileté très élevé ($\theta = 2.5$) auront une probabilité de réussir l'item de 0.99 (soit 99 %).

2.1.2.2. Paramètre de difficulté

Nous l'avons mentionné, il existe plusieurs modèles au sein de la théorie de réponse à l'item. Nous allons présenter les plus courants à savoir les modèles à un, deux ou trois paramètres pour les items dichotomiques (réponse binaire : réussi – échoué). Commençons d'abord par les modèles de réponse à l'item logistiques à un paramètre (1 PL) qui estiment uniquement le degré de difficulté de l'item comme exprimée dans l'équation (4). On parle bien des modèles à un paramètre au pluriel, car il y a théoriquement un modèle 1 PL pour chaque valeur fixée du paramètre de discrimination a_i et de la constante D. Cependant, les modèles à un paramètre les plus fréquemment utilisés sont le modèle de Rasch (avec les paramètres fixés suivants : $a_i = 1$, $c_i = 0$ et $D = 1$) et le modèle logistique normal à un paramètre (1 PL ; avec les paramètres fixés suivants : $a_i = 1$, $c_i = 0$ et $D = 1.7$). Nous allons débuter par la présentation de ce dernier dont la fonction caractéristique de l'item s'exprime donc par l'équation suivante :

$$P_i(\theta) = \frac{1}{1 - e^{-1.7(\theta - b_i)}} \qquad (5)$$

L'attribution de la valeur de 1.7 à la constante D donne aux courbes caractéristiques de l'item une allure proche d'une ogive normale avec une échelle de thêta analogue aux scores z (allant de -3 à 3). Par convention, pour les modèles à un et deux paramètres, le paramètre de difficulté b_i est défini comme « la valeur de θ pour

laquelle la probabilité de donner une réponse correcte est de 0.5 » (Laveault & Grégoire, 2014, p. 276). À la différence du p-indice calculé dans la théorie classique, il s'agit d'un paramètre dont l'interprétation se lit comme un véritable indice de difficulté. Nous allons le constater à travers une illustration. Sur la Figure 16 (p. 90) sont représentées les CCI de trois items d'un test.

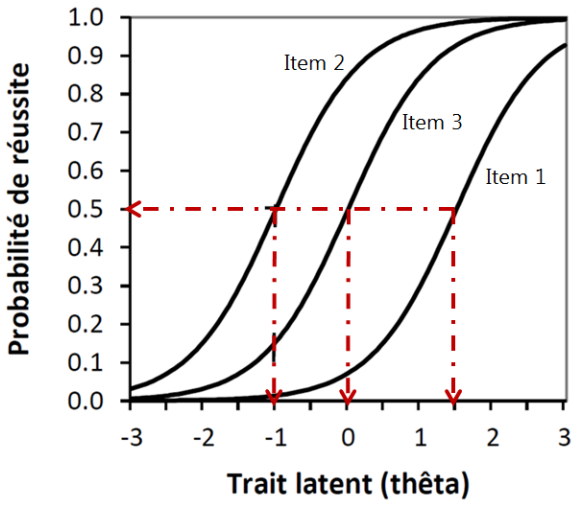


Figure 16. Trois courbes caractéristiques de l'item avec différents paramètres de difficulté ($b_1 = 1.5$, $b_2 = -1$ et $b_3 = 0$).

Pour les valeurs exactes du paramètre de difficulté b_i, il faut examiner les résultats du calcul de son estimation par un logiciel statistique. Cependant, à partir des CCI, on peut déjà avoir de manière visuelle une estimation de la valeur de b_i pour les trois items de la Figure 16. Pour cela, on regarde à quelle valeur de thêta (niveau d'habileté de l'individu sur la propriété mentale évaluée) correspond une probabilité de réussir l'item de 0.5 (soit 50 % de réussite pour l'item). Rappelons qu'un trait latent θ positif et élevé correspond à un niveau d'habileté élevé, tandis qu'un trait latent θ négatif et élevé correspond à un niveau d'habileté faible. Un trait latent θ égal à 0 correspond à un niveau moyen d'habileté. Pour l'item 1 dont la CCI est la plus à droite sur la Figure 16, un pourcentage de réussite de 50 % est atteint à $\theta = 1.5$. Cela indique un item plutôt difficile, puisqu'un individu possédant une habileté au-dessus de la moyenne de $\theta = 1.5$ a 50 % de probabilité de réussir l'item 1. Pour l'item 2 dont la CCI est la plus à gauche sur la Figure 16, un pourcentage de réussite de 50 % est atteint à $\theta = -1$. Cela indique un item plutôt facile. Pour l'item 3, un pourcentage de réussite de 50 % est atteint à $\theta = 0$, indiquant un item de difficulté moyenne. La valeur du

paramètre de difficulté b_i est exprimée sur la même échelle que le trait latent θ, ainsi $b_1 = 1.5$, $b_2 = -1$ et $b_3 = 0$. Si on ordonne les items par ordre croissant de difficulté, cela donne item 2, item 3 et enfin item 1 le plus difficile. Ainsi, un individu d'habileté moyenne ($\theta = 0$) a plus de difficulté de réussir l'item 1 ($b_1 = 1.5$) que l'item 3 ($b_3 = 0$) ; mais aussi plus difficulté à réussir l'item 3 que l'item 2 ($b_2 = -1$). Dans un modèle à un paramètre, plus la valeur du paramètre de difficulté b_i est positive et élevée, plus l'item est difficile.

De façon visuelle, plus la courbe caractéristique de l'item se trouve à droite de l'échelle du trait latent, plus l'item est difficile. À l'inverse, plus la courbe caractéristique de l'item se trouve à gauche de l'échelle du trait latent, plus l'item est facile. Par ailleurs en vertu des propriétés de la distribution normale centrée et réduite, on peut déterminer à partir de la valeur de b_i la proportion d'individus dont la probabilité de réussir l'item est d'au moins 50 %. Par exemple, pour un item possédant une valeur $b_i = -1$ (item de difficulté plutôt faible), il y a 84 % des individus (34.13 % + 34.13 % + 13.59 % + 2.14 % + 0.13 %) qui ont une probabilité d'au moins 50 % de le réussir.

Le modèle de Rasch (1960) est le plus courant des modèles à un paramètre, dans lequel on suppose une discrimination égale pour tous les items d'un test (soit $a_i = 1$, $c_i = 0$ et $D = 1$). Il est populaire par sa complexité moindre, mais le revers à sa simplicité est la contrainte forte pour une discrimination égale à 1 de tous les items du test. Comme l'illustre la Figure 17, les CCI dans le modèle de Rasch ont une allure identique, seulement décalées les unes des autres selon leur paramètre de difficulté b_i.

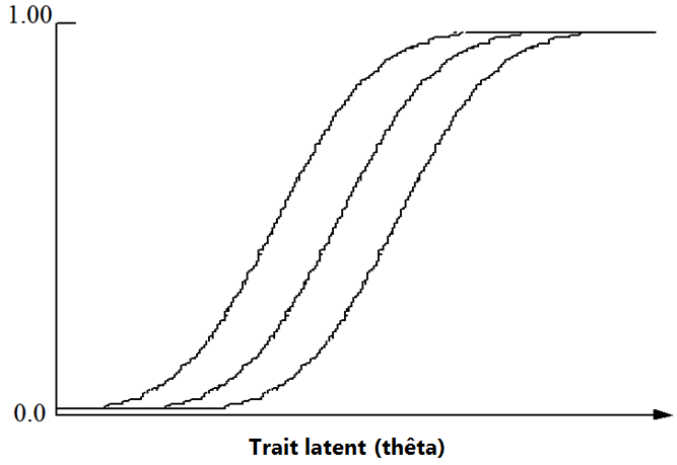


Figure 17. Modèle de Rasch.

2.1.2.3. Paramètre de discrimination

Les modèles de réponse à l'item à deux paramètres (2PL ; introduits par Birnbaum, 1968) ajoutent l'estimation du paramètre de discrimination a_i, dont voici l'équation pour le modèle logistique normal à deux paramètres (2 PL ; avec les paramètres fixés suivants $c_i = 0$ et $D = 1.7$) :

$$P_i(\theta) = \frac{1}{1 - e^{-1.7a_i(\theta - b_i)}} \qquad (6)$$

Plus un item est discriminant, plus il permet de finement distinguer les individus entre eux selon leur niveau d'habileté. De manière visuelle, le pouvoir discriminant d'un item se détermine à la pente de la droite tangente au point d'inflexion de la CCI (voir Figure 18). Le point d'inflexion étant l'endroit où la courbe passe du concave au convexe.

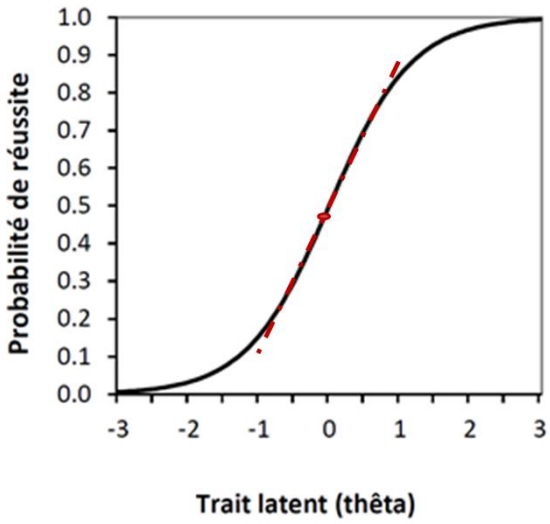


Figure 18. Courbe caractéristique de l'item avec la droite tangente (trait en pointillé) au point d'inflexion (point rouge).

Plus la pente de la droite tangente au point d'inflexion est abrupte, meilleure est la discrimination. En effet, un petit changement de niveau sur le trait latent va augmenter (ou diminuer) la probabilité de réussir l'item. Au point d'inflexion, la pente est maximale et c'est donc l'endroit où l'item est le plus discriminant. Plus la valeur du paramètre de discrimination est élevée, plus l'item discrimine finement les individus dans la zone du point d'inflexion. Si la valeur du paramètre est négative, cela indique

que l'item est mieux réussi par les individus possédant de faibles habiletés que les individus possédant des habiletés élevés sur ce qu'évalue l'item.

Les valeurs habituelles du paramètre de discrimination a_i se situent entre 0 et 2. Pour des items avec le paramètre a_i supérieur à 2, cela indique une pente extrêmement raide, ce qui est peu fréquent (Laveault & Grégoire, 2014). En outre plus la pente est raide, plus la zone avoisinant le point d'inflexion est restreinte et, donc, plus l'item discrimine très finement sur une petite étendue de l'échelle du trait latent. Dans un test, il peut être opportun d'avoir également des items qui discriminent plus modérément, mais sur une grande étendue de l'échelle thêta.

Si le point d'inflexion se situe vers des valeurs positives et élevées de thêta, cela indique que l'item discrimine finement au sein des individus possédant des habiletés élevées sur le trait latent évalué. Et inversement, si le point d'inflexion se situe vers des valeurs négatives et élevées de thêta, cela indique que l'item discrimine finement au sein des individus possédant des habiletés faibles. Sur la Figure 18 (p. 92), le point d'inflexion se situe à $\theta = 0$ et indique que la discrimination est la plus élevée au sein des individus d'habileté moyenne.

La Figure 19 (p. 94) présente l'exemple de deux items (item 1 et item 2) avec un même degré de difficulté ($b_1 = b_2 = 0$). En revanche, les items diffèrent sur l'indice de discrimination a_i. L'item 1 a un indice de discrimination ($a_1 = 2$) plus élevé que l'item 2 ($a_2 = 0.5$). Néanmoins, on peut voir sur le graphique qu'aux valeurs de thêta entre -3 et -1, l'item 2 est plus discriminant que l'item 1 dont la CCI est plate (tous les individus échouent l'item). De même pour des valeurs de thêta de +1 et +3, l'item 2 est aussi plus discriminant que l'item 1. Ainsi, l'indice de discrimination a_i n'est pas un indice global de discrimination de l'item, mais il s'interprète « comme un indice de discrimination de l'item i dans le voisinage du point d'inflexion » (Bertrand & Blais, 2004, p. 131).

À la différence des CCI des modèles à un paramètre qui sont parallèles les unes aux autres à cause d'un paramètre de discrimination fixé ($a_i = 1$), les CCI des modèles à deux paramètres peuvent se croiser puisque leur pente diffère (voir Figure 19, p. 94). L'interprétation du paramètre de difficulté b_i est alors plus nuancée que pour les modèles à un paramètre. Si l'on reprend la Figure 19 (p. 94), le paramètre de difficulté est le même pour les deux items ($b_1 = b_2 = 0$). Cependant, nous pouvons voir sur le graphique qu'il est plus difficile de réussir l'item 1 que l'item 2 pour les individus d'habileté en dessous de la moyenne ($\theta < 0$) où tous les individus à partir d'un trait

latent inférieur à -1 échouent l'item 1. En revanche, pour des individus possédant des habiletés au-dessus de la moyenne ($\theta > 0$), il est plus difficile de réussir l'item 2 que l'item 1 où tous les individus le réussissent dès un niveau de trait latent $\theta > 1$.

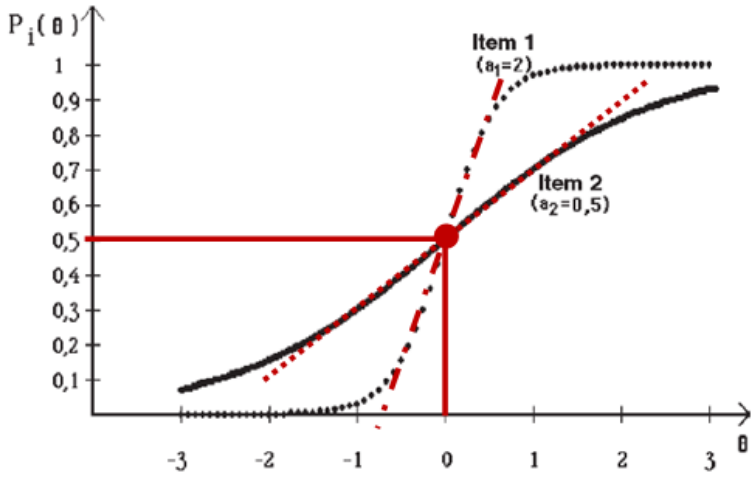


Figure 19. Deux courbes caractéristiques de l'item avec différents paramètres de discrimination ($a_1 = 2$ et $a_2 = 0.5$). *Source* : Bertrand et Blais (2004, p. 130).

Ainsi, dans les modèles à plusieurs paramètres estimés, l'examen visuel de la courbe caractéristique de l'item est nécessaire pour éviter les erreurs d'interprétation, puisque la valeur des paramètres estimés ne donne qu'une indication globale. Par ailleurs, nous verrons dans les modèles à trois paramètres que, « strictement parlant, la valeur du paramètre b_i donne une indication de la position du point d'inflexion de la CCI » (Bertrand & Blais, 2004, p. 132).

2.1.2.4. Paramètre de pseudo-chance

Dans des modèles plus complexes à trois paramètres (3 PL) librement estimés, les modèles de réponse à l'item intègrent le rôle de la chance. Avec des items à choix de réponses proposées, il peut arriver qu'en essayant de deviner au hasard la réponse correcte, on tombe en effet sur celle-ci. Ce phénomène est estimé avec le paramètre de pseudo-chance c_i (*guessing parameter* dans la littérature anglophone) qui met en évidence la probabilité de choisir la réponse correcte parmi celles proposées pour un niveau de trait latent aussi faible que possible. Du fait qu'il n'y a pas une même

probabilité de choix parmi les propositions incorrectes, on ne l'appelle pas le paramètre de chance. En effet, parmi les distracteurs, il y en a qui sont plus enclins à nous induire en erreur. Le paramètre de pseudo-chance est donc plus petit que la probabilité de répondre totalement au hasard (Hambleton et al., 1991). La valeur de ce paramètre de pseudo-chance c_i est exprimée sur la même échelle que la probabilité de réussir l'item, et varie donc de 0 (soit 0 % de probabilité de répondre correctement grâce à la devinette) à 1 (soit 100 % de probabilité de répondre correctement grâce à la devinette). Généralement, les valeurs de ce paramètre sont inférieures à 0.3 (Harris, 1989). L'équation pour le modèle logistique normal à trois paramètres (3 PL ; avec le paramètre fixé : $D = 1.7$) est la suivante :

$$P_i(\theta) = c_i \frac{1 - c_i}{1 - e^{-1.7a_i(\theta - b_i)}} \qquad (7)$$

Sur la Figure 20 est représentée la courbe caractéristique d'un item dont le paramètre de pseudo-chance est $c_i = 0.15$. Cela signifie qu'un individu possédant une habileté aussi faible que possible (ici $\theta = -3$) a néanmoins une probabilité de 15 % de réussir l'item grâce à la devinette. Le paramètre de pseudo-chance peut s'interpréter comme la probabilité minimale de réussir l'item.

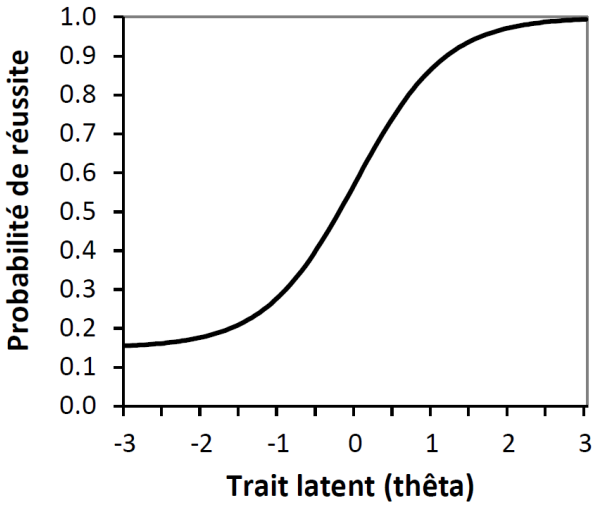


Figure 20. Courbe caractéristique de l'item avec un paramètre de pseudo-chance $c_i = 0.15$.

Sur la Figure 21 (p. 96) sont représentées trois courbes caractéristiques de l'item. Les trois items présentent un même paramètre de difficulté ($b_1 = b_2 = b_3 = 0$) et un

même paramètre de discrimination ($a_1 = a_2 = a_3 = 1$). Nous voyons que, dans un modèle à trois paramètres, le paramètre de difficulté correspond à la valeur de thêta au point d'inflexion, dont les coordonnées sont justement $[b_i\ ; (1 + c_i)/2]$.

> Ainsi, l'abscisse du point d'inflexion donne la valeur du paramètre de difficulté b_i. L'ordonnée du point d'inflexion est située à mi-distance entre c_i et 1, où c_i peut être vue comme la probabilité minimale de réussir l'item et 1, la valeur maximale que peut prendre la probabilité de réussir l'item. Incidemment, lorsque $c_i = 0$, donc dans le cas d'un modèle à deux paramètres, l'ordonnée du point d'inflexion vaut tout simplement ½. Dans ce cas, on peut interpréter le point d'inflexion d'un item i comme l'endroit où l'on passe le cap psychologique du 50 % des chances de réussir l'item i. (Bertrand & Blais, 2004, p. 137)

Quant au paramètre de pseudo-chance des trois CCI, il diffère ($c_1 = 0.5$, $c_2 = 0.2$ et $c_3 = 0$). Pour des valeurs de thêta aussi faible que possible (ici $\theta = -3$), il y a grâce à la devinette une probabilité de 50 % de réussir l'item 1, de 20 % de réussir l'item 2 et de 0 % de réussir l'item 3.

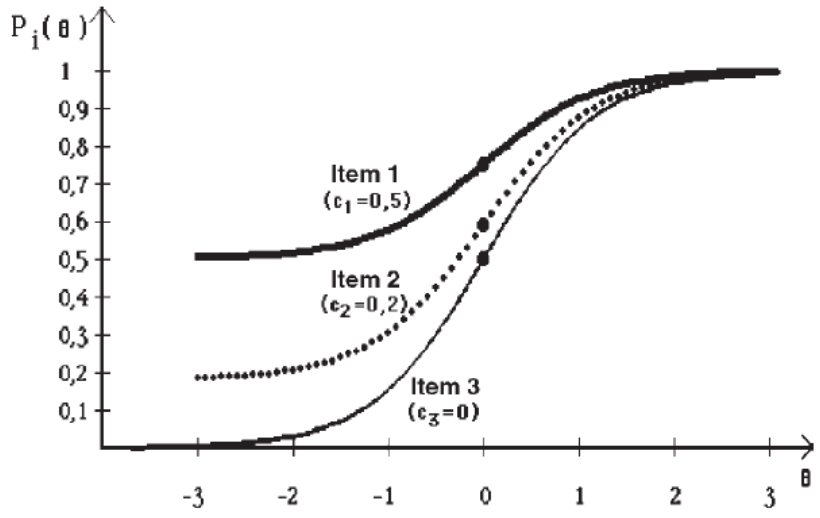


Figure 21. Trois courbes caractéristiques de l'item avec différents paramètres de pseudo-chance ($c_1 = 0.5$, $c_2 = 0.2$ et $c_3 = 0$). *Source* : Bertrand et Blais (2004, p. 135).

On peut remarquer que le paramètre de pseudo-chance n'a que peu d'impact sur la probabilité de réussir l'item pour des individus possédant une habileté élevée. En effet, à partir d'un certain niveau de thêta, l'item est de toute façon réussi quelque soit

la valeur du paramètre de pseudo-chance (voir Figure 21 où à partir de $\theta = 2$ la probabilité de réussir les items 1, 2 et 3 est de 100%).

Lorsque les trois paramètres sont librement estimés, l'allure des CCI ressemble à ce qui est représenté dans Figure 22 (p. 97). Les points noirs indiquent le point d'inflexion des CCI. Rappelons que, dans un modèle à trois paramètres, le paramètre de difficulté a_i correspond à la valeur de thêta au point d'inflexion. De manière visuelle, on peut décrire l'item 1 avec $b_1 = -2$ et $c_1 = 0.4$, l'item 2 avec $b_2 = -1.3$ et $c_2 = 0.4$, l'item 3 avec $b_3 = -0.5$et $c_3 = 0.2$, et l'item 4 avec $b_4 = 1.2$ et $c_i = 0.3$. La valeur exacte du paramètre de discrimination a_i ne peut pas être connue à partir du graphique, néanmoins, l'analyse visuelle des pentes des CCI permet de décrire la discrimination des items à certains endroits. Par exemple, l'item 4 est particulièrement discriminant pour des valeurs de thêta autour de 1.5.

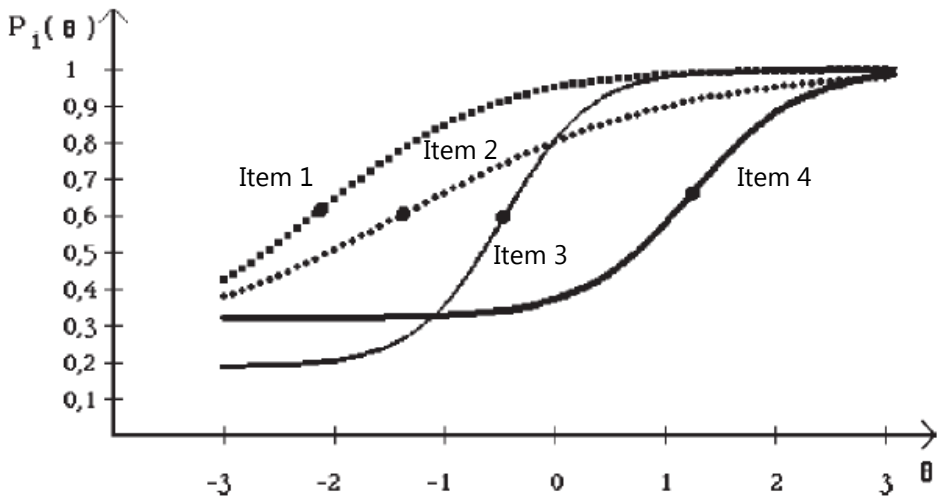


Figure 22. Quatre courbes caractéristiques de l'item avec différents paramètres de difficulté, de discrimination et de pseudo-chance. *Source* : Bertrand et Blais (2004, p. 136).

Nous venons d'étudier le comportement des items d'un test, les MRI permettent également d'examiner le comportement global du test le long de l'échelle d'habileté au moyen de la courbe caractéristique du test et la fonction d'information. Nous allons présenter d'abord la courbe caractéristique du test.

2.1.2.5. Courbe Caractéristique du Test

À partir de l'addition des courbes caractéristiques de chaque item d'un test, on peut définir une Courbe Caractéristique du Test (CCT). Concrètement, il s'agit de sommer les valeurs $P_i(\theta)$ de chaque item du test pour chaque valeur de l'échelle du trait latent θ. Nous obtenons alors en ordonnée, une échelle des $\sum P_i(\theta)$ qui correspond à une échelle du nombre d'items du test (allant de 0 à n items) et en abscisse, l'échelle du trait latent θ. La Figure 23 (p. 98) illustre la courbe caractéristique du même test dont les courbes caractéristiques de ses items sont illustrées sur la Figure 22. Ce test comporte 4 items, donc : $n = 4$ et $n/2 = 2$. Pour les quatre items de la Figure 22, si l'on somme les probabilités de réussir l'item à $\theta = -3$, on obtient $\sum P_i(-3) = 0.4 + 0.4 + 0.2 + 0.3 = 1.3$, ce qui correspond à la valeur de l'ordonnée à $\theta = -3$ de la Figure 23.

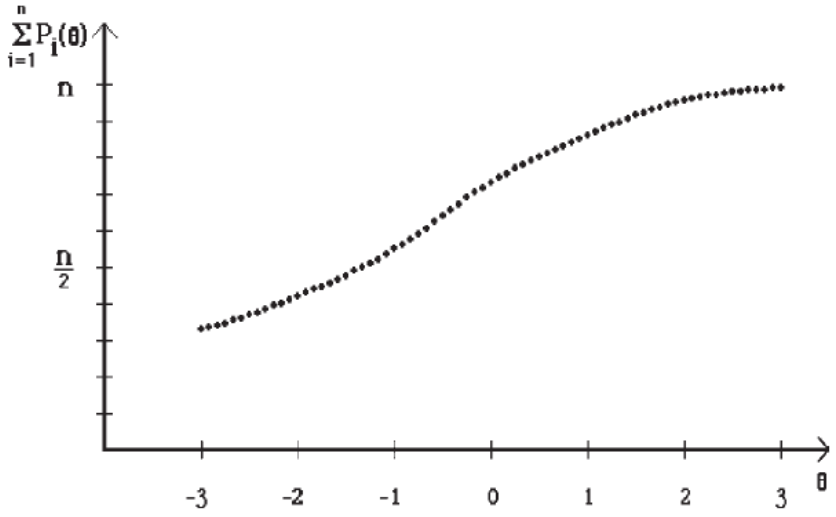


Figure 23. Courbe caractéristique du test de la Figure 22. *Source* : Bertrand et Blais (2004, p. 140).

L'échelle des $\sum P_i(\theta)$ est en fait une échelle des scores vrais V (pour la démonstration mathématique, voir annexe 4.2 de Bertrand & Blais, 2004, p. 173). Ainsi, à chaque niveau d'habileté θ correspond un score vrai variant de 0 à n ; où n est le nombre total d'items du test. Si l'on divise $\sum P_i(\theta)$ par n, puis on le multiplie par 100, on obtient le pourcentage de score vrai ou le pourcentage de contenu maîtrisé (Hambleton & Swaminathan, 1985). Avec l'exemple de la Figure 23, intéressons-nous à la situation de la moitié des items réussis pour ce test à quatre items ($n/2 = 2$). En lisant le graphique, on peut interpréter que les individus possédant une habileté $\theta = -1$ présentent 50 % de contenu maîtrisé sur le test, et plus largement, on peut

inférer que ces individus réussiraient 50 % des items dans l'univers des items d'où sont issus les items qui composent le test.

2.1.2.6. Concept d'information

Dans les MRI, on définit également une courbe d'information d'un item ou d'un test qui montre le pouvoir informatif de l'item ou du test entier tout au long de l'échelle du trait latent. Pour quelle(s) catégorie(s) d'individus, tel item ou tel test est-il le plus informatif, et donc le plus précis, pour l'évaluation de la propriété mentale ? On peut remarquer d'abord les liens entre le pouvoir informatif, le degré de difficulté et le pouvoir discriminant. Un item facile fournit peu d'information pour des valeurs de thêta élevées et positives. Tous les individus possédant une habileté élevée réussissent l'item, il n'y a donc pas de différence interindividuelle. En revanche, au sein des individus ayant une habileté faible, on peut voir une différenciation selon qu'ils réussissent ou échouent cet item facile. Une remarque similaire s'observe pour un item difficile qui donne peu d'information sur les individus possédant une habileté faible qui l'échouent tous, mais qui, en revanche, permettent de voir une différenciation au sein des individus ayant des habiletés élevées. Ainsi, la courbe d'information intègre l'information des différents paramètres estimés. L'item est plus informatif dans la zone où le degré de difficulté de l'item b_i est proche du niveau d'habileté de l'individu, ainsi que dans la zone où la discrimination a_i est élevée et enfin, plus le paramètre de pseudo-chance c_i diminue vers zéro.

Sur la Figure 24 est représenté l'exemple d'une courbe d'information de l'item. En ordonnée se trouve l'échelle des valeurs d'information et en abscisse, l'échelle du trait latent. On peut voir que le sommet de la courbe qui indique le pouvoir informatif le plus élevé se situe à $\theta = -1$ (habileté à un écart type en dessous de la moyenne). Plus globalement, l'item est particulièrement informatif pour des individus ayant une habileté entre $\theta = -2$ et $\theta = 0$.

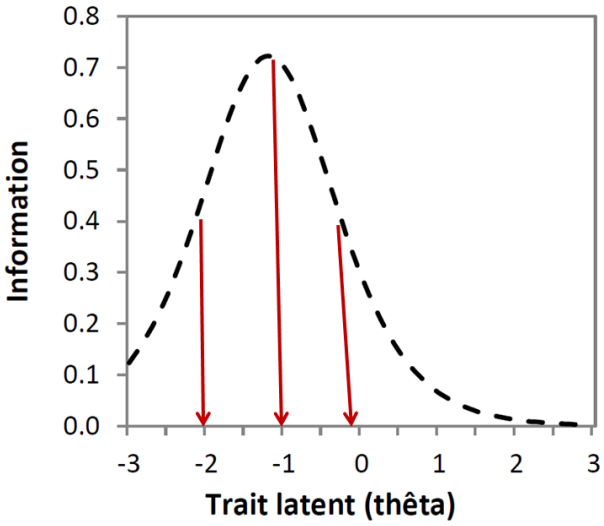


Figure 24. Courbe d'information de l'item.

La courbe d'information se définit par la fonction d'information suivante pour un modèle dichotomique à trois paramètres :

$$I_i(\theta) = D^2 a_i^2 \frac{1 - P_i(\theta)}{P_i(\theta)} \left[\frac{P_i(\theta) - c_i}{(1 - c_i)}\right]^2 \quad (8)$$

Où $I_i(\theta)$ est l'information de l'item i en fonction du trait latent, D est un facteur d'échelonnement (souvent de valeur 1.7), a_i est le paramètre de discrimination pour l'item i, $1 - P_i(\theta)$ est la probabilité d'échouer l'item i en fonction du trait latent et c_i est le paramètre de pseudo-chance pour l'item i. Au niveau du test, la valeur d'information $\sum_{i=1}^{n} I_i(\theta)$ est d'autant plus élevée que les paramètres de discrimination des items sont élevés, que le nombre d'items qui composent le test est élevé et que les paramètres de pseudo-chance sont faibles. Pour les courbes d'information du test, on peut trouver deux sommets, indiquant que le test est le plus information dans deux zones différentes de l'échelle d'habileté.

Dans les MRI, le concept d'information rejoint l'idée de précision (ou de fidélité), il y a donc un lien entre le pouvoir informatif et l'erreur type de mesure (appelée aussi erreur standard de mesure). Dans la théorie classique des tests (et également dans la théorie de la générabilisation), l'un des postulats stipule une erreur type de mesure identique pour tous les individus quelque soit leur niveau d'habileté. Les MRI

définissent, quant à eux, l'erreur type de mesure tout le long de l'échelle d'habileté des individus sur la propriété mentale évaluée selon l'équation suivante :

$$SE_i(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}} \quad (9)$$

Où $SE_i(\hat{\theta})$ est l'erreur type de mesure en fonction de l'habileté d'un individu $\hat{\theta}$. L'erreur type de mesure des MRI remplit le même rôle que l'erreur type de mesure (ETM) de la théorie classique, à la différence qu'elle varie sur l'échelle d'habileté θ des individus. L'étendue des valeurs de l'erreur type de mesure dépend de la longueur du test et des propriétés des items. Ainsi, l'erreur type de mesure diminue (1) plus le test comporte d'items, (2) lorsque le degré de difficulté de l'item est proche de l'habileté du sujet, (3) dans les zones où la discrimination est élevée, et (4) plus le paramètre de pseudo-chance diminue vers zéro (Hambleton et al., 1991).

La Figure 25 illustre la relation entre l'erreur type de mesure du test et son pouvoir informatif. En ordonnée, on a deux échelles avec différents ordres de grandeur : d'un côté la valeur d'information pour la courbe d'information (trait foncé), et de l'autre la valeur de l'erreur de mesure pour la courbe de l'erreur de mesure (trait clair). En abscisse, on a l'échelle du trait latent (ici allant de -4 à 4). Pour $\theta = 1$, le test est le plus informatif, indiquant que le test est particulièrement informatif pour les individus avec une habileté moyenne forte. De plus, on observe une relation inverse entre les deux courbes. L'erreur type de mesure est minimale dans la zone la plus informative du test à $\theta = 1$, et maximale dans une zone la moins informative à $\theta = -4$.

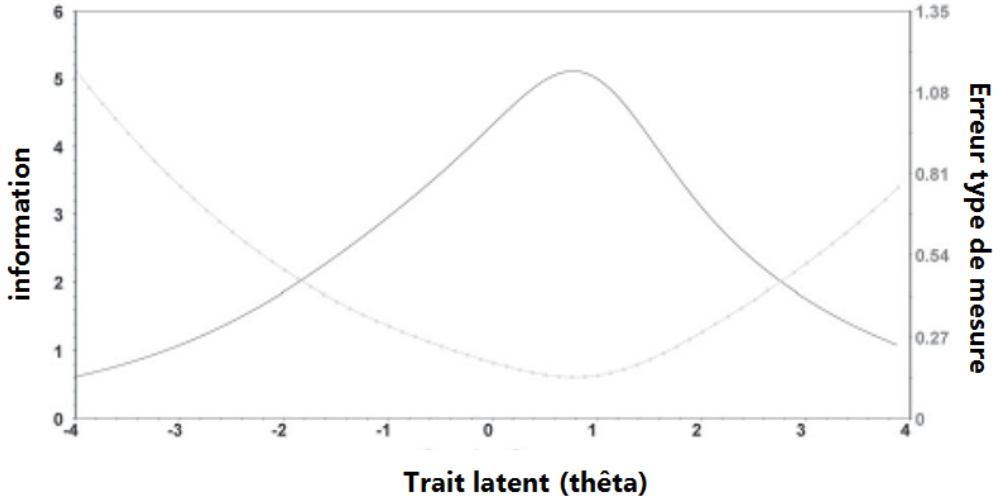


Figure 25. Relation entre les courbes d'information (trait foncé) et d'erreur standard de mesure (trait clair).

En mettant en relation les concepts d'erreur de mesure et d'information, on peut donc dire qu'à la zone où le test est le plus informatif, il est aussi le plus précis. En somme, l'erreur type de mesure dans les MRI est une mesure d'incertitude associée à la mesure.

Le concept d'information peut également renseigner sur l'information de deux (ou plus) tests, qui évaluent la même propriété mentale. Hambleton et al. (1991) parlent alors d'efficacité relative (*relative efficiency*). La comparaison des courbes d'information permet de voir quel test est le plus informatif (et donc le plus précis) pour quel niveau d'habileté des individus. L'efficacité relative à un niveau d'habileté $RE(\theta)$ est le rapport entre l'information d'un test I_A par rapport à un autre test I_B calculé selon l'équation suivante :

$$RE(\theta) = \frac{I_A(\theta)}{I_B(\theta)} \qquad (10)$$

2.1.2.7. MRI pour items polytomiques

Les analyses sur les modèles à un, deux et trois paramètres que nous venons de présenter concernent les items au format dichotomique (réponse binaire : réussi –

échoué). Les MRI proposent également des modèles pour le cas d'items au format polytomique (p. ex., le modèle des réponses nominales de Bock, le modèle à réponse graduée de Samejima, les modèles de réponse de Muraki). Il peut y avoir des items à cotation polytomique sur une échelle ordinale (opinion ou degré d'appréciation de Likert) ou sur une échelle nominale (réponse à choix multiples). Dans les subtests du WISC-IV (Cubes, Similitudes, Vocabulaire et Compréhension), la cotation polytomique des items s'apparente à une échelle ordinale allant graduellement d'échoué à parfaitement réussi. Le modèle adéquat à appliquer aux subtests du WISC-IV est par conséquent, le modèle gradué de Samejima (1997). Ce dernier est le seul modèle pour items polytomiques que nous développerons à la suite.

À l'instar des MRI pour le cas des items dichotomiques, les MRI d'items polytomiques cherchent également à modéliser la probabilité de fournir une réponse correcte (ou une certaine appréciation) selon le trait latent θ. Cependant, au lieu d'une seule fonction et d'une seule courbe caractéristique pour chaque item, l'analyse fournit une fonction et une courbe distinctes pour chacune des options de l'item. De ce fait, on doit préciser à quelle option (modalité, catégorie) correspond la courbe caractéristique de l'item.

La Figure 26 (p. 104) illustre le cas d'un item avec quatre options de réponse dont une seule réponse est correcte (échelle nominale). Dans cet exemple, la courbe caractéristique de l'option A désigne la bonne réponse, tandis que les autres courbes caractéristiques se rapportent aux options de réponse incorrecte (distracteur, leurre). Le comportement de la courbe caractéristique de l'option A montre que la probabilité de choisir cette option augmente avec le niveau d'habileté des individus. De même, plus le niveau d'habileté augmente, plus les probabilités de choisir les options de réponse incorrecte B, C et D diminuent. Si on compare les courbes caractéristiques de l'option A et C, on observe un point de basculement à $\theta = 0$. Les individus ayant une habileté en dessous de la moyenne ($\theta < 0$) ont une probabilité beaucoup plus élevée de choisir l'option C que A. Par contre, les individus ayant une habileté au-dessus de la moyenne ($\theta > 0$) ont une probabilité plus forte de choisir l'option A que C. Les options B et D sont moins populaires que les options A et C, et de plus, elles ne sont choisies que par les individus possédant des niveaux d'habileté faible ($\theta < 0$). Au trait latent $\theta = 0$, la probabilité de choisir les options A et C sont de 50 %, tandis que les probabilités sont de 10 % pour l'option D et quasi nulles pour l'option B. Toutes choses étant égales par ailleurs, on peut donc considérer les individus qui choisissent les options incorrectes B et D comme moins habiles que les individus qui choisissent l'option aussi incorrecte C.

Les courbes caractéristiques des options incorrectes de l'item donnent de l'information sur les leurres, et notamment sur les individus qui les choisissent. L'information concerne surtout le groupe des individus les moins habiles qui sont plus sujets à être influencés par les distracteurs que les individus les plus habiles sur le trait latent.

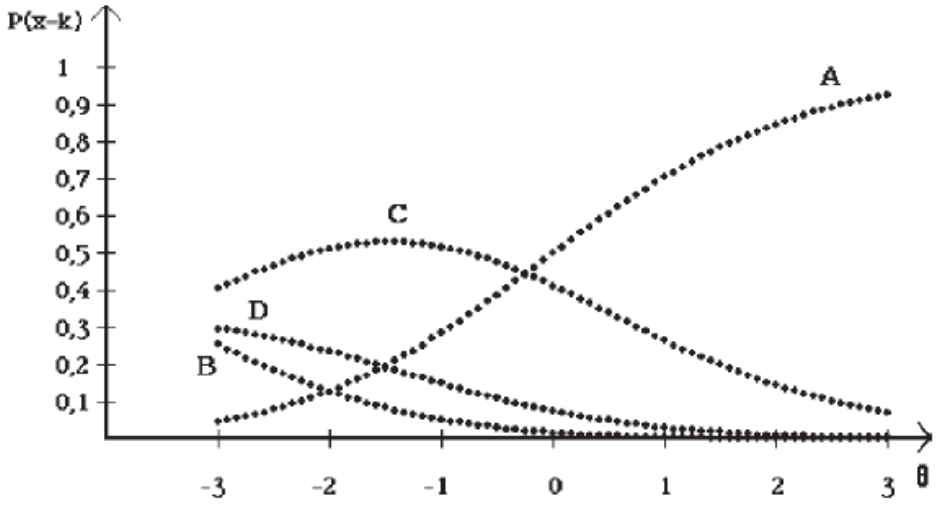


Figure 26. Courbe caractéristique pour chacune des quatre options de l'item. *Source* : Bertrand et Blais (2004, p. 157).

Dans le cas d'items polytomiques sur une échelle ordinale, le modèle gradué de Samejima est fréquemment utilisé. Ce modèle s'accommode d'un nombre d'options de réponse variable d'un item à l'autre. Les fonctions des courbes caractéristiques suivant ce modèle se définissent selon l'équation suivante :

$$P_i^*(k/\theta) = \frac{1}{1 - e^{-Da_i(\theta - b_{ik})}} \qquad (11)$$

Où $P_i^*(k/\theta)$ est la probabilité de choisir une des options supérieures à la k-ième de l'item i pour un individu possédant un certain degré d'habileté θ, k est une des options que l'item comporte (pouvant valoir 0, 1, 2, … $m_i - 1$), e est la constante de Neper (valant env. 2.71828), D est le facteur d'échelonnement, a_i est le paramètre de discrimination pour l'item i, et b_{ik} est le paramètre de localisation pour l'item i. Dans ce modèle, chaque item i est caractérisé par un seul paramètre de discrimination a_i (*slope*) – qui est donc une indication générale de la pente de toutes les courbes caractéristiques des options de l'item – et autant de valeurs de paramètres de localisation b_{ik} (*threshold*) qu'il y a d'options pour l'item moins une. La valeur du

paramètre de localisation b_{ik} représente le point sur l'échelle du trait latent où la probabilité de choisir l'option k et les options au-dessus, est supérieure à 50 %. Par exemple, si on prend le subtest Vocabulaire du WISC-IV. Il est possible d'obtenir 0 (échoué), 1 (partiellement réussi) et 2 points (parfaitement réussi). Le modèle gradué de Samejima nous fournit pour chaque item de Vocaublaire trois courbes caractéristiques, soit une pour chaque gradation de réponse (0, 1 et 2 points). Un seul paramètre de discrimination a_i renseigne sur la discrimination de l'item. En revanche, deux valeurs de paramètres de localisation b_{ik} sont à disposition. La valeur du paramètre de localisation b_{i1} renseigne sur le niveau d'habileté pour lequel la probabilité d'obtenir 1 point ou plus est supérieure à 50 % et le paramètre de localisation b_{i2} renseigne sur le niveau d'habileté pour lequel la probabilité d'obtenir 2 points est supérieure à 50 %. Plus les valeurs des paramètres de localisation b_{ik} sont faibles, plus le seuil sur le trait latent pour lequel la probabilité de choisir l'option qui correspond au maximum de point, est bas.

2.1.2.8. Estimation des paramètres

Jusqu'ici, nous avons présenté différentes équations de fonctions caractéristique de l'item, caractéristique du test et d'information. Dans ces équations, les paramètres de difficulté b_i, de discrimination a_i ou de pseudo-chance c_i doivent être estimés (ou fixés). Connaissant leur valeur, il est simple de les injecter dans les équations présentées, mais comment ces paramètres sont-ils estimés ? A nouveau, nous n'entrerons pas dans des détails techniques, nous restons à une compréhension conceptuelle de la procédure. Au départ, les probabilités de réussir l'item en fonction de l'habileté et des caractéristiques de l'item ne sont pas connues. On ne dispose que d'un large échantillon de réponses obtenues sur un échantillon d'items. À partir des réponses aux items par les sujets de l'échantillon, la démarche itérative d'estimation se réalise « *by incorporating information about items into the ability-estimation process and by incorporating information about the examinees abilities into item-parameter-estimation process* » (Hambleton et al., 1991, p. 8). Pour des modélisations polytomiques ou multidimensionnelles, le nombre de paramètres librement et simultanément estimés est important, ce qui complexifie la procédure d'estimation. D'ailleurs, le principal problème auquel on est confronté dans une procédure itérative aussi complexe, est celui de la non-convergence des estimations du fait qu'il y ait trop de paramètres librement estimés. La précision de l'estimation des paramètres va

dépendre de plusieurs facteurs : quel modèle de réponse à l'item est choisi, combien de paramètres sont estimés pour l'item, quel est l'ajustement des scores au critère d'unidimensionnalité, combien d'items comporte le test et quelle est la taille d'échantillon (Harris, 1989). Il existe plusieurs méthodes statistiques d'estimation (méthode du maximum de vraisemblance, méthode Newton-Raphson, méthode bayésienne du maximum marginal de vraisemblance, etc.) dont le choix est souvent restreint à celles réalisables par le logiciel utilisé pour les analyses. Concrètement, la procédure d'estimation est très technique et demande une familiarité avec les possibilités et les limites des programmes informatiques qui effectuent les analyses.

2.1.2.9. Conditions d'application des MRI

À la différence de la théorie classique, les MRI modélisent « la relation entre la probabilité pour un sujet de réussir un item (et non plus la fréquence de réussite dans un groupe) et sa position sur une variable latente (et non plus son score sur une variable observable) » (Huteau & Lautrey, 2003, p. 87). Pour estimer de tels modèles, cela exige un échantillon de sujets et d'items d'un test très importants. Avec un paramètre librement estimé, le modèle de Rasch est le plus commode au niveau du critère du nombre de données suffisantes. En effet, ce modèle nécessite une taille d'échantillon entre 100-200 sujets et une longueur de test habituel. Pour les modèles à deux paramètres, il faudrait disposer d'un test avec une longueur d'au moins 30 items et 500 sujets (Hulin, Lissak, & Drasgow, 1982). Quant aux modèles à trois paramètres, il s'agit de disposer d'un test avec une longueur d'au moins 50 items et de 1000 sujets (Hulin et al., 1982). Le nombre d'items d'un test peut difficilement être augmenté, en revanche, il y a un peu plus de souplesse pour augmenter le nombre de sujets dans une recherche. Pour des analyses à l'aide des MRI, cela demande donc de recruter un nombre très important de sujets afin d'aider à l'estimation des paramètres.

Hormis les critères pratiques de taille d'échantillon et d'items du test, au niveau du cadre théorique, les MRI sont applicables lorsque trois conditions sont satisfaites : la propriété d'invariance ainsi que deux concepts étroitement liés, la propriété d'indépendance locale et la propriété d'unidimensionnalité. Les hypothèses d'indépendance locale et d'unidimensionnalité ne sont pas propres à la théorie de réponse à l'item, mais relèvent plus largement de l'utilisation même des tests. Théoriquement, un test regroupe un ensemble d'items qui contribuent à évaluer une seule propriété mentale et qui ne devraient pas se biaiser les uns les autres. En

préambule à toutes analyses dans le cadre de la théorie de réponse à l'item, il s'agit donc de s'assurer de l'adéquation du contexte d'application.

La propriété d'invariance postule que les estimations liées aux items (paramètres de difficulté, de discrimination, etc.) et aux individus (niveau d'habileté sur la propriété mentale évaluée) sont indépendantes de l'échantillon particulier d'individus ou de l'échantillon d'items à partir duquel elles sont réalisées. En fait, « la propriété d'invariance est principalement ce qui permet à la TRI d'étaler sa supériorité par rapport aux autres propositions de modélisation, comme la théorie classique des tests ou la théorie de la généralisabilité » (Bertrand & Blais, 2004, p. 187). Nous avons déjà mentionné le problème de dépendance circulaire dans la théorie classique. En effet, l'estimation des indices (p-indice, d-indice, coefficient de fidélité, etc.) est dépendante à la fois de la distribution des habiletés dans un échantillon particulier et du set d'items d'un test qui sont utilisés pour les estimer. Ainsi, la généralisation des résultats à d'autres échantillons n'est pas de facto et doit être soutenue par une étude sur chaque échantillon. Dans les MRI, « les estimations du ou des paramètres associés aux items sont indépendantes du groupe de sujets qui est la cible de l'opération de mesure et les estimations du ou des paramètres associés aux sujets sont indépendantes du groupe d'items inclus dans l'opération de mesure » (Bertrand & Blais, 2004, p. 183). La Figure 27 (p. 108) illustre la propriété d'invariance qu'assurent les MRI. Dans cette figure est représentée la distribution des habiletés de deux groupes de sujets (groupe 1 et groupe 2). Les individus du groupe 1 sont en moyenne moins habiles que les individus du groupe 2. Malgré la différence entre la performance moyenne des deux groupes, le modèle de réponse à l'item donne une seule et même CCI pour l'item i. Ainsi, les individus possédant une même habileté sur le trait latent ont la même probabilité de réussir l'item quelque soit les caractéristiques du groupe dont ils sont issus.

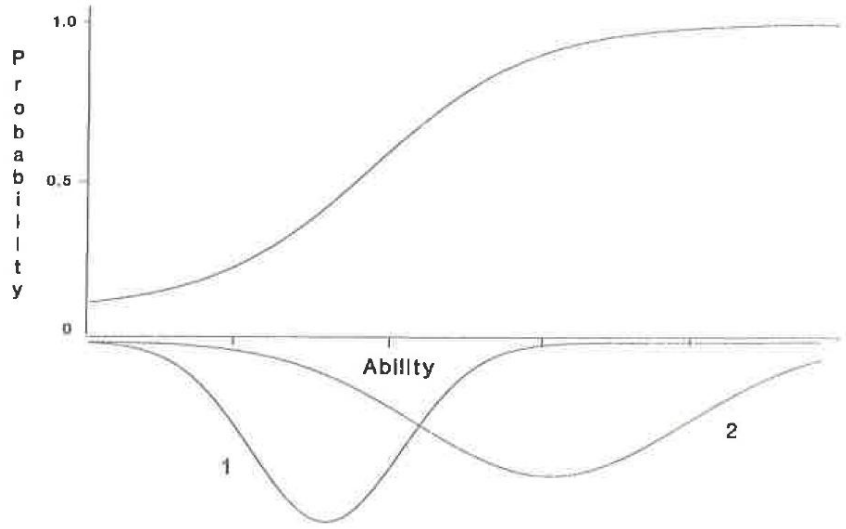


Figure 27. Courbe caractéristique de l'item i et distributions d'habileté de deux groupes (1 et 2).

Source : Hambleton et al. (Hambleton et al., 1991, p. 8).

Chaque item d'un test est conçu pour ne pas apporter d'information qui orienterait le sujet sur la réponse à un autre item. Il s'agit de la condition d'indépendance locale des items. Selon cette condition, la performance (réussite ou échec) sur un item d'un test n'influence pas la performance sur les autres items du test. Seul le trait latent évalué (niveau d'habileté sur la propriété mentale) explique la performance de l'individu sur les items du test. Sur le plan statistique, cela signifie que les corrélations entre les réponses aux items d'un test sont nulles pour une valeur de trait latent fixée comme le soulignent Lord et Novick :

> La performance d'un individu dépend du seul trait si, étant donné la valeur observée pour ce trait, rien d'autre ne peut contribuer à nous informer sur la performance au test. Le trait latent est le seul facteur important et, lorsque la position de l'individu sur l'échelle de ce trait latent est connue, le comportement est aléatoire, au sens de l'indépendance statistique. (1968, cités par Bertrand & Blais, 2004, pp. 201–202)

Ainsi, le respect de la condition d'indépendance locale des items assure la validité des estimations liées aux individus (échelle des thêta).

En étroit lien avec l'indépendance locale, l'unidimensionnalité est l'un des postulats importants sur lesquels reposent les MRI, ainsi que le fondement de l'utilisation des tests. Pourtant, rares sont les situations réelles de testing qui produisent des données en accord avec ce présupposé. Relevons d'abord qu'il n'y a pas de

véritable définition opérationnelle et concrète du terme d'unidimensionnalité ni de consensus sur les méthodes pour évaluer sa présence ou son absence dans un ensemble d'items (Hambleton & Rovinelli, 1986; Hattie, 1985; McDonald, 1981). On définit un ensemble d'items comme étant unidimensionnel dès lors que ceux-ci contribuent à évaluer une seule et même dimension (propriété mentale, attribut psychologique) qui sous-tend la performance du sujet au test. Cette définition plutôt générale entretient une certaine confusion dans le choix de méthodes appropriées pour l'évaluation de l'unidimensionnalité. Par exemple, certains commettent l'erreur d'interpréter l'alpha de Cronbach comme un bon indice de l'unidimensionnalité d'un test. Or le calcul d'un alpha de Cronbach présuppose l'unidimensionnalité des items pour être pertinent. Il s'agit d'un coefficient qui évalue le degré de covariation entre les items d'un test. Des items unidimensionnels conduisent à une valeur élevée pour l'alpha de Cronbach, toutefois, une valeur élevée de ce coefficient ne garantit pas l'unidimensionnalité des items. Selon Hattie (1985) :

> *Alpha can be high even if there is no general factor since (1) it is influenced by the number of items and parallel repetitions of items, (2) it is increases as the number of factors pertaining to each item incrases, and (3) it decreass moderately as the item communalities increase.* (p. 158)

Hattie dénombre plus d'une trentaine d'indices utiliser pour évaluer l'unidimensionnalité, qu'il regroupe sous cinq catégories de méthodes : (1) basées sur les patterns de réponse (p. ex., indice d'homogénéité de Loevinger), (2) sur la fidélité (p. ex., alpha de Cronbach ou Kuder-Richardson 20), (3) sur l'analyse en composantes principales (p. ex., part de variance expliquée par la première composante, nombre de valeurs propres >1), (4) sur les analyses factorielles (p. ex., indice omega, analyse factorielle non linéaire) et (5) sur les modèles en traits latents (1985, pp. 141–142, Table 1). Les techniques des analyses factorielles non linéaires, sont parmi les plus recommandées comme le suggère McDonald :

> It is reasonable to assert that a set of *n* tests or a set of *n* binary items is unidemensionnal if and only if it fits a non-linear factor model with one common factor. (1981, pp. 104–105)

Implicitement exigée pour l'utilisation des tests, l'unidimensionnalité est très importante dans la clinique. En effet, l'interprétation qui est réalisée sur un test repose sur la propriété mentale que le test est censé évaluer (c.-à-d. sur la variance pertinente). Si les items du test évaluent plusieurs traits, plusieurs habiletés, l'interprétation d'un test et la mise en relation des résultats à différents tests deviennent très délicates, voire

divinatoires. En effet, comment déterminer à quelles propriétés mentales, le sujet a eu recours, et dans quelle mesure chacune a contribué au score ? Si le test n'est pas unidimensionnel, obtenir des scores identiques au test ne reflèterait pas nécessairement une même habileté pour deux individus.

Dans le sens strict, aucun test ne peut prétendre à la propriété d'unidimensionnalité. Nous l'avons mentionné, les parts de variance d'un score de test se répartissent en variance pertinente (ce qu'est censé évaluer le test), mais aussi en variance non pertinente (ce qui intervient dans la performance, mais qui n'entre pas dans l'interprétation de base du test) et en variance d'erreur (ce qui affecte aléatoirement la mesure). Évidemment, on cherche à construire des tests qui maximalisent la part de variance pertinente, toutefois, il est impossible de réduire à zéro les parts de variance non pertinente et d'erreur. Jusqu'à un certain point, il y aura toujours plusieurs éléments en jeu qui contribuent à la performance sur un test.

Face à la complexité des situations réelles qui ne permettent pas d'éliminer la variance non pertinente et l'erreur de mesure, un assouplissement de l'hypothèse d'unidimensionnalité est accepté si on identifie une dimension dominante qui explique la performance et les réponses au test (Bertrand & Blais, 2004; Frenette et al., 2007). De plus, des auteurs préconisent, au lieu de se demander : « est-ce qu'un test est unidimensionnel ou non ? », de reformuler la question ainsi : « est-ce qu'il y a des critères de décision qui nous permet de déterminer à quel degré un ensemble d'items est proche de l'unidimensionnalité ? » (Hattie, 1985). Les analyses sur les données récoltées évaluent une unidimensionnalité sur un plan statistique ; le psychologue et les concepteurs des tests doivent réfléchir à leur position théorique pour évaluer l'unidimensionnalité aussi sur un plan conceptuel. Par exemple, un test de raisonnement arithmétique évalue-t-il une ou plusieurs dimensions ?

Les conditions d'application des MRI ne sont pas nombreuses ni plus exigeantes que la théorie classique. Toutefois, au vu d'une demande moins importante en taille d'échantillon et en longueur de test, les méthodes développées par la TCT sont les plus usuelles en comparaison des MRI. Dans le cadre du présent travail, les MRI sont utilisés pour l'étude du fonctionnement différentiel des items. Cette analyse est en lien avec les préoccupations autour de la validité de l'évaluation et, plus spécifiquement, de son équité pour tous les individus. Nous développerons sur l'équité de l'évaluation dans la dernière partie du présent chapitre. Avant cela, un rappel des qualités métrologiques recherchées pour un test permettra d'expliciter le sens de ces notions qui parsèment de façon récurrente le présent travail. La validité est un des concepts centraux en

psychométrie, néanmoins dans le présent travail, elle ne sera que brièvement abordée. Quant à la fidélité, elle fera l'objet d'un chapitre à part étant notre sujet principal.

2.2. HOMOGÉNÉITÉ

L'évaluation de l'homogénéité permet de répondre à la question suivante : est-ce que tous les items d'un test évaluent la même propriété mentale qu'ils sont censés mesurer ? Un test regroupe un ensemble limité d'items pour évaluer la propriété mentale pour lequel il est construit. Les items d'un test sont sélectionnés en fonction d'un certain nombre de critères (p. ex., difficulté, longueur souhaitée du test). L'homogénéité des items est une des propriétés métrologiques importantes.

Une fois qu'un test est administré, un score total est calculé à partir des points obtenus sur les items réussis. Le score total représente une estimation du niveau du sujet sur la propriété mentale. Pour justifier cette inférence, les items qui composent le test doivent être homogènes et ainsi tous contribuer dans le même sens à l'estimation de la propriété mentale. S'il s'agit d'une batterie de tests, l'homogénéité s'évalue aussi entre les scores des subtests qui contribuent à un même indice.

Le terme d'homogénéité est souvent utilisé en synonyme d'unidimensionnalité, on l'évalue donc avec les mêmes méthodes mentionnées dans la précédente discussion sur l'unidimensionnalité. Toutefois, certains auteurs donnent une nuance spécifique à ce terme, qui est alors utilisé pour une situation d'égalité des intercorrélations entre items. Dans ce sens, « *a perfectly homogeneous test is one in which all the items intercorrelate equally. That is, the items all measure the construct or constructs equally* » (Hattie, 1985, p. 157). Dans une telle situation, on risque d'avoir un test qui évalue un champ très restreint, avec des items qui ne sont que des reformulations les uns des autres. Utilisée dans ce sens précis, une trop grande homogénéité des items d'un test n'est pas souhaitable (R. B. Cattell & Tsujioka, 1964).

2.3. SENSIBILITÉ

L'évaluation de la sensibilité d'un item (ou du score total du test) répond à la question suivante : est-ce que les items (le test) permettent de discriminer les individus les uns par rapport aux autres ? Un test sert à mettre en lumière des différences

interindividuelles. En effet, à partir des performances/réponses du sujet, on cherche à le situer par rapport aux performances/réponses d'autres individus de son groupe de référence. Les items ont un pouvoir discriminant élevé s'ils permettent de finement situer les individus selon leur performance au test. Un item que tout le monde réussit/échoue n'aide pas à interpréter l'habileté du sujet, sa sensibilité est nulle et l'item est inutile. En revanche, une faible (et non nulle) discrimination ne pose pas problème si l'item sert à discriminer à l'intérieur des individus très faibles ou très forts. Par exemple, les items du début et de la fin du test ont une faible sensibilité, puisqu'ils sont très faciles (pour mettre à l'aise le sujet) ou très difficiles (pour discriminer au sein des très forts).

Nous les avons mentionnés, la théorie classique calcule un indice de difficulté (p-indice) et un indice de discrimination (D-indice). Par rappel, le D-indice résulte de la différence des p-indices entre deux groupes extrêmes d'un échantillon (c.-à-d. les ~30 % les plus faibles vs les ~30 % les plus forts). Plus il y a une grande différence dans la proportion de réussite chez le groupe des forts et chez le groupe des faibles, plus le D-indice est élevé et l'item discrimine bien les individus des deux groupes. Il y a une relation entre la difficulté de l'item et sa discrimination : plus un item est facile ou difficile, moins il est discriminant. Les meilleurs items sont ceux avec des p-indices autour de .50 et un D-indice ≥ .40.

L'évaluation de la sensibilité n'est pas seulement importante dans la sélection des items au moment de la construction d'un test. On en tient compte également dans la formulation d'hypothèses lors de l'interprétation des résultats à un test. On examine alors l'étendue des scores. Par exemple, pour les subtests du WISC-IV, l'étendue théorique des notes standards est de 1 (performance extrêmement faible) à 19 (performance extrêmement élevée). Sachant que la moyenne de la distribution des notes standards est de 10 et son écart type est de 3, on peut discriminer les performances entre -3 et +3 écarts types. La Figure 28 (p. 113) présente la table de conversion des notes brutes en notes standards pour la tranche d'âge des 16 ans 8 mois à 16 ans 11 mois.

Table A.1 Conversion des notes brutes totales aux subtests en notes standard, par groupe d'âge (*suite*)

Ages 16:8–16:11																
Notes standard	CUB	SIM	MCH	IDC	COD	VOC	SLC	MAT	COM	SYM	Notes standard	CIM	BAR	INF	ARI	RVB
1	0-36	0-16	0-9	0-11	0-38	0-31	0-14	0-17	0-16	0-16	1	0-18	0-61	0-12	0-19	0-11
2	37	17	10	12	39-41	32	15	18	17	17-18	2	19	62-64	13	20	12
3	38	18-19	11	13	42-44	33	16	19	18-19	19	3	20	65-70	14	21	13
4	39	20	12	14-15	45-47	34	17	20	20-21	20-21	4	21-22	71-74	15	22	14
5	40	21	13	16	48-50	35-36	18	21	22-23	22-23	5	23-24	75-77	16-17	23	15
6	41-43	22-23	14	17	51-54	37-38	-	22	24-25	24-26	6	25-26	78-80	18	24	16
7	44-46	24	15	18	55-59	39-41	19	23	26-27	27	7	27	81-86	19	25	17
8	47-48	25-26	16	19	60-64	42-43	20	24	28-29	28-29	8	28	87-93	20	-	-
9	49-51	27-28	17-18	20	65-68	44-45	-	25-26	30	30-32	9	29	94-99	21-22	26	18
10	52-53	29	19-20	21	69-73	46-47	21	27	31	33-35	10	30	100-108	23-24	27	19
11	54-56	30-31	21-22	22	74-78	48	22	28-29	32	36-37	11	31-32	109-115	25	28	20
12	57-58	32	23-24	23-24	79-83	49	23	30	33	38	12	33	116-120	26	29	21
13	59-61	33-34	25-26	25	84-86	50-51	24	31	34	39	13	-	121-123	27	30	22
14	62-64	35	27-28	26-27	87-89	52-53	25	32	35	40	14	34	124-125	28	31	23
15	65-66	36	29-30	28	90-93	54-56	26	33	36	41-42	15	35	126	29	32	24
16	67	37	31	-	94-96	57-58	27	34	37	43	16	36	127	30	33	-
17	68	38	32	-	97-99	59	28	35	38	44	17	37	128	31	34	-
18	-	39	-	-	100-102	60	29	-	39	45	18	38	129	32	-	-
19	-	40-44	-	-	103-119	61-68	30	-	40-42	46-60	19	-	130-136	33	-	-

Figure 28. Table de conversion des notes aux subtests du WISC-IV pour les 16 ans 8 mois à 16 ans et 11 mois. *Source* : Manuel d'administration et de cotation (Wechsler, 2005a, p. 236).

On peut voir parmi les subtests obligatoires un manque de sensibilité de Cubes (CUB), Mémoire des chiffres (MCH), Identification de concepts (IDC), Séquence lettres-chiffres (SLC) et Matrices (MAT). Le manque de sensibilité se traduit par une étendue plus restreinte des notes standards pour les performances élevées. Par exemple, un sujet qui obtient la note brute maximale de 28 points au subtest IDC a une note standard de 15, et non de 19. Cela indique que, pour ce groupe d'âge, les items les plus difficiles de ce subtest ne permettent pas de discriminer finement au sein des individus les plus forts, qui à partir d'une certaine habileté les réussissent tous.

2.4. STANDARDISATION – ÉTALONNAGES

Un test est administré selon une procédure standardisée. Le psychologue doit donner les consignes au plus près de celles fournies par le manuel et toujours en respecter le sens, ce qui ne signifie pas les énoncer de manière stéréotypée ou mécanique. Les conditions de passation (p. ex., salle lumineuse et calme, individuelle vs collective), le setting (p. ex., disposition spatiale, matériel, chronométrage) ainsi que les

règles de cotation sont également à respecter. La standardisation est essentielle pour permettre une comparaison entre la performance d'un sujet avec les performances de son groupe de référence. Elle contrôle que tous les individus soient placés dans des conditions comparables. Elle limite les erreurs de mesure et les biais, mais ne peut pas les supprimer complètement. Grâce à la standardisation, on peut supposer que, « toutes choses étant égales par ailleurs », les différences interindividuelles observées sur les résultats du test sont des différences sur la propriété mentale évaluée.

Dans la définition d'un test, l'interprétation des scores se réalise dans la comparaison du sujet et de son groupe de référence. En effet, les scores bruts s'étendent selon un nombre de points définis par la longueur du test. Ils n'ont pas de valeur universelle. L'interprétation des scores bruts d'un test dépend des caractéristiques de leur distribution dans la population (moyenne, dispersion, fréquence). Généralement, la distribution des scores bruts d'un test psychologique s'approche de la loi normale. L'étalonnage d'un test permet alors de construire une échelle de référence (de moyenne et d'écart type connus) à partir de la distribution des scores issus d'un échantillon représentatif de la population à qui s'adresse le test. La population d'étalonnage sert à constituer des normes. Généralement, les tests cognitifs proposent des normes pour différents groupes d'âge et différents groupes cliniques, tandis que les tests de personnalité proposent des normes différentes entre hommes et femmes.

Comme le score d'un test ne renseigne pas sur le niveau absolu du sujet sur une propriété mentale, mais sur un score relatif qui le situe par rapport à son groupe de référence, il est important se référer aux normes adéquates et récentes. Par exemple, on ne peut pas utiliser des normes américaines, si on utilise une adaptation en français du test. Les normes des tests utilisés en Suisse Romande sont pour la grande part issues d'un étalonnage sur la population française. Des recherches sont nécessaires pour explorer le degré de comparabilité entre les individus de différentes populations, puisque le coût de constitution d'un échantillon d'étalonnage n'encourage pas les maisons d'édition à établir des normes pour différentes populations.

La révision des normes d'un test permet de réétalonner sur un échantillon relativement contemporain au sujet évalué afin que la comparaison entre eux soit pertinente. Au fil du temps, les caractéristiques d'une population évoluent, changent, se modifient tant et si bien qu'elle s'éloigne progressivement de la population des normes établies. Un tel phénomène d'obsolescence des normes pour les tests d'aptitudes est mis en évidence à la suite des travaux publiés de James Robert Flynn (1984). Sur une

période de quarante-six ans (1932 à 1978), ce dernier constate que, le niveau intellectuel moyen de la population américaine augmente de 13.8 points de QI (évalué avec différentes révisions du Stanford-Binet et des échelles de Wechsler), soit un gain annuel de 0.30 point de QI. En référence à cet auteur, on nomme « effet Flynn » l'augmentation observée des scores aux tests cognitifs pour des individus qui ont passé deux versions d'un test sur plusieurs années d'écart. Des travaux sur d'autres cohortes se sont intéressés aux différences de gains entre les domaines (raisonnement verbal, raisonnement visuospatial, aptitude verbale, aptitude numérique, etc.), entre les hommes et les femmes, entre des groupes de différents niveaux de QI. Dans les travaux qui explorent des facteurs explicatifs de l'effet Flynn, l'une des hypothèses est les progrès dans l'éducation et l'accès aux connaissances pour une large frange de la population. L'interprétation de l'effet Flynn suscite bien des débats, cependant, il a une conséquence dans l'évaluation de l'intelligence au moyen d'un test : la tendance à surestimer la performance du sujet à mesure du vieillissement des normes. Pour assurer la pertinence de la comparaison entre le sujet testé et les individus qui ont constitué les normes, des réétalonnages doivent être proposés régulièrement (environ tous les dix ans). La révision des normes s'intègre généralement dans une révision globale du test. En effet, les items doivent également être actualisés pour tenir compte des changements culturels, technologiques et sociétaux. À l'heure actuelle, de plus en plus de tests sont proposés sur des supports informatiques et numériques (p. ex., tablette).

2.5. VALIDITÉ

Considérée comme le concept le plus fondamental et le plus important de la psychométrie, la validité de l'interprétation des scores d'un test représente le grand défi des concepteurs de tests comme en témoigne Kelly (1927) : « *the establishment of the fact that a given test is valid for specifically named purpose is at present one of the most, if not in fact the most, difficult of the problems confronting the test deviser* » (pp. 30-31). Dans sa définition la plus classique, le concept de validité renvoie au « *degree to which a test or examination measures what it purports to measure* » (Ruch, 1924, cité par P. E. Newton, 2012, p. 3). La validité renseigne sur ce que le test évalue et dans quelle mesure cette évaluation permet d'interpréter les différences interindividuelles dans les scores comme des différences réelles sur la propriété mentale censée être évaluée par le test. L'évaluation de la validité ne repose pas sur une seule analyse, mais combine les analyses provenant de différentes sources. Chaque résultat qui contribue à

définir la validité de l'interprétation des scores d'un test est une preuve de validité. On recourt à une démarche tant quantitative (p. ex., analyses statistiques) que qualitative (p. ex., évaluation écrite sur les items par des experts) pour obtenir une preuve de validité. Une preuve de validité apportée par une étude sur un échantillon ne peut pas être automatique généralisée à d'autres contextes d'utilisation du test et à d'autres populations. En effet, « la validité d'un test doit être établie empiriquement pour chacun des usages auxquels le test est destiné » (Anastasi, 1994, p. 131). Par exemple, lorsqu'un test est adapté dans une autre langue, on ne peut pas s'appuyer sur ce qui a été démontré dans les recherches de validation de la version d'origine. Étant donné la proximité des termes, définissons ce qui distingue la validité de la validation.

Le concept de validité renvoie à une propriété. La validité renseigne dans quelle mesure le test évalue de façon appropriée ce qu'il est censé évaluer ainsi que dans quelle mesure les inférences à partir des résultats au test sont pertinentes. Dans le présent travail, notre définition de la validité est en termes de proportion de la variance pertinente dans la variance totale des scores observés[11]. Pour rappel, le score d'un test peut se décomposer en variance pertinente (la propriété mentale qu'il évalue et sur laquelle porte l'interprétation), en variance non pertinente (tout ce qu'il évalue d'autre, qui n'est pas la propriété mentale sur laquelle porte l'interprétation du test) et en variance d'erreur. La proportion de la variance pertinente se rapporte à la validité, tandis que la proportion de la variance pertinente et de la variance non pertinente se rapportent à la fidélité.

Quant au concept de validation, il renvoie à une activité de développement de méthodes et d'accumulation de preuves empiriques de validité de l'évaluation pour un test et pour telle finalité ou tel contexte d'utilisation. La validation d'un test est un long processus d'enrichissement des données empiriques de validité. Les études de validation sont notamment guidées par l'exploration des différentes inférences liées à l'utilisation des tests psychologiques. Du recueil de la réponse sur un item de test à la décision d'intervention, on réalise quatre types d'inférences d'après Kane (2006). La première inférence se réalise dans le passage d'une réponse ou d'un comportement sur des items à un nombre de points, qui sont ensuite transformés en un score standardisé (*scoring inferences*). L'interprétation du score total au test en termes de propriété mentale constitue la deuxième inférence (*generalisation inferences*). Le score total au test est considéré comme une estimation du niveau d'habileté du sujet sur la propriété mentale qu'évalue le test. À partir d'un échantillon de réponses/comportements sur des

[11] Voir section 2.1.1.1, p. 73.

items, on calcule un score total qu'on traduit comme un indicateur du fonctionnement de l'individu sur une propriété mentale, qui n'est en fait pas directement observée. À partir de l'estimation du niveau de fonctionnement sur la propriété mentale évaluée, une troisième inférence réalise une extrapolation sur des difficultés ou des forces dans des activités associées dans différentes sphères (p. ex., scolaire, professionnelle, sociale) de l'individu (*extrapolation inferences*). La quatrième et dernière inférence apparaît dans les utilisations ou les interprétations qui dépassent ce que permet le test (*decision rules inferences*). Bien souvent les concepteurs de tests sont peu clairs sur les limites de leur test et laissent la responsabilité aux utilisateurs d'évaluer le bon usage du test. Toutes ces inférences, sur lesquelles repose l'utilisation d'un test, soulignent l'importance d'apporter des preuves de validité.

Pour certains auteurs (p. ex., Messick), un test n'est pas valide ou non dans l'absolue ; mais il présente un certain degré de validité par rapport à une utilisation spécifique. Selon cette position, il s'agit d'un abus de langage de parler de « validité d'un test ». La validité est une propriété de l'interprétation des scores du test, et non du test lui-même. En effet, « on ne valide [pas] un instrument de mesure mais les mesures qu'il permet d'obtenir . . ., celles-ci dépendent non seulement des caractéristiques de l'instrument, mais aussi des caractéristiques des sujets auxquels cet instrument est appliqué et du contexte dans lequel il est utilisé » (Dickes et al., 1994, p. 49). Pour d'autres auteurs (p. ex., Borsboom), il s'agit bien de parler de la validité d'un test. Pour ces auteurs, le score du test est déterminé par la propriété mentale évaluée par le test (modèle réflectif). La propriété mentale évaluée – et qui existe – est à l'origine du score sur le test, et donc les variations dans la propriété mentale amènent à des différences dans les scores au test. Ainsi, « *validity is a property of tests: a valid test can convey the effect of variation in the attribute one intends to measure* » (Borsboom, Mellenbergh, & van Heerden, 2004, p. 1067). Pour d'autres auteurs encore (p. ex., Newton), la validité porte sur l'ensemble de la procédure d'évaluation qui amène à une prise de décision (*property of assessment-based descision-making procedure*, P. E. Newton, 2012, p. 18). Il ne s'agit pas uniquement de déterminer si les interprétations des scores du test sont valides ou non, mais si l'instrument choisi et son utilisation dans l'évaluation réalisée sont appropriés. Entre les tenants de ces trois conceptions de la validité, le débat est actuellement en cours et alimente en fait des discussions passées. En effet, si l'on retrace l'histoire des discussions sur la validité, on s'aperçoit que le concept connaît régulièrement une redéfinition. Bien qu'intéressante, un historique complet de l'évolution du concept de validité dépasse le cadre du présent travail (lire p. ex.,

Borsboom et al., 2004; Cronbach & Meehl, 1955; Kane, 2001; Messick, 1989; P. E. Newton, 2012). Pour brièvement résumer, dans les premières définitions de la validité au début du 20e siècle, il s'agit d'une propriété du test : la validité examine dans quelle mesure un test évalue ce qu'il prétend évaluer. Par la suite, la relation entre le test et un critère est au centre de la validité. À travers la convergence des résultats sur le test et un critère, on cherche non seulement à montrer que le test évalue bien ce qu'il est censé évaluer (validité convergente), mais aussi à déterminer son pouvoir prédictif (validité prédictive). Au milieu du 20e siècle, les critiques grandissant sur les tests et leur utilisation insistent à davantage étayer leur validité. On cherche à évaluer la cohérence entre la structure interne du test et la théorie à laquelle il se réfère (validité de construit), la relation entre le contenu du test et la propriété mentale évaluée (validité de contenu) et toujours la relation du test avec d'autres variables (validité critérielle). Les nombreux travaux élargissent le champ de la validité conduisant progressivement à un morcellement du concept en autant de types de validité que de méthodes de validation. Vers la fin des 50, un mouvement de réunification est amorcé. La validité devient un concept global sous lequel sont rassemblés différents arguments de validité pour l'interprétation d'un test. Les *Standards for educational and psychological testing* qui publient les dernières recommandations pour l'utilisation des tests, adoptent une approche basée sur les arguments de validité (*argument-based approach*). Ils relèvent cinq sources pour les preuves de validité : (a) les preuves de validité basées sur le contenu du test, (b) les preuves de validité basées sur la relation à d'autres variables, (c) les preuves de validité basées sur la structure interne du test, (d) les preuves de validité basées sur les processus de réponse et, (e) les preuves de validité basées sur les conséquences sociales de l'évaluation (pour une description détaillée voir Sireci & Sukin, 2013). Suivant leur approche basée sur les arguments de validité, les *Standards* proposent la définition suivante :

> A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses. . . . Ultimately, the validity of an intended interpretation . . . relies on all the available evidence relevant to the technical quality of a testing system. This includes evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all examinees. (AERA et al., 1999, p. 17, cités par Sireci & Sukin, 2013)

Dans cette définition, il est relevé l'importance de la fidélité des scores et de l'évaluation de l'équité dans les preuves de validité pour une interprétation pertinente

des scores d'un test. Nous reviendrons largement sur la fidélité des scores dans le chapitre suivant. Pour terminer les considérations psychométriques sur l'évaluation, nous allons discuter de l'équité et des biais dans les tests qui sont en lien avec l'étude du fonctionnement différentiel que nous avons menée sur les items du WISC-IV.

2.6. ÉQUITÉ DE L'ÉVALUATION

Avec leur utilisation étendue dans les processus d'évaluation et de sélection d'individus, les tests sont devenus une aide à la prise de décision dans différentes sphères de la société (éducation, santé, légale, ressources humaines, etc.). Étant donné les enjeux et les répercussions liés à certaines décisions, le recours aux tests n'est pas un acte anodin. Bon nombre de critiques sont rattachées aux tests, notamment en ce qu'ils « affectent l'image de soi, autorisent des pronostics à partir de critères discutables, favorisent une classification rigide et définitive des individus » (Bourguignon, 2003, p. 103). Le bien-fondé des critiques soulève des réflexions sur l'éthique dans le testing et l'évaluation psychologique. Parmi les directives sur l'évaluation psychologique, il est clairement stipulé la responsabilité du psychologue de sélectionner et d'utiliser les tests à bon escient selon l'individu à évaluer.

> *Psychologists should select and use tests or assessments with members of populations for whom adequate reliability and validity of the test scores has been established. If the reliability and validity of the test scores has not been examined or verified for a particular population, psychologists are obligated to describe the strengths and limitations of the interpretations and recommendations derived from the test or assessment results.* (Ethical Standard 9.02b cité par Leong, Park, & Leach, 2013, p. 268)

Dans le processus de prise de décision, les tests vont aider à tester les hypothèses du clinicien, l'aider à établir un diagnostic et à formuler des pronostics. Les résultats de tests donnent des pistes à l'élaboration d'un projet (thérapeutique, professionnel, éducatif) qui implique l'avenir du sujet. Pour cela, les différences dans les scores au test doivent indiquer les différences entre les individus sur ce que le test évalue. Nous avons vu que la validation d'un test permet l'interprétation valide de ses résultats dans les contextes étudiés. Dans les préoccupations de l'éthique sociale de l'usage des tests, l'évaluation de la validité porte plus précisément sur l'équité de l'évaluation (*fairness in assessment*). Cette dernière est principalement étudiée dans les comparaisons entre les individus issus de la « culture majoritaire » et des individus issus

de « cultures minoritaires ». La culture majoritaire se réfère à la population pour laquelle le test a été conçu en première instance et sur laquelle il a étalonné. Étant donné les différences culturelles entre la population majoritaire et minoritaire, le matériel du test (p. ex., formulation des consignes, contenu d'un item, etc.) peut ne pas être approprié pour les individus des cultures minoritaires sur qui l'on utilise néanmoins le test. Dans la situation d'évaluation d'individus provenant d'autres cultures se pose donc la question de l'équité dans l'évaluation, mais pas seulement. Des problèmes d'équité peuvent porter sur des différences de sexe, d'âge, etc. En fait, la validité de l'interprétation des scores est compromise dès lors que, des individus de groupes différents (culture, ethnie, sexe, âge, etc.), ayant la même habileté sur ce qu'évalue le test n'obtiennent pas le même score. Dans cette situation, les différences mises en évidence par le test ne sont pas en lien avec ce qu'il évalue. Il n'y a plus d'équivalence des scores pour des individus de groupes différents ayant le même niveau sur la propriété mentale évaluée. Sans équivalence des scores, aucune comparaison ne peut être réalisée entre les individus de groupes différents. Ainsi, *« a test that is fair . . . reflects the same construct(s) for all test takers, and scores from it have the same meaning for all individuals in the intended population; a fair test does not advantage or disadvantage some individuals because of characteristics irrelevant to the intended construct »* (The Standards for Educational and Psychological Testing, 2014, p. 50). Dans le présent travail, nous allons donc nous intéresser à la principale menace sur l'équité : les biais.

2.6.1. NOTION DE BIAIS

De façon générale, on parle de biais pour relever toute menace à la validité des comparaisons entre les scores d'individus qui diffèrent sur des variables sociodémographiques. En psychométrie, un biais est un « *systematic inaccuracy of assessment* » de sorte que le score du test n'a pas la même signification lorsqu'on compare des groupes différents (Millsap & Everson, 1993, p. 297). Plus précisément, « on dit qu'une mesure est biaisée dès lors qu'elle ne mesure pas, ou qu'imparfaitement, ce qu'elle est censée mesurée. On est en présence d'un biais lorsque la mesure met en évidence des différences entre des groupes de sujets et que ces différences ne peuvent être mises en relation avec la ou les variables mesurées » (Vrignaud, 2002, p. 626). D'après van de Vijver et Leung (1997), on peut relever trois principales sources de biais : le biais de construit, le biais de méthode et le biais d'items.

Les biais de construit (ou de concept) surviennent des différences culturelles dans les définitions des construits psychologiques. Par exemple, le concept d'intelligence ne recouvre pas exactement les mêmes conduites selon les cultures. Le biais observé est d'autant plus important que les deux cultures sont éloignées l'une de l'autre. Les biais de méthode proviennent des caractéristiques du test et de sa passation. Il s'agit notamment de biais liés à la façon de répondre du sujet (p. ex., biais d'acquiescement, biais de réponses extrêmes, biais de réponses centrales, désirabilité sociale), aux normes, aux conditions de passations ou à l'examinateur (p. ex., son apparence, son attitude, etc.). Quant aux biais d'items qui nous intéressent en particulier, ils concernent toute menace qui peut biaiser un item. Il y a plusieurs méthodes pour détecter des biais d'item, nous en présentons deux approches : l'analyse des droites de régression et le fonctionnement différentiel des items selon les modèles de réponse à l'item.

2.6.2. MÉTHODE DES DROITES DE RÉGRESSION

L'analyse des droites de régression permet de détecter un biais d'items qui conduit à un biais de prédiction (ou aussi appelé biais du test). Le biais de prédiction porte spécifiquement sur la relation du test et d'un critère. L'une des utilités des tests réside dans les prédictions qu'ils permettent de formuler sur un critère. Par exemple, les études montrent que les résultats aux tests de QI sont de bons prédicteurs de la réussite scolaire, des performances professionnelles et de l'adaptation sociale (Brody, 1997; Neisser et al., 1996; Sternberg et al., 2001). Lorsqu'on étudie les biais de prédiction, la question qu'on se pose est de savoir si le test permet des prédictions pour tous les individus.

Tout d'abord, il est important de souligner qu'une différence de performances moyennes entre deux groupes différents n'est pas dans le sens psychométrique ce qu'on entend par un biais. Pour l'expliquer, nous allons nous référer à la situation représentée par la Figure 29 (p. 122). Il s'agit d'un cas de figure d'absence de biais entre deux groupes différents d'individus (groupe A et groupe B). Le nuage des points pour chaque groupe est représenté par une ellipse. On constate que le groupe A a des performances moyennes moins élevées que le groupe B. Néanmoins, il n'y pas de biais puisque pour un même score sur le test (trait en pointillé), les individus du groupe A et du groupe B ont le même niveau sur le critère, et inversement. Les différences mises en évidence par le test conduisent également et de façon proportionnelle à des différences sur le critère. Il est normal, par exemple, d'observer de meilleures

performances moyennes à un test logico-mathématique chez le groupe d'élèves qui sont dans le cours avancé de mathématiques (groupe B) par rapport au groupe d'élèves qui suit le cours standard de mathématiques (groupe A).

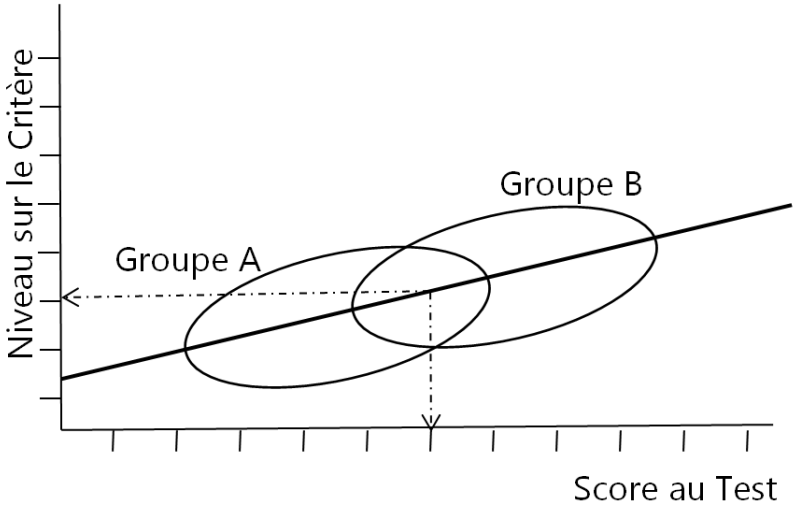


Figure 29. Illustration d'absence de biais.

En revanche, on est en présence d'un biais de prédiction si, ayant le même score au test (ou au critère), deux individus dont l'un appartient au groupe A et l'autre au groupe B, n'ont pas le même niveau sur le critère (ou au test). On peut distinguer trois situations : le biais de la pente, le biais de l'intercept (ou biais de l'ordonné à l'origine) ou les deux à fois. Le biais de la pente est en lien avec la validité qui examine la relation entre test-critère et montre une différence de sensibilité dans la prédiction test-critère. Par exemple, un test logico-mathématique pourrait bien prédire la performance des garçons à un examen de maths, mais être peu utile pour prédire les performances des filles. Dans ce cas, la pente de la droite de régression des garçons serait plutôt raide, tandis que la pente de la droite de régression des filles serait plutôt plate. Le biais de l'intercept montre une différence de niveau initial sur le test qui amène à systématiquement sous-évaluer ou surévaluer la performance au critère de tous les membres d'un groupe possédant une certaine caractéristique. Dans la Figure 30 (p. 123), les droites de régression illustrent la situation d'un biais de la pente et de l'intercept. À nouveau, une ellipse représente le nuage des points pour chaque groupe. Dans cette situation, on s'aperçoit que, pour un même score au test (trait en pointillé), les individus des groupes A et B n'ont pas les mêmes scores au critère. Pour un même score sur le test (trait en pointillé), le groupe B obtient des scores plus élevés sur le

critère que le groupe A. De même, pour un même score sur le critère, les individus des groupes A et B n'ont pas les mêmes scores au test. Les droites de régression révèlent un biais de prédiction. Si le test est étalonné sur des individus du groupe A (groupe majoritaire), il y aura systématiquement une sous-évaluation des performances des individus du groupe B (groupe minoritaire) sur le critère. Des comparaisons entre les individus des deux groupes ne sont pas valides ; il n'y a pas équivalence des scores.

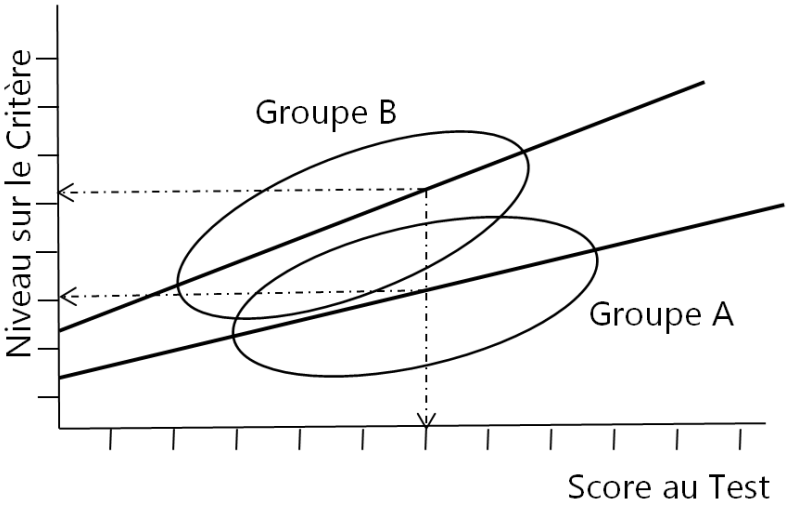


Figure 30. Illustration d'un biais de la pente et de l'intercept.

2.6.3. FONCTIONNEMENT DIFFÉRENTIEL DES ITEMS

L'analyse du fonctionnement différentiel des items (FDI, ou DIF dans la littérature anglophone pour *differential item functioning*) cherche à détecter les éventuels items biaisés d'un test. Elle n'est pas rattachée à un cadre de mesure, même si elle est souvent associée aux modèles de réponse à l'item (MRI). En fait, étant moins connoté, le terme de fonctionnement différentiel des items – introduit par Holland et Thayer (1988) – tend à remplacer le terme de biais. Toutefois, le terme de fonctionnement différentiel se réfère plus spécifiquement à des procédures statistiques qui déterminent le comportement des items entre des groupes distincts, tandis que le terme de biais renvoie aux différences entre groupes de manière plus générale qu'au seul niveau des items (Teresi & Jones, 2013). Pour nos analyses sur les items du WISC-IV, nous étudions une méthode de FDI fondée sur les modèles de réponse à l'item. La présentation qui suit se limite donc aux approches basées sur les MRI.

On repère d'éventuels items biaisés, lorsque les différences sur les items du test entre groupes différents ne sont pas constantes pour tous les items. À l'instar de Bertrand et Blais (2004), nous définissons deux critères pour déterminer qu'un item est biaisé envers un groupe : (a) deux individus d'habileté équivalente sur la propriété mentale évaluée par le test, mais issus de deux groupes distincts, n'ont pas la même probabilité de réussir un même item et (b) la différence de probabilité de réussir un même item dépend d'une autre variable que la propriété mentale évaluée par le test. À l'inverse, « un item est considéré comme non biaisé lorsque la probabilité de réussir cet item est la même pour tous les sujets de la population possédant la même aptitude, indépendamment de leur sous-groupe d'appartenance » (Osterlind, 1989 cité par Laveault & Grégoire, 2014, p. 231).

Nous n'entrerons pas dans les aspects techniques des méthodes d'analyse du fonctionnement différentiel ; notre présentation vise à la compréhension conceptuelle. Dans les approches selon les modèles de l'item, un fonctionnement différentiel des items (FDI) apparaît lorsque les courbes caractéristiques des item (CCI) d'un test ne sont pas équivalentes pour tous les individus de deux groupes différents. La Figure 31 (p. 125) illustre la comparaison entre deux courbes caractéristiques d'un même item d'un test de mathématiques. Pour rappel, une CCI représente la probabilité de réussir l'item en fonction de l'habileté sur le trait latent. Sur la Figure 31 (p. 125), l'une des CCI provient d'un échantillon d'anglophones (trait plein), tandis que l'autre provient d'un échantillon de francophones (trait en pointillé). La version originale du test est adiminstrée à l'échantillon d'anglophones, tandis qu'une une adaptation en français est administrée à l'échantillon de francophones. On peut voir que la probabilité de réussir l'item pour un niveau de trait latent n'est pas la même dans les deux groupes. Par exemple, pour un niveau de trait latent $\theta = 0$, la probabilité de réussir l'item est d'environ 70 % chez les francophones, alors qu'elle est d'environ 50 % pour les anglophones.

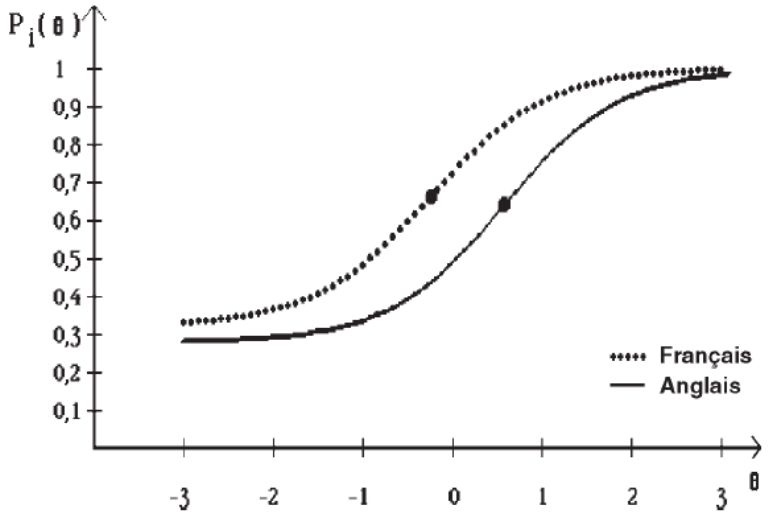


Figure 31. Courbe caractéristique d'un item pour le groupe de francophones (trait en pointillé) et pour le groupe d'anglophones (trait plein). Le point d'inflexion de la courbe est indiqué par un point noir. *Source* : Bertrand et Blais (2004, p. 284).

Un FDI se définit donc comme une différence de probabilité de réussir un item pour des individus d'habileté égale, mais appartenant à des groupes distincts (Bertrand & Blais, 2004). Dans le cas d'une absence de FDI, les CCI de chaque groupe sont proches, voire se superposent. Le FDI est donc « une notion statistique, une certaine valeur relative à une différence de probabilités » (Bertrand & Blais, 2004, p. 285). La présence d'un FDI satisfait le critère (a) de la définition d'un item biaisé. Pour satisfaire le critère (b), le FDI observé doit révéler que l'écart entre les probabilités de réussir l'item des deux groupes est signifiant par rapport à l'interprétation des scores du test. De manière statistique, on peut déterminer une valeur seuil à partir de laquelle l'aire entre les deux CCI révèle un écart significatif. Mais ce n'est pas suffisant, il faut également une réflexion approfondie sur les données pour savoir si les différences de probabilités résultent d'une autre variable que celle évaluée par le test. Nous reprenons l'exemple de la Figure 31 pour expliciter notre propos. L'exemple porte sur un test de mathématiques en anglais qui a été adapté en français. Les items sont des problèmes de maths. Les CCI entre les deux groupes montrent un FDI pour l'item considéré. Selon une valeur seuil définie, l'écart entre les CCI est considéré assez important pour être significatif. Est-ce que l'item est biaisé ? On ne peut pas encore l'affirmer. Cela dépend encore de l'interprétation qu'on donne au score du test. Le français est une langue moins synthétique que l'anglais. Le passage de l'anglais au français peut conduire à un

allongement de l'énoncé des problèmes de maths. Si le score du test est interprété comme un indicateur de l'habileté en mathématiques, l'item est biaisé. Si en revanche, les compétences en lecture font partie des habiletés secondaires et légitimes évaluées par le test, alors l'item n'est pas biaisé. En effet, on peut argumenter que la réussite en mathématiques inclut de réussir sur des problèmes de maths, qui eux-mêmes demandent une certaine compétence en lecture. Dans l'adaptation en français, l'item évalue principalement les aptitudes mathématiques mais secondairement des aptitudes en lecture interviennent également, tandis que, dans la version anglaise, la demande en lecture pour le test est moindre.

La détection d'un biais ne repose donc pas que sur une analyse statistique, mais demande une réflexion contextuelle sur les données. De façon plus générale, les données psychométriques s'interprètent en tenant compte d'un cadre théorique, d'un contexte et des finalités de l'évaluation psychologique. Le prochain chapitre portera sur le sujet principal du présent travail : la fidélité des scores d'un test. Nous verrons que plusieurs facteurs liés à la méthode d'estimation de la fidélité, mais aussi liés aux calculs du risque peuvent intervenir dans l'interprétation d'un coefficient de fidélité d'un test.

3. FIDÉLITÉ DES SCORES D'UN TEST

Dans la première partie de ce chapitre, le concept de fidélité des scores d'un test sera présenté. Tout d'abord, nous définirons la fidélité dans le cadre de la théorie classique des tests dans lequel elle est un des concepts-clés (voir section 3.1). Ensuite, nous nous étendrons sur les différentes méthodes pour l'estimation de la fidélité (voir section 3.2). L'estimation de la fidélité fournit un coefficient de corrélation dont l'interprétation demande de tenir compte de plusieurs facteurs (voir section 3.3) Les notions d'erreur de mesure et d'intervalle de confiance seront développées, pour rendre compte de l'utilisation sur le plan clinique des données de fidélité des scores d'un test.

Dans la seconde partie du chapitre, nous focaliserons notre attention sur le cœur du présent travail, à savoir l'échelle de Wechsler pour les enfants (WISC-IV) et l'estimation de la stabilité de ses scores. Nous définirons les types d'évaluation de la stabilité qu'on relève dans les études (voir section 3.4.1). Le chapitre se terminera sur une revue de la littérature de la stabilité tant à court qu'à long terme du WISC-IV (voir sections 3.4.3 et 3.4.4).

3.1. DÉFINITION DE LA FIDÉLITÉ DES SCORES

Étant donné que les tests psychologiques sont utilisés et interprétés pour comprendre le fonctionnement d'un individu, leurs résultats se doivent d'être fiables et reproductibles. Un test n'aura aucune utilité, si les résultats d'un individu amènent à différentes hypothèses ou conclusions d'une passation à l'autre. Plusieurs variables peuvent affecter les résultats d'un test au cours du processus d'évaluation, telles des variations situationnelles (p. ex., conditions environnementales), des variations induites par l'expérimentateur (p. ex., attitude, erreurs de cotation), des variations instrumentales (p. ex., format des items, contenu des items) et des variations dues au sujet lui-même (p. ex., fatigue, motivation, anxiété, familiarité avec la tâche). Dans un sens général, la fidélité est un indicateur du degré de reproductibilité d'un score (quel que soit ce qu'il mesure). Nous allons davantage définir ce concept, mais tout d'abord soulevons un raccourci de langage à son propos. Couramment, on parle de la fidélité comme d'une propriété du test en déclarant : « ce test est fidèle » ou « la fidélité du test est… ». Or la fidélité est une propriété des scores d'un test, et non une propriété

inhérente au test lui-même. Il ne s'agit pas ici d'une subtile nuance ; la distinction est importante dans ses implications quant à l'utilisation et à l'interprétation des données de fidélité pour un test. Si la fidélité est une propriété du test, cela signifierait qu'elle peut être établie définitivement pour les tous contextes d'utilisation et pour tous les individus à qui le test est administré. Nous pourrions nous référer à une valeur applicable en toutes circonstances. Or, la fidélité n'est pas une propriété absolue et immuable du test. Nous le verrons longuement dans ce chapitre, la fidélité pour un test donné n'est pas caractérisée par un score unique. Dans la théorie classique des tests (TCT), il existe plusieurs méthodes d'estimations de la fidélité, dont chacune tient compte de différentes sources d'erreur. Nous le verrons également, même si la fidélité est estimée à partir d'un large échantillon dans des conditions standardisées, elle varie néanmoins d'un échantillon à l'autre en fonction des caractéristiques de l'échantillon testé et du contexte d'évaluation. On ne peut pas transférer les données de fidélité étudiées d'un échantillon à l'autre, ni les utiliser dans d'autres contextes que ceux étudiés. Cela rend nécessaire des études comme la nôtre avec des sujets tout-venant ainsi que sur d'autres populations. Ainsi, lorsqu'on parle de « fidélité d'un test », il s'agit toujours d'entendre « fidélité des scores d'un test ».

Revenons à présent à la définition de la fidélité, qui est le concept sur lequel la théorie classique des tests s'est construite. Nous rappelons que, selon l'équation de base de la TCT, le score observé (X) d'un individu à un test est égal à l'addition de son score vrai (V) et de l'erreur de mesure (E). Le score vrai (V) est une entité théorique qui représente le score moyen qu'obtiendrait le sujet s'il passait un même test une infinité de fois. Il est constant pour un individu et pour un test donné, mais différent d'un individu à l'autre et d'un test à l'autre. On ne peut pas directement l'observer. Nous n'avons accès qu'au score observé (X) qui est le score obtenu par un sujet à une passation du test. Pour un sujet, le score observé varie d'une passation à l'autre d'un même test à cause de l'erreur. En effet, l'erreur de mesure (E) est une entité constituée des fluctuations qui surviennent de manière aléatoire d'une passation à l'autre du test. Comme nous en avons déjà parlé[12], on peut distinguer deux types d'erreur : l'erreur systématique et l'erreur aléatoire. Sur des mesures répétées à un même test, l'erreur systématique affecte les mesures dans le même sens et avec la même intensité. Pour tous les individus d'un échantillon, elle va surévaluer (ou sous-évaluer) leur score observé de manière constante et prévisible. Ce type d'erreur ne peut pas être détecté par les méthodes d'estimation de la fidélité (corrélations), car elle ne révèle pas une

[12] Voir section 2.1.1.1, p. 73.

inconsistance dans les mesures répétées. Quant à l'erreur aléatoire, elle affecte les mesures dans un sens et dans un autre et avec une intensité différente. D'une passation à l'autre, l'influence de l'erreur aléatoire sur la mesure est imprévisible ; elle peut être positive (surévaluer) ou négative (sous-évaluer) d'une passation à l'autre. Pour évaluer la fidélité comme entendue dans la théorie classique des tests, on en tient compte que de l'erreur aléatoire.

Compte tenu des effets d'apprentissage notamment, il n'est pas sensé d'effectuer une infinité de passations d'un même test sur un même individu. Les mesures répétées se réalisent sur un groupe d'individus à qui on fait passer le test. La fidélité porte donc sur la variabilité des scores observés des individus d'un échantillon. Suivant l'équation de base, la variance d'un échantillon (s_x^2) se traduit ainsi :

$$s_x^2 = s_v^2 + s_e^2 \qquad (12)$$

Où s_v^2 est la variance des scores vrais et s_e^2 est la variance des erreurs. La variance vraie des scores représente la variabilité du score vrai au sein des individus de l'échantillon à qui on passe le test, tandis que la variance d'erreur représente la variabilité des erreurs de mesure de l'échantillon. Partant de cette nouvelle équation, la question que pose l'évaluation de la fidélité est la suivante : dans quelle mesure les différences individuelles sur les scores observés au test relèvent-elles de « vraies » différences sur la propriété mentale évaluée et dans quelle mesure celles-ci relèvent-elles d'erreurs dues au hasard ? En termes plus techniques, la fidélité permet d'évaluer quelle proportion de la variance totale des scores observés à un test est expliquée par la variance vraie des scores et quelle proportion est expliquée par la variance d'erreur.

$$r_{xx'} = r_{vx}^2 = \frac{s_v^2}{s_x^2} \qquad (13)$$

La fidélité est exprimée sous la forme du carré de la corrélation entre le score vrai et le score observé. Elle peut également être traduite par un rapport entre la variance des scores vrais (s_v^2) et la variance des scores observés (s_x^2). Plus la variance des scores observés diffère de la variance des scores vrais, plus la fidélité décrira un certain degré d'écart entre le score observé et le score vrai des individus. À l'inverse, plus les deux variances ont des valeurs proches, plus la fidélité décrira des scores observés et des scores vrais similaires pour les individus de l'échantillon. Ainsi, « meilleure est la fidélité, meilleur sera la prédiction du score vrai à partir du score observé » (Laveault & Grégoire, 2014, p. 112). Dans la situation d'un test dont les scores sont parfaitement fidèles, 100 % de la variance totale est attribuable à la variance des scores vrais. Le

score observé est dépourvu d'erreur de mesure, et coïncide exactement avec le score vrai. Évidemment, il s'agit d'une situation idéale et hypothétique. Dans la réalité, l'erreur ne peut jamais être totalement éliminée. On peut décomposer la variance des scores vrais et transformer l'Équation (13) pour faire apparaître la variance d'erreur (s_e^2) :

$$r_{xx'} = \frac{s_x^2 - s_e^2}{s_x^2} = \frac{s_x^2}{s_x^2} - \frac{s_e^2}{s_x^2} = 1 - \frac{s_e^2}{s_x^2} \quad (14)$$

Selon cette nouvelle formulation de la fidélité, plus le résultat de la fraction est élevé (c.-à-d. plus la variance d'erreur est importante par rapport à la variance des scores observés), plus la fidélité sera faible.

Comme le score vrai des individus n'est pas directement accessible, la fidélité s'opérationnalise par un coefficient de fidélité qui traduit la relation entre les scores observés de mêmes sujets sur deux tests considérés comme parallèles, et non entre le score vrai et le score observé des sujets. Un test administré à deux reprises, de même que deux versions d'un test ou deux moitiés d'un test, peuvent être considérés comme des tests parallèles. Pour rappel, deux tests parallèles sont supposés présenter (1) un contenu et une difficulté équivalents, (2) des scores vrais qui corrèlent[13], (3) un écart type égal ($\sigma_1 = \sigma_2$) ainsi que (4) une variance d'erreur égale ($\sigma_{e1}^2 = \sigma_{e2}^2$). En vertu des propriétés des tests parallèles, la corrélation entre les scores observés à deux tests parallèles permet de rendre compte de la proportion de variance totale des scores observés qui est expliquée par la variance du score vrai et par la variance d'erreur.

3.2. MÉTHODES D'ESTIMATION DE LA FIDÉLITÉ

Dans le cadre de la théorie classique des tests, plusieurs procédures sont décrites pour l'estimation de la fidélité en fonction de la source d'erreur prise en compte. (a) Le même test est administré à deux reprises aux mêmes sujets dans un intervalle de temps court ou long (fidélité par la méthode test-retest). (b) On administre deux formes parallèles d'un test aux mêmes sujets dans un laps de temps plus ou moins court (fidélité par la méthode des formes parallèles avec/sans délai). (c) La cotation d'un test est effectuée par au moins deux évaluateurs (fidélité par la méthode interjuges). (d) On administre un test à des sujets, puis on partitionne l'ensemble des

[13] Les scores vrais à chaque test parallèle sont égaux, puisque le score vrai d'un même sujet à deux tests parallèles est théoriquement identique.

items en deux parties (fidélité par la méthode bissection) ou (e) on partitionne le test en autant de parties qu'il y a d'items (fidélité par la méthode des covariances). Le Tableau 2 résume les différentes sources de variance d'erreur dont tient compte chaque méthode d'estimation.

Tableau 2

Sources de variance d'erreur selon la méthode d'estimation de la fidélité (selon Anastasi, 1994)

Variance d'erreur	Méthode d'estimation	Type de coefficient	Setting
Échantillonnage temporel	Test-retest	Stabilité	2 passations avec les mêmes individus qui passent le même test
Échantillonnage des contenus	Formes parallèles immédiates	Équivalence	2 passations à la suite avec les mêmes individus qui passent 2 versions du test
Échantillonnage temporel et des contenus	Formes parallèles différées	Équivalence et stabilité	2 passations avec les mêmes individus qui passent 2 versions du test
Différences entre cotateurs	Interjuges	Équivalence interjuges	Au moins 2 cotateurs
Échantillonnage des contenus	Bissection	Consistance interne / *split half*	1 passation
Échantillonnage des contenus	Covariances	Alpha de Cronbach ou Kuder-Richardson	1 passation

Les méthodes d'estimation de la fidélité des scores apportent des informations différentes les unes des autres. En fonction des objectifs du test, certaines méthodes sont plus appropriées. Par exemple, s'agissant d'un test qui porte sur des traits ou des comportements stables, il est important d'évaluer la fidélité des scores dans le temps. À cause du coût, le choix des fidélités à estimer se limite souvent aux méthodes qui demandent un test et une passation. Nous verrons par la suite que ces méthodes ne permettent pas à proprement parlé l'estimation de la fidélité des scores. Dans une étude sur la fréquence d'utilisation de chaque méthode, Hogan, Benjamin et Brezinski (2000) consultent un registre qui recense des informations sur 2'078 tests apparus dans 37 revues scientifiques (en psychologie, science de l'éducation ou sociologie) entre 1991 et 1995. Le registre fournit des informations sur le nom du test, ce qu'il évalue, le

nombre et le format des items, le temps de passation et des données psychométriques. Hogan, Benjamin et Brezinski constituent leur échantillon en intégrant de manière systématique tous les trois tests dans la liste, sélectionnant ainsi un total de 696 tests pour leur étude. Ils relèvent que, pour la majorité des 696 tests étudiés, un seul type de fidélité (75 %) est reporté. Moins fréquent est de reporter deux types d'estimation de la fidélité (17 %), trois et plus (2 %) ou aucun (6 %). En regardant en détail le type de méthode, l'alpha de Cronbach triomphe loin devant, étant reporté pour plus de 66 % des tests de l'étude, tandis que le coefficient test-retest n'est fourni que pour 19 % des tests et le coefficient *split-half* pour 4.1 % des tests. Plus généralement, les auteurs notent un certain manque de clarté dans les articles sur la méthode d'estimation utilisée. Les articles énoncent des termes génériques tels que « coefficient de fidélité » ou « coefficient de consistance interne » sans autre précision sur la méthode d'estimation. De plus, peu d'études renseignent de manière détaillée sur les caractéristiques de l'échantillon sur lequel l'estimation de la fidélité est réalisée. Nous l'avons déjà souligné, la fidélité des scores est relative à la composition de l'échantillon testé, il est donc important de le décrire.

3.2.1. MÉTHODE TEST-RETEST

La procédure test-retest consiste à administrer un même test à deux reprises aux mêmes sujets après un certain délai de temps. L'intervalle de temps entre les deux passations peut être court (quelques jours) ou long (plusieurs années). On calcule un coefficient de fidélité entre les scores obtenus au test (temps 1) et au retest (temps 2) par chaque sujet. Ce coefficient, appelé *coefficient de stabilité*[14] ou c*oefficient test-retest*, renseigne sur la stabilité des différences interindividuelles dans le temps. Parmi les variables qui peuvent fluctuer au cours du temps, il y a par exemple, l'état physique et mental du sujet, les conditions de passation ou l'influence imputable à l'examinateur. Une faible corrélation entre le score au test et au retest signale que « l'effet du passage du temps s'ajoutera à l'erreur de mesure » (Laveault & Grégoire, 2014, p. 114). En outre, elle doit également questionner sur l'hypothèse de stabilité conférée à la propriété mentale évaluée par le test.

L'application de cette méthode rencontre des limites, puisqu'elle suppose l'absence d'effet d'apprentissage, de souvenirs liés à la première passation ou de

[14] Pour des intervalles de moins de deux mois entre les deux passations, certains parlent d'un coefficient de confiance (Bernaud, 2014).

changement important dans la propriété mentale évaluée (p. ex., un apprentissage différentiel entre temps). D'après Duff et al. (2011), l'effet d'apprentissage se définit comme une amélioration des performances dans les tests cognitifs à la suite de l'exposition répétée d'un même matériel de test. Induit par la répétition du test, et non par une réelle augmentation du niveau d'habileté, l'effet d'apprentissage est considéré comme une source d'erreur, d'autant qu'il peut masquer ou minimiser un déclin cognitif. Il s'observe de façon marquée sur des intervalles test-retest courts (quelques jours à quelques mois). Dans des travaux qui comparent des groupes contrôle et clinique (Cooper, Lacritz, Weiner, Rosenberg, & Cullum, 2004; Duff et al., 2008; Duff, Westervelt, McCaffrey, & Haase, 2001), il est intéressant de relever la présence d'un effet d'apprentissage chez les sujets sains, tandis que chez les sujets patients, l'effet d'apprentissage est moins prononcé, voire absent. La présence d'un effet du retest peut conduire à la violation du postulat d'indépendance entre l'erreur et le score vrai[15]. Par exemple dans le cas où les sujets les plus forts à la première passation sont ceux qui, au moment de la seconde passation, se rappellent le mieux des questions posées ou de la stratégie opportune à appliquer.

Contrairement à l'effet d'apprentissage et autres fluctuations attribuables à un état temporaire du sujet ou au fruit du hasard, un changement réel dans la propriété mentale ne fait pas partie des sources d'erreur de la mesure. Il révèle chez l'individu une différence ancrée du niveau d'habileté (p. ex., un changement développemental, un apprentissage). Par exemple lors d'une remédiation mise en place, on s'attend à ce que l'intervention conduise à un changement entre le pré-test et le post-test. Lorsqu'un changement de performances s'observe entre deux passations d'un même test, il est important de discerner entre l'effet d'apprentissage et un changement réel, même si la distinction est peu aisée.

Avec la méthode test-retest se pose la question de la durée du délai. En général, elle est à déterminer en regard des tâches du test, des changements développementaux de la population étudiée et du temps estimé suffisant pour limiter les influences de l'expérience d'une première passation. En cas d'un intervalle trop court, on est confronté à une possible mémorisation des items ou des stratégies opportunes à appliquer, tandis qu'en cas d'un intervalle trop long, on court le risque d'un réel changement de niveau sur la propriété mentale. Ainsi, les intervalles trop courts tendent à surestimer la corrélation entre les scores des deux passations car, théoriquement, les sujets sains bénéficient plus ou moins uniformément d'un effet

[15] Voir 2.1.1.2, p. 78.

d'apprentissage. La durée de l'intervalle test-retest est déterminée pour être assez longue afin d'estomper l'effet d'apprentissage, mais pas trop pour ne pas risquer une sous-estimation de la fidélité à cause d'un réel changement.

Dans les tests cognitifs qui nous intéressent tout particulièrement, des études montrent que les tâches impliquant des habiletés de compréhension-connaissance présentent une meilleure stabilité des scores que les tâches impliquant du raisonnement fluide et de la résolution de problème (Calamia, Markon, & Tranel, 2012; Dikmen, Heaton, Grant, & Temkin, 1999; Schwartzman, Gold, Andres, Arbuckle, & Chaikelson, 1987). De même, pour des intervalles de 3 à 6 mois, les gains à la seconde passation tendent à être plus importants pour les épreuves simples de vitesse de traitement que pour les épreuves verbales de vocabulaire ou de culture générale (Calamia et al., 2012; Estevis, Basso, & Combs, 2012).

3.2.2. MÉTHODE DES FORMES PARALLÈLES IMMÉDIATES/DIFFÉRÉES

La méthode des formes parallèles (ou des formes équivalentes) consiste à administrer deux versions similaires d'un même test (forme A et forme B) aux mêmes individus. Lors de la première passation, une moitié de l'échantillon passe la forme A, tandis que l'autre moitié passe la forme B, et inversement lors de la seconde passation (principe de l'ordre contrebalancé). Bien que composées d'items différents, les deux formes du test doivent être équivalentes quant à leur nombre d'items, à leur consigne, à leur contenu, à l'étendue de leur niveau de difficulté, etc. L'intervalle entre les deux passations est très court – généralement à la suite – pour la procédure immédiate, et de quelques jours à plusieurs semaines pour la procédure avec délai.

La corrélation dans cette méthode examine la relation entre les scores obtenus par chaque sujet aux deux versions du test. Pour la méthode sans délai, on contrôle uniquement les fluctuations dues à l'échantillonnage des items, et le coefficient calculé est appelé *coefficient d'équivalence*. Une faible corrélation traduit un manque de parallélisme (ou un faible degré d'équivalence) entre les deux formes. Pour la méthode avec délai, le coefficient de fidélité s'appelle plus précisément *coefficient d'équivalence et de stabilité*. Les formes parallèles différées sont considérées comme la meilleure méthode d'estimation de la fidélité (Dickes et al., 1994). L'estimation de la fidélité au moyen de cette procédure produit des valeurs de coefficients plus faibles comparativement aux autres méthodes, puisqu'on « cumule les erreurs aléatoires de

mesure imputables aux différences d'échantillonnage des items entre les deux tests parallèles et les erreurs aléatoires de mesure imputables à l'effet du temps » (Laveault & Grégoire, 2014, p. 117).

Lorsqu'on doit évaluer régulièrement un sujet avec un même instrument, l'utilisation de la forme parallèle du test atténue l'effet d'apprentissage (sans néanmoins l'éliminer). Cependant, très peu de tests disposent de formes parallèles. En soi, il est déjà difficile et couteux de concevoir un bon test, alors deux !

3.2.3. MÉTHODE DE BISSECTION

La méthode de bissection (ou appelée aussi méthode des moitiés, méthode du partage ou méthode du *split-half*) consiste à diviser les items d'un test en deux parties égales (p. ex., items pairs vs items impairs, la première moitié d'items vs la seconde moitié d'items). Les deux parties s'apparentent à deux formes équivalentes/parallèles d'un même test. À partir d'une seule passation d'un seul test, on peut alors calculer deux scores totaux ; un score total pour chacune des parties. La corrélation entre les scores totaux des deux parties renseigne sur la consistance interne (ou la cohérence interne) du test entier. Le coefficient calculé est appelé *coefficient split-half* (ou de consistance interne). Il est élevé si les deux parties du test sont consistantes, c'est-à-dire si les items des deux parties contribuent dans le même sens à l'évaluation de la propriété mentale.

La méthode de bissection est facile à mettre en œuvre, seulement elle ne permet pas d'estimer la fidélité dans le sens de la reproductibilité des scores. En effet, il n'y a pas de mesures répétées dans le temps ; il s'agit d'individus testés à un seul moment de temps. De plus, la méthode porte sur la fidélité des scores totaux de deux moitiés d'un test, et non sur celle du score total d'un test entier. C'est la fidélité au niveau du score total au test qui nous intéresse. Davantage qu'une estimation de la fidélité, cette méthode sert donc à se prononcer sur le degré d'uniformité et de cohérence des parties constituant le test. Il s'agit donc d'une méthode d'évaluation de la consistance interne des items du test, et non de la fidélité au sens de la répétabilité et de la stabilité des mesures comme nous l'avons définie. Toutefois, plus un test possède des items consistants, plus il tend à être fidèle.

Plus on recueille un large échantillon des comportements de l'individu, plus l'évaluation que nous en faisons est fidèle. On comprend alors que la réduction du

nombre d'items sous-estime la fidélité par rapport à la fidélité estimée sur davantage d'items. L'estimation de la consistance interne par deux moitiés de test sous-estime donc la fidélité du test entier. On doit corriger cette sous-estimation avec par exemple, la correction de Spearman-Brown (W. Brown, 1910; Spearman, 1910).

$$r_{xx'} = \frac{2r_{AB}}{1 + r_{AB}} \quad (15)$$

Où $r_{xx'}$ représente le coefficient de consistance interne attendu du test entier et r_{AB} est le coefficient de corrélation entre les deux moitiés du test. Lorsque les deux moitiés du test ne forment pas des tests strictement parallèles et qu'elles présentent une forte différence de variances, la correction de Rulon (1939) est alors préférée.

3.2.4. MÉTHODE DES COVARIANCES

Une autre méthode d'estimation de la fidélité au moyen d'une seule passation et d'une seule version d'un test est réalisée avec la méthode des covariances. Il s'agit de partitionner le test en autant de parties que de nombre d'items qui le composent. Tous les items sont analysés deux à deux. En revanche, on tient compte de la performance du sujet sur chacun des items du test, et non sur deux moitiés séparées comme dans la méthode de bissection. Plus la covariation entre les pairs d'item est élevée, meilleure est la consistance de l'ensemble des items du test. Des items consistants sont des items qui contribuent d'une façon cohérente (ou dans le même sens) à l'évaluation de la propriété mentale. À nouveau, avec cette méthode, il s'agit de l'évaluation de la consistance interne des items d'un test davantage que de l'évaluation de la fidélité du score au test.

Pour cette méthode, on peut calculer un coefficient à partir des formules développées par Kuder et Richardson (1937), ou plus usuellement le coefficient alpha de Cronbach (Cronbach, 1951), qui notons-le, « repose sur un postulat fort que chaque item est parallèle aux autres (même degré de difficulté, même variance) » (Laveault & Grégoire, 2014, p. 120). Ce qui est rarement évalué au préalable. Nous l'avons déjà mentionné, l'alpha de Cronbach est utilisé à tort comme un indicateur de l'unidimensionnalité des items, alors que l'application de l'alpha de Cronbach présuppose l'unidimensionnalité.

3.2.5. MÉTHODE INTERJUGES

La fidélité interjuges (ou intercorrecteurs) examine les variations aléatoires entre les évaluateurs/ cotateurs d'un test. On demande à des juges de coter de façon indépendante un même protocole et on regarde ensuite au moyen des corrélations quel(s) score(s) présente(nt) une concordance élevée. Si la corrélation entre les évaluateurs est bonne, cela signale une équivalence interjuges. On parle également de fidélité interobservateurs s'il s'agit d'observer un comportement à l'aide d'une grille par exemple. L'information donnée par cette évaluation permet, d'une part de savoir si les critères de cotation ont besoin d'être affinés et, d'autre part, de connaître le degré d'objectivité dans les scores calculés. Sans surprise, les épreuves impliquant de formuler une appréciation ou une inférence pour coter les réponses du sujet révèlent plus particulièrement de la variance interexaminateurs (p. ex., des tests de compréhension verbale, les tests projectifs, tests de créativité).

Notons que nous présentons la méthode interjuges, car les ouvrages de psychométrie l'intègrent dans les méthodes d'estimation de la fidélité. Néanmoins, cette méthode n'estime pas la fidélité telle que nous l'avons définie au début du présent chapitre. Cette méthode évalue un degré de consensus entre les cotateurs et ne porte aucunement sur les différences interindividuelles sur ce qu'évalue le test. Le coefficient d'équivalence interjuges ne rend donc pas compte de la part de variance des scores vrais dans la variance totale des scores observés.

3.3. INTERPRÉTATION DE LA FIDÉLITÉ DES SCORES D'UN TEST

Comme définie, la fidélité rend compte des proportions de variance vraie des scores et de variance d'erreur qui sont expliquées dans la variance totale. Le coefficient de corrélation qui l'opérationnalise est généralement traduit en pourcentage de variance commune entre le score vrai et le score observé. Par exemple, un coefficient de fidélité de .70 signifie qu'il y a 70 % de variance commune entre le score vrai et le score observé. Autrement dit, 70 % de la variance des scores au test correspond à la variance vraie sur la propriété mentale évaluée. En nous référant à la relation qui relie les différentes variances (c.-à-d. $s_x^2 = s_v^2 + s_e^2$), nous pouvons également déduire que 30 % de la variance dans les scores observés est due à la variance d'erreur.

3.3.1. FACTEURS INFLUENÇANT SUR L'ESTIMATION DE LA FIDÉLITÉ

Dans la théorie classique des tests, l'estimation de la fidélité est réalisée par la méthode des corrélations. En tant que coefficient de corrélation, le coefficient de fidélité est influencé par les facteurs qui agissent sur la corrélation (p. ex., l'étendue des scores dans l'échantillon). À ces facteurs s'ajoutent d'autres en lien avec l'instrument de mesure (p. ex., longueur du test, difficulté des items). Nous allons développer ces facteurs afin de souligner à nouveau que l'interprétation de la fidélité ne porte pas uniquement sur la valeur d'un coefficient, mais qu'elle demande une réflexion plus large autour de la méthode d'estimation.

3.3.1.1. Étendue des différences interindividuelle

Comme on le sait, la corrélation est affectée par la variabilité au sein de l'échantillon. En effet, elle rend compte de la relation linéaire entre la variation réciproque (c.-à-d. covariation) de deux variables manifestes. Pour cela, elle nécessite un certain degré de dispersion sur chacune des variables. Dans la population générale, l'étendue des scores observés est relativement large. En revanche, dans un échantillon, il peut arriver que les individus sélectionnés forment un groupe plus homogène (en âge, en niveau de *g*, etc.) que la population générale dont ils sont issus. Il s'agit d'une situation de restriction des scores, et donc de réduction des différences interindividuelles sur le test. Dans ce cas, les scores des individus de l'échantillon se concentrent sur une zone plus restreinte que l'étendue possible qu'on aurait observé sur la population générale. L'homogénéité au sein des scores observés gomme les différences d'habileté entre les individus, affaiblissant la fidélité qui est alors sous-estimée. Il peut également arriver le cas inverse : l'échantillon est plus hétérogène que la population générale. Dans ce cas-là, la fidélité est surestimée. Pour expliciter les effets de l'homogénéité ou de l'hétérogénéité d'un échantillon sur le coefficient de fidélité, nous allons détailler chacun des deux cas de figure avec la situation de l'évaluation de la fidélité par une procédure test-retest.

Dans la procédure test-retest, la corrélation entre les performances à la première passation et à la seconde passation renseigne sur la stabilité du classement des individus aux deux passations. En effet, plus les individus occupent un rang similaire aux deux passations, plus la corrélation est élevée. À cause des erreurs de mesures qui entachent forcément le score observé au test, les individus n'obtiennent pas

exactement les mêmes scores aux deux passations, mais cela ne conduit pas forcément à un changement de position dans leur groupe. Prenons d'abord le cas le plus fréquent : un échantillon trop homogène (p. ex., composé des meilleurs élèves d'une classe). Dans cette situation, une petite variation dans les performances des individus peut les faire bouger dans le classement et complètement changer l'ordre des individus d'une passation à l'autre. Le coefficient test-retest calculé est alors plus faible que si les individus montraient une plus grande différenciation. Si notre échantillon est trop hétérogène, la distance entre les niveaux de performances des individus est très éloignée. Une variation – même importante – des performances d'une passation à l'autre ne conduit pas à un changement de place dans le classement ; la corrélation entre les deux mesures apparaît alors très élevée. Dans un échantillon trop hétérogène, le coefficient de fidélité est donc surestimé. On tend à conclure que le test est fidèle, alors que, d'une passation à l'autre, un même individu peut avoir des performances très éloignées. Sauf que comme les individus sont trop différents les uns des autres, les importantes variations de performances n'amènent pas forcément à un changement de rangs. La plupart des individus gardent leur place dans le classement ; la corrélation apparaît alors plus élevée que si les individus montraient une différenciation moins importante.

L'échantillon d'étalonnage/de standardisation d'un test est constitué pour être représentatif de la population à qui s'adresse le test. Il est généralement d'une taille importante et présente une étendue des scores observés la plus large possible. C'est à partir de cet échantillon que les normes du test sont établies. Pour les études de fidélité, l'échantillon est souvent constitué d'un nombre moindre d'individus ou d'individus relativement homogènes (p. ex., les étudiants de psychologie). Pour corriger la variabilité dans l'échantillon d'étude par rapport à la variabilité dans l'échantillon de standardisation, une formule est proposée par Magnusson (1967), appelée la correction de Magnusson qui s'énonce comme suit :

$$r_{UU'} = 1 - \frac{s_X^2(1 - r_{XX'})}{s_U^2} \qquad (16)$$

Où le coefficient corrigé pour l'échantillon d'étude ($r_{UU'}$) tient ainsi compte de la variance de l'échantillon de standardisation (s_X^2), du coefficient de fidélité de l'échantillon de standardisation($r_{XX'}$) et de la variance de l'échantillon d'étude (s_U^2). Nous appliquerons la correction de Magnusson sur les coefficients test-retest de notre étude afin de corriger toute éventuelle restriction des différences interindividuelles par rapport à l'échantillon de standardisation de l'adaptation en français du WISC-IV.

3.3.1.2. Longueur du test

Le nombre d'items influe directement sur la précision de la mesure. De même que, pour un coefficient de corrélation, l'estimation des caractéristiques de la population générale est d'autant plus précise que le sous-échantillon tiré est grand, le coefficient de fidélité des scores d'un test augmente à mesure que le test comporte d'items évaluant le même attribut. En effet, « la somme des erreurs aléatoires de mesure devrait tendre vers zéro lorsqu'un grand nombre d'items est utilisé » (Laveault & Grégoire, 2014, p. 126).

Afin de calculer l'influence du nombre d'items sur le coefficient de fidélité, la formule de Spearman-Brown[16] qu'on applique pour corriger la sous-estimation du coefficient *split half* peut être généraliser de la manière suivante :

$$r_{xx'} = \frac{kr_{jj'}}{1 + (k-1)r_{jj'}} \qquad (17)$$

Où $r_{xx'}$ représente le coefficient de fidélité attendu du test modifié, k est le facteur d'allongement du test (p. ex., $k = 2$ dans le cas de la méthode de bissection) et $r_{jj'}$ est le coefficient de fidélité initial du test. Ainsi, on peut calculer le coefficient de fidélité attendu d'un test sur lequel on aurait ajouté ou supprimé une proportion k d'items parallèles aux autres items du test. Si nous voulions savoir la proportion d'items (de même difficulté et de même contenu) à rajouter pour augmenter jusqu'au degré de fidélité visé, on peut également isoler k dans la formule (17) et obtenir l'équation suivante :

$$k = \frac{r_{xx'}(1 - r_{jj'})}{r_{jj'}(1 - r_{xx'})} \qquad (18)$$

3.3.1.3. Difficulté d'un test

Lorsqu'un test est trop facile ou trop difficile, la distribution des scores sur le test ne suit plus une distribution normale (c.-à-d. symétrique de part et d'autre de la moyenne, voir Figure 32, p. 141). Les résultats sur le test vont se décaler vers les scores maximaux du test et entraîner une distribution asymétrique négative dans la situation d'un test trop facile (voir Figure 32a). À l'inverse, les résultats sur le test vont se décaler vers les scores minimaux du test et entraîner une distribution asymétrique positive dans

[16] Voir Équation (15), p. 132.

la situation d'un test trop difficile (voir Figure 32c). Dans les cas de distributions asymétriques, la corrélation de Bravais-Pearson ne peut plus atteindre sa valeur maximale de 1, même théoriquement. Le coefficient de fidélité se voit donc affaibli si la difficulté du test est trop basse ou trop élevée pour l'échantillon qui sert à son estimation.

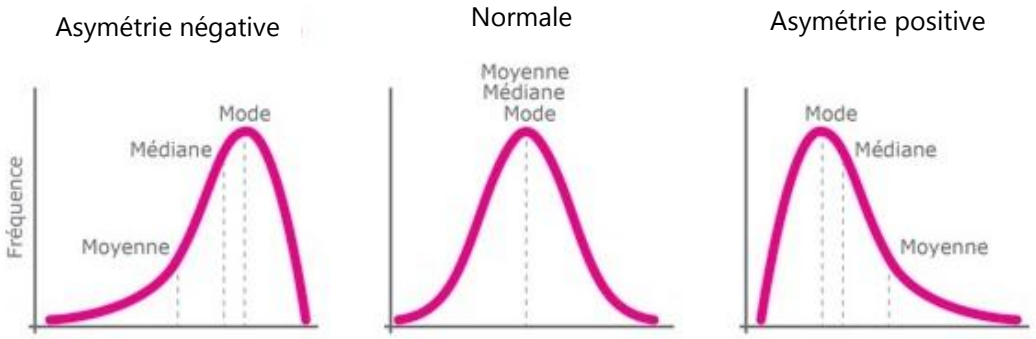


Figure 32. Distributions asymétrique négative (32a), normale (32b) et asymétrique positive (32c). Source : WikiStat (https://www.tns-ilres.com/cms/Home/WikiStat/Asymetrie-et-aplatissement).

Nous venons de voir les facteurs qui influencent la valeur du coefficient de fidélité, soit en la surestimant soit en la sous-estimant. Avant même de s'intéresser à la valeur du coefficient, le travail d'interprétation débute en appréciant les possibles influences desdits facteurs ainsi que de la méthode d'estimation utilisée. De plus, nous rappelons que la fidélité est établie à la fois pour les scores d'un test et pour l'échantillon sur lequel elle a été estimée. Sans étude, on ne peut pas généraliser une donnée de fidélité à d'autres échantillons et d'autres contextes d'évaluation. Dans la suite, nous allons nous pencher sur l'interprétation à donner aux valeurs des coefficients de fidélité.

3.3.2. SEUILS POUR LES COEFFICIENTS DE FIDÉLITÉ

À la lecture des données psychométriques d'un manuel de test, le clinicien s'interroge forcément sur l'interprétation à donner à telle ou telle valeur d'un coefficient de fidélité. Certains seuils sont proposés dans la littérature pour guider l'interprétation, néanmoins, ils ne doivent pas être considérés comme des standards absolus. En effet, il s'agit davantage de repères, puisque ces valeurs seuils varient légèrement d'un auteur à l'autre. Les auteurs s'accordent sur une recommandation : la

détermination d'un seuil de fidélité dépend du contexte de l'évaluation et de l'utilisation qui est faite des résultats (Abell, Springer, & Kamata, 2009; Nunnally & Bernstein, 2010; Salvia, Ysseldyke, & Bolt, 2012; Thorndike & Thorndike-Christ, 2010). Des auteurs comme Cicchetti (1994) et, Murphy et Davidshofer (2001) donnent des lignes directrices, à prendre à titre indicatif. Pour eux, un coefficient de fidélité inférieur à .60 est à considérer comme insuffisant. Il est faible si avoisinant .70, modérément élevé à partir de .80 et, peut être considéré comme élevé à .90 et plus. Pour des observations sur un groupe d'individus et à des fins de recherche, certains auteurs considèrent un coefficient de .70 comme un minimum (Abell et al., 2009; R. M. Thorndike & Hagen, 1977; Wasserman & Bracken, 2013), tandis que d'autres déterminent le minimum à .80 (Nunnally & Bernstein, 2010). Lorsqu'on utilise le coefficient de fidélité à des fins cliniques, des valeurs plus élevées que .80 sont recommandées (Wasserman & Bracken, 2013). Nunnally et Bernstein suggèrent au moins des valeurs de .90, voire .95, lorsqu'il s'agit de décisions à fort enjeu pour l'individu (p. ex., diagnostic, placement spécialisé, sélection dans un programme personnalisé).

Comme le psychologue doit apprécier selon les situations le seuil acceptable, il est important de comprendre le sens d'une valeur de coefficient de fidélité. Pour aider à l'interprétation, on peut traduire la fidélité en proportion de variance vraie. Par exemple, un coefficient de fidélité de .80 indique que 80 % de la variabilité dans les scores observés au test est expliquée par de la variance vraie sur la propriété mentale évaluée par le test. Autrement dit, 80 % des différences interindividuelles mises en évidence par le test sont attribuables à de « vraies » différences entre les individus sur ce qu'évalue le test. En outre, on déduit que 20 % de la variabilité dans les scores observés au test est expliquée par de la variance d'erreur. Selon la situation en jeu, accepter un risque d'erreur de 20 % (soit 1 chance sur 5 de se tromper) peut être inacceptable. Même si de prime abord un coefficient de fidélité de .80 peut paraître élevé, il n'est parfois pas raisonnable de prendre une décision sur la base des uniques résultats d'un tel test.

Étant un coefficient de corrélation, le coefficient de fidélité peut aussi se traduire en terme de changement de rang des individus dans un groupe. Dans cette perspective, on peut se demander quelle est la probabilité d'un changement de position associée à un coefficient de fidélité. Reprenons la situation présentée par Thorndike et Hagen (1977) ; l'ensemble des résultats sont reportés dans le Tableau 3 (p. 143). Il s'agit d'une situation d'évaluation de deux sujets A et B à qui l'on administre un même test à deux reprises. Lors de la première passation de test, l'individu A obtient un score qui le situe

au rang percentile de 75 (soit parmi les 25 % des meilleures performances sur le test) et l'individu B, quant à lui, obtient un score qui le situe dans le percentile 50 (soit parmi les performances moyennes sur le test). Quelle est la probabilité que A et B interchangent leur position si l'on répète la mesure ?

Tableau 3

Pourcentage de fois où un renversement de position entre A et B se produit à la suite d'une mesure répétée pour des scores initialement au rang percentile 75 et 50 (R. M. Thorndike & Hagen, 1977, p. 93)

	Pourcentage de renversement de position suite à une mesure répétée du test		
Coefficient de fidélité	Score d'un individu	Moyenne des scores d'un groupe de 25	Moyenne des scores d'un groupe de 100
.00	50.00	50.00	50.00
.40	40.30	10.90	0.70
.50	36.80	4.60	0.04
.60	32.50	1.20	
.70	27.10	0.10	
.80	19.70		
.90	8.70		
.95	2.20		
.98	0.05		

Si le test a une fidélité nulle (r = 0), il y a exactement 50 % de chance que A et B interchangent leur position lors de la seconde passation. Si le test a une fidélité de .70, la probabilité d'un changement de rang entre A et B serait de 27.1 %. Pour une fidélité de .80, la probabilité d'un renversement de position est de 19.7 % et finalement, la probabilité est de 2.2 % pour une valeur de fidélité de .95.

Thorndike et Hagen (1977) montrent également que les probabilités d'un renversement de position entre les rangs percentiles 50 et 75 sur un groupe de 25 individus et sur un groupe de 100 individus. Par exemple, A est une classe de 25 élèves qui présentent des performances se situant en moyenne au rang percentile de 75, tandis que B est une classe de 25 élèves qui présentent des performances se situant en moyenne au rang percentile de 50. Quelle est la probabilité que les performances moyennes de la classe A et de la classe B permutent leur position si l'on répète la mesure ? Pour un coefficient nul, la probabilité demeure de 50 % de chance d'un renversement de position. En revanche, avec un coefficient de .70, la probabilité devient

très faible à 0.1 % de chance. Plus les conclusions portent sur un groupe nombreux, plus la sécurité dans ces conclusions augmente rapidement avec le coefficient de fidélité. Cela souligne l'importance d'apporter des résultats non seulement sur la fidélité au niveau du groupe, mais également sur la fidélité au niveau de l'individu. Thorndike et Hagen concluent ainsi :

> *A test with relatively low reliability will permit us to make useful studies of and draw accurate conclusions about groups, especially groups of substantial size, but quite high reliability is required if we are to speak with confidence about individuals.* (1977, p. 94)

Dans une étude sur la relation entre la fidélité des scores d'un test et la prise de décision, Charter et Feldt (2001) examinent différents niveaux de fidélité et comparent les pourcentages d'individus ayant été correctement identifiés comme ayant besoin d'une prise en charge clinique (vrais positifs) ou n'ayant pas besoin d'une prise en charge (vrais négatifs) ainsi que les pourcentages des individus qui sont identifiés à tort comme ayant besoin d'un traitement (faux positifs) ou identifiés à tort comme n'ayant pas besoin d'un traitement (faux négatifs). Charter et Feldt (2001) présentent deux situations : un test A avec une fidélité de .90 et un test B avec une fidélité de .70. Les deux tests A et B ont une moyenne de 100 et un écart type de 20. Pour déterminer si l'individu a besoin d'un traitement, le critère est qu'il obtient des performances inférieures à 74. Ce cut-off à 74 correspond à -1.3 écart type de la distribution du test (c.-à-d. 100 - 74 = 26 => 26/20 = 1.3). Les scores au test inférieurs à 74 situe l'individu parmi les 10 % les plus faibles dans la population. Selon une distribution normale bivariée, Charter et Feldt (2001) montrent que si le score du test a une fidélité parfaite de 1, on trouverait 10 % de vrais positifs (ceux correctement identifiés comme ayant besoin d'une prise en charge clinique), 90 % de vrais négatifs (ceux correctement identifiés comme n'ayant pas besoin d'une prise en charge), 0 % de faux positifs (ceux identifiés à tort comme ayant besoin d'un traitement) et 0 % de faux négatifs (ceux identifiés à tort comme n'ayant pas besoin d'un traitement). Dans ce cas idéal de fidélité parfaite, le test permettrait de prendre 100 % de décisions correctes. Pour la situation du test A qui a une fidélité de .90, il y a 7.8 % de vrais positifs, 87.8 % de vrais négatifs, 2.2 % de faux positifs et 2.2 % de faux négatifs. Dans ce cas, 4.4 % des décisions sur la base de ce test sont incorrectes (soit 2.2 % de faux positifs + 2.2 % de faux négatifs). Pour des situations plus fréquentes comme celle du test B qui a une fidélité de .70, il y a 6 % de vrais positifs, 86 % de vrais négatifs, 4 % de faux positifs et 4 % de faux négatifs. Dans ce cas, 8 % des décisions sont incorrectes (soit 4 % de faux

positifs + 4 % de faux négatifs). Sans surprise, le risque d'erreur sur les décisions augmente avec la diminution de la fidélité du score du test.

Charter et Feldt (2001) montrent également une relation en lien avec le choix du cut-off. Dans le premier exemple, le cut-off pour décider de la nécessité d'une prise en charge est défini aux performances de 74 et inférieures, ce qui représentent les 10 % des performances les plus faibles pour des scores du test de moyenne 100 et d'écart type 20. Généralement, on définit les seuils des performances faibles à -1 écart type (ce qui en représente les 15.9 % des performances les plus faibles), -1.5 écart type (ce qui en représente les 6.7 % des performances les plus faibles) ou -2 écarts types en dessous de la moyenne (ce qui en représente les 2.3 % des performances les plus faibles). En nous référant aux propriétés de la distribution normale, si le test a une fidélité du score parfaite (r = 1), nous trouverions donc des pourcentages de vrais positifs de 15.9 % pour un cut-off à -1 écart type, de 6.7 % pour un cut-off à -1.5 écart type et de 2.3 % pour un cut-off à -2 écarts types, ce qui correspondraient à chaque fois à 100 % de prises décisions correctes pour le test. Nous avons vu précédemment que la diminution de la fidélité des scores du test augmente le risque de se tromper (faux positif et faux négatif). Charter et Feldt (2001) montrent que le choix du cut-off influence aussi la probabilité de se tromper. Si le cut-off est proche de la moyenne, le pourcentage de chance de prendre une décision correcte est plus élevé que pour un cut-off plus éloigné de la moyenne. Charter et Feldt (2001) donnent l'exemple d'un test qui a une fidélité de .80. Pour un cut-off à -1 écart type, il y a 72 % des 15.9 % qui ont besoin de traitement qui le recevront effectivement au lieu de 100 % des 15.9 % avec une fidélité parfaite (r = 1). Pour un cut-off à -1.5 écart type, il y a 65 % des 6.7 % qui ont besoin de traitement qui le recevront effectivement au lieu de 100 % des 6.7 % avec une fidélité parfaite. Enfin, si on choisit un cut-off à -2 écarts types, il y a 57 % des 2.3 % qui ont besoin de traitement qui le recevront effectivement au lieu de 100 % des 2.3 % avec une fidélité parfaite. Ainsi, « *when the reliability is held constant a cut-off score close to the mean is more efficient (higher correct classifications) than a cut-off farther from the mean* » (Charter & Feldt, 2001, p. 533). En conclusion, les résultats de Charter et Feldt (2001) conduisent à une valeur de fidélité de .98 ou plus si on souhaite s'assurer de 90 % de décisions correctes pour les individus en nécessité de traitement quel que soit le cut-off. Avec des fidélités inférieures à .98, il faut tenir compte à la fois de la fidélité et du cut-off choisi pour évaluer le risque d'erreur. La méthode proposée par Charter et Feldt (2001) illustre la difficulté de l'interprétation des données psychométriques,

lorsqu'on souhaite les utiliser pour des cas individuels. Bien conscients de la complexité, ils constatent :

> *If test score interpretation were a science we would not need highly trained experts for the job; a monkey with the ability to recognize numbers and enter them into a computer could do it.* (Charter & Feldt, 2001, p. 536)

L'utilisation d'un coefficient de fidélité ne repose pas sur l'application de critère-seuil. Il y a toujours une réflexion sur la méthode qui a permis son estimation, sur l'échantillon étudié et sur les enjeux du contexte d'utilisation du test. Outre la difficulté à interpréter un coefficient de fidélité, il peut aussi être mal aisé de le mettre en relation avec le score au test d'un individu particulier. En effet, le coefficient de fidélité n'est pas immédiatement parlant pour le clinicien face au score observé d'un sujet à un test. Pour répondre aux besoins d'une interprétation des scores individuels, l'erreur type de mesure s'utilise pour construire un intervalle de confiance autour du score observé. Nous allons expliciter ces deux concepts dans ce qui suit.

3.3.3. ERREUR DE MESURE ET INTERVALLE DE CONFIANCE

Nous l'avons bien compris, le score d'un test est d'autant plus fidèle qu'il s'agit d'une mesure peu entachée d'erreur de mesure. Les erreurs de mesure sont inévitables d'une passation à l'autre et leurs origines sont variées. On les définit comme toutes fluctuations aléatoires dans la mesure qui ne rendent pas compte de différences interindividuelles sur ce que le test prétend évaluer. Toutefois, rappelons que la théorie classique des tests (TCT) postule des distributions normales pour les scores observés et pour les erreurs de mesure lors de mesures répétées sur un même sujet et avec un même test. Cela signifie que si l'on fait passer le test un grand nombre de fois à un même individu, la distribution de ses scores observés est normale et a pour moyenne son score vrai. De même, la distribution des erreurs sur ces mesures répétées est supposée suivre la loi normale et avoir une moyenne qui tend vers zéro. Au niveau d'un échantillon, la TCT suppose que tous individus (quel que soit leur niveau d'habileté sur la propriété mentale évaluée) ont la même dispersion des erreurs pour un test particulier. À partir de ces postulats, on définit un écart type des erreurs pour l'échantillon, appelé erreur type de mesure. L'Équation (14)[17] peut alors se dériver

[17] Voir p. 126.

jusqu'à obtenir la formule de l'erreur type de mesure (*standard error of measurement* dans la littérature anglophone) :

$$\frac{s_e^2}{s_x^2} = 1 - r_{xx'} \tag{19}$$

$$s_e^2 = s_x^2(1 - r_{xx'}) \text{ d'où } s_{em} = s_x\sqrt{1 - r_{xx'}} \tag{20}$$

L'erreur type de mesure (s_{em} ou ETM pour la notation française) est basée sur la métrique du score observé. Plus un test est fidèle (c.-à-d. $r_{xx'}$ proche de 1), « plus la variance des scores observés est due à la variance des scores vrais et non à des fluctuations du hasard » (Laveault & Grégoire, 2014, p. 112). Ainsi, plus le coefficient de fidélité $r_{xx'}$ est proche de 1, plus l'erreur type de mesure tend vers zéro et plus la variance dans les scores observés est expliquée par la variance des scores vrais (ou variance vraie). À l'inverse, plus le coefficient de fidélité $r_{xx'}$ est faible, plus l'erreur type de mesure tend vers la valeur de l'écart type de la distribution des scores observés. Dans la situation extrême d'une fidélité nulle ($r_{xx'} = 0$), toute la variance dans les scores observés n'est alors que de la variance d'erreur.

Alors que le coefficient de fidélité informe sur la fidélité du score d'un test dans une perspective de comparaison interindividuel, l'utilité de l'ETM est plus parlante lorsqu'on s'intéresse au score d'un individu particulier. En effet, ce dernier peut s'interpréter comme une marge d'erreur autour du score observé et permet ainsi de construire un intervalle de confiance. L'utilisation d'un intervalle de confiance autour du score observé au test est une conséquence de la fidélité imparfaite des scores. Sachant que tout score observé est entaché d'erreur, le clinicien ne peut pas s'appuyer sur un score unique.

Illustrons l'utilisation de l'ETM et de l'intervalle de confiance par l'exemple d'un enfant qui a obtenu un score QI Total de 129 au WISC-IV. Tout d'abord, on introduit dans la formule de l'ETM[18], le coefficient de fidélité fourni dans le manuel d'interprétation du WISC-IV[19] et l'écart type de la distribution des scores QI. On obtient le calcul suivant : $\text{ETM} = 15\sqrt{1 - .94} = 3.67$. L'erreur type de mesure autour du score observé est donc de ± 3.67, soit un intervalle de confiance [125 ; 133]. Il s'agit ici d'un intervalle de confiance associé à une probabilité de 68 %. En effet, nous l'avons déjà dit, l'ETM représente l'écart type de la distribution théorique des erreurs liées à une mesure

[18] Voir Équation (20), p. 143.

[19] Voir Annexe F.

qu'on répèterait une infinité de fois sur un individu donné ; la distribution est présumée gaussienne. Théoriquement et en vertu des propriétés de la courbe normale, 68 % des scores observés d'un sujet doivent tomber dans l'intervalle de ± 1 ETM autour de son score vrai. De même, théoriquement 95 % des scores observés doivent tomber dans l'intervalle de ± 2 ETM autour de son score vrai et 99 % des scores observés doivent tomber dans l'intervalle de ± 3 ETM autour de son score vrai. Comme le score vrai est inaccessible, l'ETM est appliqué sur le score observé. Pour construire un intervalle de confiance avec des probabilités telle que 90 %, 95 % ou 99 %, il faut multiplier l'ETM par la valeur critique de *z* associée au niveau de confiance choisi, soit respectivement 1.645, 1.96 ou 2.58.

Du fait que l'ETM est appliqué sur le score observé, et non sur le score vrai, certaines précisions sont à relever. Reprenons l'exemple de notre enfant qui a un QI Total de 129. Si l'on répète la mesure, on s'attend à ce que 95 % de ses QIT observés se situent à l'intérieur de l'intervalle de ± 2 ETM autour de son score vrai. Or, il nous est impossible de connaître le score vrai. La mesure effectuée par un test est un score observé, qui étant biaisé, ne corrèle pas parfaitement avec le score vrai (Laveault & Grégoire, 2014 ; Nunnally & Bernstein, 2010). En effet, les scores des tests cognitifs n'échappent pas au phénomène statistique de la régression vers la moyenne. C'est en observant la taille d'enfants dont les parents sont plus grands ou plus petits que la moyenne que Galton découvre ce phénomène, qu'il appelle la loi de régression filiale (*law of filial regression*). Il remarque que les enfants de parents de grande taille sont souvent plus grands que la moyenne de la population, mais plus petits que leurs parents. À l'inverse, les enfants de parents de petite taille sont plus petits que la moyenne, mais plus grands que leurs parents. Plus généralement, le phénomène de régression vers la moyenne traduit une plus grande probabilité des scores extrêmes à se rapprocher de la moyenne plutôt qu'à s'en éloigner lors de mesures ultérieures. Pour tenir compte de ce biais, l'ETM est remplacée par l'erreur type d'estimation (ETE).

$$ETE = r_{xx'}\left(\sigma\sqrt{1 - r_{xx'}}\right) = r_{xx'}(ETM) \tag{21}$$

La particularité d'un intervalle de confiance construit avec l'ETE est d'être centré sur un score vrai estimé (V_{est}) qu'on obtient par le calcul suivant :

$$V_{est} = \bar{X} + r_{xx'}\left(X - \bar{X}\right) \tag{22}$$

Où $\bar{X}$ est la moyenne théorique des scores observés, X est le score observé au test et $r_{xx'}$.est le coefficient de fidélité. Notre QIT observé de 129 devient un QIT vrai estimé : $V_{est} = 100 + .94(129 - 100) = 127.26$. Pour un niveau de confiance de 95 %, l'ETE est de ±6.76 (soit $.94 \times 3.67 = 3.45 => 3.45 \times 1.96 = 6.76$). Si on applique l'ETE calculé pour un niveau de confiance à 95 % au score vrai estimé, cela donne les bornes arrondies [120 ; 134] autour du QIT observé de 129. Étant centré sur le score vrai estimé (et non le score observé), l'intervalle de confiance calculé avec l'ETE est asymétrie par rapport au score observé lorsqu'il s'agit de scores éloignés de la moyenne. Pour des scores observés au-dessus de la moyenne (comme notre exemple), la borne gauche est plus large, tandis que la borne droite est plus large pour des scores observés en dessous de la moyenne. Par exemple, pour un QIT observé de 60, cela donne les bornes arrondies [55 ; 70] pour un niveau de confiance à 95 %. L'asymétrie sera d'autant plus marquée que les scores sont très éloignés de la moyenne et/ou que la fidélité diminue.

Même si la procédure qui tient compte du biais de régression vers la moyenne (c.-à-d. ETE) est plus rigoureuse que la procédure de l'erreur type de mesure (ETM), il demeure néanmoins certaines limites à l'utilisation des intervalles de confiance. L'une des limites est le postulat d'homoscédasticité selon lequel tous les individus (quel que soit leur niveau d'habileté sur la propriété mentale, quel que soit leur état au moment de la passation, etc.) sont touchés par une erreur de mesure de même ampleur. En effet, on applique la même erreur type de mesure (ETM ou ETE) sur chaque score observé, lorsqu'on calcule un intervalle de confiance. Ce postulat n'est pas forcément avéré ni réaliste pour tous les individus et dans toutes les situations. Rappelons que l'ajustement pour un éventuel manque d'homoscédasticité est, en revanche, possible dans le cadre de la théorie des modèles à l'item, puisqu'on peut calculer différentes erreurs types de mesure en fonction du niveau d'habileté sur la propriété mentale évaluée et des caractéristiques de l'item.

Une autre limite majeure est l'interprétation erronée qu'on prête aux intervalles de confiance. Si l'on reprend l'exemple de notre enfant avec un QIT de 129, on glisse souvent de l'interprétation correcte : « Si on répétait une infinité de fois la mesure, on s'attend à ce que 95 % de ses QIT observés se situent à l'intérieur de l'intervalle de ±6.76 autour de son score vrai » à l'énoncé erroné : « l'intervalle de confiance à 95 % autour de son score observé signifie qu'il y a 95 % de chances que son score vrai soit compris entre 120 et 134 ». Dans l'énoncé erroné, on cherche à probabiliser les valeurs possibles du paramètre (c.-à-d. le score vrai) à partir des données recueillies lors d'une seule mesure. Or, l'approche fréquentiste ne permet pas de se prononcer sur « *the*

probability that a particular, observed confidence interval contains the true value » (Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2015, p. 105). En effet, la probabilité fréquentiste – formalisée $P(D|H)$ – porte sur les probabilités d'échantillonnage, conditionnelles au paramètre estimé (ici le score vrai). Une fois le score observé (c.-à-d. l'événement réalisé), il n'y a plus aucune probabilité, sinon une probabilité soit de 1 (contient le score vrai) soit de 0 (ne contient pas le score vrai). Ce n'est pas notre ignorance quant à la valeur du score vrai qui crée la probabilité, mais la procédure d'échantillonnage (c.-à-d. le choix aléatoire de l'échantillon qui servira à estimer l'intervalle de confiance). Inconnu mais constant, le score vrai n'est ni un paramètre aléatoire ni déterminé par l'observation des données. En revanche, les bornes des intervalles sont « des grandeurs aléatoires, qui varient d'un échantillon à un autre » (Lecoutre, 2005, p. 32). Ainsi, l'interprétation fréquentiste correcte d'un Intervalle de Confiance à 95 % (IC 95 %) s'énonce comme suit : « 95 % des intervalles calculés sur l'ensemble des échantillons possibles (tous ceux qu'il est possible de tirer) contiennent la vraie valeur » (Lecoutre, 2005, p. 93). Dit autrement, « si on répétait un grand nombre de fois l'expérience et si on notait à chaque fois l'intervalle ainsi trouvé, alors dans [X %] des cas en moyenne, la vraie valeur . . . se trouverait dans l'intervalle » (D'Estampes, Garel, & Saint Pierre, 2003, p. 77). On peut donc assigner une probabilité sur la procédure d'échantillonnage par laquelle on estime les intervalles de confiance (*long-run repetition of the same experiment*), mais on ne peut pas probabiliser sur un intervalle spécifique (*single case*). De plus, le cadre fréquentiste ne permet pas d'inférences en *post-data* (c.-à-d. une fois les données observées). Il s'avère donc « *incorrect to interpret a CI as the probability that the true value is within the interval . . . . As is the case with p-values, CIs do not allow one to make probability statements about parameters or hypotheses* » (R. Hoekstra, Morey, Rouder, & Wagenmakers, 2014, p. 1159).

Les interprétations erronées sur les intervalles de confiance sont largement diffusées dans les ouvrages de référence en psychométrie et au sein de la communauté scientifique. Morey et al. (2015) relèvent trois catégories d'erreurs dans l'interprétation des intervalles de confiance. Ils nomment la première : la *Fundamental Confidence fallacy*. Cette erreur fondamentale nous conduit à déclarer à tort que l'intervalle calculé à partir d'une unique mesure a une probabilité de X % de contenir le vrai score. Une fois le score observé obtenu (c.-à-d. l'événement réalisé), il n'y a plus réellement de probabilités. La probabilité que le score vrai soit d'être compris dans les limites d'un intervalle de confiance qui entoure un score observé au test est de 100 % (le score vrai

est dedans) ou 0 % (le score vrai n'est pas dedans). C'est avant d'obtenir le score observé au test qu'il y a une probabilité de X % que l'intervalle de confiance à calculer contient le score vrai. La deuxième erreur est nommée la *Precision fallacy*. Elle nous conduit à déclarer à tort que l'étendue de l'intervalle calculé est une indication du degré de précision du paramètre estimé. Plus l'intervalle de confiance calculé est étroit, plus on est proche de l'estimation du score vrai. Cela peut sembler une affirmation correcte puisque, plus la fidélité est élevée, plus l'intervalle de confiance calculé est petit. Sauf qu'à nouveau, nous ne pouvons pas savoir si le score vrai est à l'intérieur ou non de l'intervalle calculé. Si le score vrai est dans l'intervalle calculé, alors effectivement l'affirmation est correcte, en revanche, si le score vrai ne l'est pas dans l'intervalle calculé, l'affirmation est complètement erronée. Enfin, la troisième erreur est nommée la *Likelihood fallacy*. Cette dernière nous conduit à déclarer à tort qu'à l'intérieur de l'intervalle calculé se trouvent les valeurs les plus plausibles du score vrai. Toujours pour la même raison, l'affirmation est incorrecte puisqu'on ne peut jamais savoir si le score vrai est ou non à l'intérieur de l'intervalle calculé à partir d'un unique score observé.

Si la confusion persévère dans l'esprit des utilisateurs des intervalles de confiance, la question était déjà débattue autrefois par les fondateurs des statistiques modernes, dont le célèbre mathématicien Jerzy Neyman (1894 – 1981) :

> *Consider now the case when a sample . . . is already drawn and the [confidence interval] given . . . . Can we say that in this particular case the probability of the true value of [the parameter] falling between [the limits] is equal to [X %]? The answer is obviously in the negative. The parameter . . . is an unknown constant and no probability statement concerning its value may be made.* (Neyman, 1937, p. 349)

Pour le théoricien Neyman, il n'y a aucune ambigüité. Le cadre des statistiques fréquentiste ne permet pas l'interprétation usuelle qu'on fait des intervalles de confiance autour d'un score observé.

Si les postulats fréquentistes ne coïncident pas avec l'interprétation intuitive qu'on désire attribuer à un intervalle de confiance spécifique, les inférences bayésiennes, en revanche, probabilisent explicitement les valeurs possibles du paramètre, une fois les données recueillies (Lecoutre, 2005). Au lieu de considérer les probabilités d'échantillonnage $P(D|H)$ comme le fait les inférences fréquentistes, les inférences bayésiennes combinent l'information provenant de trois sources : (1) la probabilité de la croyance a priori (*prior belief*, souvent abrégé *prior* dans la littérature anglophone), (2) les données de l'expérience et (3) la probabilité de la croyance a

posteriori qui est une mise à jour des croyances initiales compte tenu des données recueillies. La croyance a priori (*prior*) reflète la probabilité d'un événement d'après les connaissances établies au sujet dudit événement par les études et des expériences antérieures. L'apport de nouvelles informations grâce aux données observées actualise cette croyance initiale qui devient une croyance a posteriori. Dans sa forme simple, le théorème de Bayes s'exprime comme suit :

$$P(H|D) = \frac{P(H) \times P(D|H)}{P(D)} \tag{23}$$

Avec $P(H|D)$ qui représente la probabilité a posteriori de l'hypothèse sachant les données, $P(H)$ représente la probabilité a priori de l'hypothèse, $P(D|H)$ représente la fonction de vraisemblance de l'hypothèse et $P(D)$ est la probabilité a priori des données. L'approche bayésienne « exprime donc directement l'incertitude sur la vraie valeur . . . par des probabilités [$P(H|D)$], conditionnelles aux données » (Lecoutre, 2005, p. 96). Tenant compte des données recueillies, l'approche bayésienne permet de choisir un intervalle au vu de l'échantillon particulier observé, qui est appelé intervalle de crédibilité. Contrairement à l'intervalle de confiance, l'intervalle de crédibilité nous permet de déclarer qu'il y a X % de chances que le score vrai soit compris entre les bornes de l'intervalle calculé à partir des données observées. En probabilisant sur le paramètre au vu des données, les inférences bayésiennes apportent une interprétation directe et naturelle aux intervalles de confiances. Nous en restons à ce bref survol, car un développement théorique, philosophique ou mathématique des statistiques bayésiennes dépasserait le cadre du présent travail.

Dans la pratique actuelle, il y a donc un télescopage entre les inférences fréquentistes dans lesquelles sont estimés les intervalles de confiance et les inférences bayésiennes dans lesquelles ils sont interprétées. Des auteurs expliquent ce non-sens comme « le produit d'une évolution complexe, au cours de laquelle les idées des fondateurs ont été en partie *occultées* et *mélangées*, ce qui a donné naissance à un mode de pensée "hybride" qui, malgré la nature inconciliable de ses éléments, perdure » (Capel, Monod, & Müller, 1997, p. 133). En l'état des choses, nous rejoignons Laveault et Grégoire (2014), « l'avantage majeur à déterminer un intervalle de confiance autour de la note obtenue est de relativiser cette dernière note. Le praticien prend ainsi mieux conscience de la marge d'erreur que comporte la mesure recueillie » (p. 129). On ne peut donc pas formuler de probabilité sur l'inclusion du score vrai dans les limites d'un intervalle de confiance appliqué autour d'un score observé. Toutefois, l'étendue de

l'intervalle de confiance donne un ordre de grandeur sur l'erreur de mesure associée au score observé, et donc sur la fidélité du score.

3.4. FIDÉLITÉ – STABILITÉ DES SCORES

Après cette première moitié de chapitre consacrée à la fidélité dans son ensemble, nous allons maintenant recentrer sur la fidélité des scores entre deux temps de mesures. S'agissant d'évaluer la fidélité des scores dans le temps, cela relève plus particulièrement de la notion de stabilité. Certains auteurs marquent une distinction entre le concept de fidélité et le concept de la stabilité (Baltes, Reese, & Nesselroade, 1997; Gergen, 1982). Pour ces derniers, le terme de stabilité ne s'attribue pas aux scores du test, mais à la propriété mentale évaluée par le test. Dans la distinction des deux concepts, l'un porte le focus sur l'instrument de mesure (fidélité), tandis que l'autre porte sur le phénomène que l'on évalue (stabilité). Ainsi, la fidélité est une propriété psychométrique et renseigne sur la répétabilité des scores du test, tandis que la stabilité est la propriété d'un processus psychologique et renseigne sur la répétabilité de ce que le test évalue (Nesselroade, Pruchno, & Jacobs, 1986). Pour notre propos, nous ne ferons pas cette distinction dans les termes. Nous considérons que le terme stabilité précise l'aspect de la fidélité qui évalue la répétabilité des scores dans le temps.

Dans un premier temps, nous définirons les types d'évaluation de la stabilité qu'on peut retrouver dans les recherches, à savoir la stabilité absolue (*mean level change*), la stabilité différentielle (*rank order consistency*) et la stabilité intra-individuelle absolue (*individual difference in change*). Dans un second temps, nous ferons une revue de la littérature centrée sur le WISC-IV et les études sur la stabilité de ses scores.

3.4.1. TYPES D'ÉVALUATION DE LA STABILITÉ

Pour la plupart des recherches menées sur la stabilité des scores d'un test, les résultats fournissent des coefficients de fidélité et renseignent donc au niveau groupal (stabilité différentielle). S'agissant d'instruments utilisés en clinique, il nous semble essentiel d'apporter des données également au niveau individuel. Dans notre étude, la

stabilité à long terme des scores est évaluée sur les deux niveaux afin de donner une vision aussi complète que possible.

Sur le plan interindividuel, deux aspects peuvent être évalués : la stabilité absolue et la stabilité différentielle. Dans une procédure test-retest, la stabilité absolue repose sur la comparaison des différences de moyennes entre une première et une seconde passation d'un même test aux mêmes individus. Il s'agit de déterminer, au niveau du groupe, si les moyennes des différents scores du test sont équivalentes entre les deux passations. La stabilité différentielle se réfère aux coefficients de corrélations calculés entre les scores de la première et de la seconde passation. Il s'agit de déterminer si l'ordre des individus est similaire entre les deux passations.

Sur le plan intra-individuel, on peut relever : la stabilité intra-individuelle absolue, la stabilité des forces et des faiblesses ainsi que la stabilité catégorielle. La stabilité intra-individuelle absolue examine dans quelle mesure une performance change d'une passation à l'autre. Il s'agit de déterminer le pourcentage d'enfants qui présentent des performances équivalentes entre les deux passations. Les auteurs considèrent un intervalle (±5 points, ±9 points, etc.) à l'intérieur duquel les différences de performances entre les deux passations sont équivalentes. La stabilité catégorielle est une évaluation du pourcentage d'enfants qui maintiennent leur performance dans la même catégorie descriptive d'une passation à l'autre (p. ex., performance faible, moyenne ou élevée). La stabilité des forces et des faiblesses examinent le pourcentage d'enfants qui d'une passation à l'autre maintient une force/moyenne/faiblesse pour tel indice.

3.4.2. STABILITÉ DES SCORES DU WISC

Dans l'ensemble, les travaux sur la stabilité des échelles de Wechsler ont été menés avec différents échantillons d'enfants (principalement des groupes cliniques) et différents intervalles test-retest (variant de quelques jours à plusieurs années). Selon la durée du délai retest, on parle d'études à court terme ou à long terme sans qu'il y ait de limite strictement définie dans la littérature pour distinguer le passage de l'un à l'autre. Toutefois, dans leurs travaux sur la stabilité à long terme du WISC-III, Canivez et Watkins (2001) relèvent qu'on n'observe pratiquement plus d'effets d'apprentissage (ou effet de pratique) lorsque les intervalles test-retest sont supérieurs à une année. Dans un autre article, ces mêmes auteurs parlent alors de long terme lorsque l'intervalle

dépasse 1 an (Watkins & Canivez, 2004). De même, Sattler (2008) considère comme étant du court terme un intervalle inférieur à une année. Dans la continuité de ces travaux, nous proposons de situer dans le court terme les intervalles test-retest inférieurs à une année, tandis que les intervalles test-retest égaux ou supérieurs à une année sont situés dans le long terme (≥ 1 an).

Les travaux sur les précédentes versions du WISC (WISC, WISC-R, WISC-III) documentent sur la stabilité des scores avec différents délais et différents groupes cliniques. Dans le système d'éducation spécialisée américain, les réévaluations des enfants et des adolescents s'effectuent régulièrement tous les trois ans, rendant aisée la récolte d'un grand nombre de données longitudinales avec des populations cliniques. La plupart des études américaines sur les précédentes éditions du WISC trouvent des coefficients test-retest supérieurs à .70 pour le score QIT (p. ex., Bauman, 1991; Canivez & Watkins, 1998, 2001; Elliott et al., 1985; Oakman & Wilson, 1988; Stavrou, 1990; Truscott, Narrett, & Smith, 1994; Vance, Blixt, Ellis, & Debell, 1981). Quant aux études avec des échantillons non cliniques, elles sont rares, et portent principalement sur la stabilité à court terme (Wechsler, 1949, 1974, 1991). Beaucoup de changements au niveau de la structure interne et des subtests qui composent les indices ont été apportés sur la 4^{e} édition, rendant obsolètes les résultats et constatations sur les précédentes éditions du WISC. Pour rappel, la fidélité des scores d'un test est dépendante de la composition d'un test et de l'échantillon testé. Comme la structure du test (items modifiés, subtests remplacés, etc.) est différente d'une édition à l'autre, la fidélité doit être estimée à nouveau. En fournissant des données sur la stabilité des scores du WISC-IV, avec un échantillon non clinique et non américain, notre étude contribue également à vérifier la généralisation des résultats américains à d'autres populations.

3.4.3. STABILITÉ À COURT TERME DU WISC-IV

Dans le Tableau 4 sont présentés les coefficients test-retest de différentes études sur la stabilité à court terme du WISC-IV. La première étude est menée sur 243 enfants issus de l'échantillon de standardisation de la version américaine du WISC-IV (Williams, Weiss, & Rolfhus, 2003). Les 243 enfants tout-venant ont été testés à deux reprises après un intervalle moyen de 32 jours. Les performances moyennes des indices varient de 99.8 (IMT) à 102.4 (IVT) lors de la première passation, et de 102.1 (ICV) à 109.5 (IVT) lors de la seconde passation. Les différences de moyennes entre les deux

passations montrent une augmentation des performances moyennes pour tous les scores. Les tailles de la différence (ou taille d'effet)[20] varient de 0.08 (Compréhension) à 0.60 (Complètement d'images) pour les subtests, et de 0.18 (ICV) à 0.51 (IVT) pour les indices. Les coefficients de corrélation corrigés révèlent une stabilité à court terme de .86 à .93 pour l'ensemble des indices. Quant aux scores des subtests, leurs coefficients de stabilité varient entre .76 et .92. Globalement, les coefficients de fidélité des subtests sont moins élevés que ceux des indices auxquels ils contribuent. Nous l'avons expliqué, toute chose étant égale par ailleurs, plus l'étendue des scores possibles est large, et plus élevé est le coefficient de corrélation. Pour la population américaine, les indices présentent une stabilité différentielle à court terme qui permet des prédictions au niveau groupal. En revanche, un effet d'apprentissage apparaît, et de façon non négligeable pour les indices de Raisonnement Perceptif et de Vitesse de Traitement.

Pour la version française du WISC-IV, un petit échantillon de 93 enfants tout-venant âgés de 6 à 15 ans est vu à deux reprises avec un intervalle de temps variant de 16 à 50 jours (moyenne = 27 jours) ; (Wechsler, 2005b). Les performances moyennes des indices varient de 100.1 (IMT) à 102.0 (IRP) lors de la première passation, et de 102.8 (ICV) à 113.6 (IVT) lors de la seconde passation. Conformément aux résultats américains, on relève également une augmentation de performances moyennes pour tous les scores (à l'exception de Compréhension) ; (voir Tableau 4, p. 158). Les tailles de la différence varient de -0.01 (Compréhension) à 0.73 (Complètement d'images) pour les subtests, et de 0.14 (ICV) à 0.81 (IVT) pour les indices. Les coefficients de fidélité corrigés de cette étude varient de .78 (IMT) à .91 (QIT). De même que, pour l'échantillon américain, la stabilité des subtests est également moins élevée que celle des indices (de .64 à .83). Notons que les coefficients de stabilité sur la version française sont légèrement plus faibles que ceux de la version américaine. La stabilité différentielle à court terme permet des prédictions sur le classement des individus au sein de leur groupe d'âge d'une passation à l'autre. Par contre, il faut tenir compte d'un effet d'apprentissage qui est prononcé pour les indices de Raisonnement Perceptif et de Vitesse de Traitement.

Dans une école privée du Midwest, Ryan, Glass et Bartels (2010) testent un petit échantillon de 43 enfants tout-venant âgés de 6 à 10 ans avec un intervalle d'environ 11 mois entre les deux passations. Soulignons d'abord que les performances moyennes des indices varient de 106.67 (IMT) à 112.56 (IRP) lors de la première passation, et de

[20] Selon Cohen (1977), la taille d'effet est considérée comme négligeable pour une valeur de $d < 0.2$, petite pour un d entre ≥ 0.2 et < 0.5, modérée pour un d entre ≥ 0.5 et < 0.8 et grande pour un $d \geq 0.8$.

108.09 (ICV) à 113.26 (QIT) lors de la seconde passation. Dans l'échantillon de cette étude, les performances moyennes des enfants sont donc à un niveau plus élevé que celles des enfants de l'échantillon de standardisation. Les auteurs réalisent des *t*-tests pour échantillons appariés qui ne montrent pas de différences significatives entre la performance moyenne de la première et de la seconde passation, et cela pour l'ensemble des scores. Les moyennes entre les deux passations sont donc équivalentes. À l'exception de l'IRP et de l'IVT, les coefficients de corrélation corrigés sont supérieurs à .70 pour la plupart des indices (voir Tableau 4, p. 158). De nouveau, les coefficients de stabilité des subtests sont moins élevés que ceux des notes composites (de .26 à .84). Au niveau individuel, Ryan et al. observent pour le QI Total, que 25.6 % des enfants augmentent leurs performances de plus de cinq points, tandis que 16.3 % des enfants diminuent de plus de cinq points. Légèrement plus de la moitié des enfants (58.1 %) présentent donc des scores compris entre ±5 points entre le test et le retest (voir Tableau 6, p. 164). Le choix d'un intervalle de ±5 points équivaut à un intervalle de ±2ETM sur les données américaines. Théoriquement, 95 % des enfants devraient avoir des performances entre les deux passations comprises à l'intérieur de cet intervalle. Avec pertinence, Ryan et al. relèvent que « *the FSIQ is less stable than one might infer from examination of the large stability coefficient (i.e., .88) and the small mean practice effect (i.e., 1.63 points)* » (2010, p. 72). Néanmoins, ils concluent que « *only the FSIQ has sufficient stability for interpretation in individual cases* » (2010, p. 71).

Dans toutes ces études dans lesquelles l'intervalle de temps entre les deux passations est inférieur à une année, on remarque une augmentation des notes composites due principalement à un effet d'apprentissage (Ryan et al., 2010; Wechsler, 2005b; Williams et al., 2003). L'augmentation des performances entre deux passations est plus marquée pour l'IRP et l'IVT que pour l'IMT et l'ICV (Ryan et al., 2010; Wechsler, 2005b). Analysant le sous-échantillon des 243 enfants issus de l'échantillon d'étalonnage américain à qui le WISC-IV est administré à environ un mois d'intervalle, Flanagan et Kaufman (2009) observent que les gains dus à l'effet d'apprentissage sont plus importants pour les enfants de 6-7 ans, mais qu'ensuite l'importance des gains diminue avec l'âge. Pour le QIT par exemple, les gains moyens sont de +8.3 points pour les 6-7 ans, +5.8 points pour les 8-11 ans et de +4.3 points pour les 12-16 ans. De son côté, Ryan et al. (2010) décrivent que les enfants ayant les meilleures performances à la première passation bénéficient davantage d'une seconde passation du WISC-IV, indépendamment de leur âge ou de leur grade scolaire au moment de la passation initiale.

Tableau 4

Coefficients test-retest, différences de moyennes et d de Cohen pour trois études sur la stabilité à court terme du WISC-IV

	Williams et al. (2003) (32 jours, *N* = 243)				Wechsler (2005b) (27 jours, *N* = 93)				Ryan et al. (2010) (10.88 mois, *N* =43)			
	r_{12}	r_c	Δ*M*	*d*	r_{12}	r_c	Δ*M*	*d*	r_{12}	r_c	Δ*M*	*d*
Subtest												
CUB	.81	.82	1.20	0.41	.72	.81	1.60	0.61	.67	.72	0.56	0.21
SIM	.81	.86	0.60	0.24	.77	.77	0.70	0.21	.63	.66	0.51	0.17
MCH	.81	.83	0.50	0.18	.59	.71	0.80	0.27	.74	.76	0.30	0.11
IDC	.71	.76	0.80	0.29	.57	.64	0.90	0.34	.22	.26	0.17	0.07
COD	.81	.84	1.40	0.48	.75	.72	2.20	0.72	.42	.40	0.24	0.08
VOC	.85	.92	0.30	0.13	.80	.82	0.40	0.17	.81	.84	-0.02	-0.01
SLC	.75	.83	0.40	0.16	.66	.71	0.40	0.14	.49	.52	0.30	0.11
MAT	.77	.85	0.60	0.23	.68	.67	1.00	0.33	.40	.47	-0.53	-0.21
COM	.72	.82	0.20	0.08	.80	.79	0.00	-0.01	.49	.47	-0.23	-0.08
SYM	.68	.80	1.10	0.41	.70	.83	1.80	0.63	.41	.50	0.31	0.31
CIM	.82	.84	1.80	0.60					.46	.46	0.72	0.25
BAR	.78	.79	1.10	0.37					.51	.52	0.89	0.33
INF	.83	.89	0.40	0.16					.66	.74	0.17	0.07
ARI	.75	.79	0.60	0.23					.68	.75	0.39	0.16
RSV	.75	.82	0.80	0.31					.61	.65	0.44	0.18
Composite												
ICV	.89	.93	2.10	0.18	.88	.88	2.10	0.14	.75	.76	0.25	0.02
IRP	.85	.89	5.20	0.39	.80	.83	8.20	0.59	.58	.68	0.42	0.04
IMT	.85	.89	2.60	0.20	.73	.78	3.60	0.24	.73	.75	2.05	0.14
IVT	.79	.86	7.10	0.51	.79	.83	12.00	0.81	.49	.54	2.32	0.20
QIT	.89	.93	5.60	0.46	.89	.91	8.30	0.59	.80	.88	1.63	0.15

Note. r_{12} = coefficient de stabilité non corrigé ; r_c = coefficient de stabilité corrigé ; Δ *M* = différence de moyennes ; *d* = taille de la différence ; CUB = Cubes ; SIM = Similitudes ; MCH = Mémoire des chiffres ; IDC = Identification de concepts ; COD = Code ; VOC = Vocabulaire ; SLC = Séquence lettres-chiffres ; MAT = Matrices ; COM = Compréhension ; SYM = Symboles ; CIM = Complètement d'images, BAR = Barrage ; INF = Information ; ARI = Arithmétique ; RSV = Raisonnement verbal ; ICV = Indice de Compréhension Verbale ; IRP = Indice de Raisonnement Perceptif ; IMT = Indice de Mémoire de Travail ; IVT = Indice de Vitesse de Traitement ; QIT= QI Total.

3.4.4. STABILITÉ À LONG TERME DU WISC-IV

Étant donné les décisions qui peuvent découler des résultats des tests d'intelligence, la nécessité d'explorer la stabilité à long terme de leurs scores est admise, cependant, le risque de mortalité expérimentale et le coût élevé n'encouragent pas à entamer une étude longitudinale. À notre connaissance, peu d'études ont estimé la stabilité à long terme des scores du WISC-IV, voire aucune pour l'adaptation française. Dans la littérature, on peut s'appuyer sur trois études : Lander (2010), Watkins & Smith (2013) et Bartoi et al. (Bartoi et al., 2015).

Les travaux de Lander (2010) sont menés sur un échantillon de 131 enfants américains (75 garçons et 56 filles) présentant des troubles des apprentissages. À la première passation, l'âge des enfants varie de 6 à 13 ans, avec un âge moyen de 8.28 ans (écart type non renseigné). À la seconde passation, l'âge des enfants varie de 8 à 16 ans, avec un âge moyen de 11.17 ans (écart type non renseigné). L'intervalle test-retest moyen est de 2.89 ans (écart type non renseigné). Les performances moyennes des indices varient de 84.94 (QIT) à 90.4 (IVT) lors de la première passation, et de 84.39 (QIT) à 88.83 (IRP) lors de la seconde passation. Étant donné leurs difficultés d'apprentissage, leurs performances sont en moyennes plus faibles que celles des enfants tout-venant de l'échantillon de standardisation. La comparaison de moyennes entre les deux passations ne montre aucune différence statistiquement significative pour les indices, sauf pour l'Indice de Vitesse de Traitement. La moyenne de l'IVT diminue significativement au retest (-2.14 points), cependant, la taille de l'effet est négligeable (d =-.18). Sur un délai test-retest de plus de deux ans et avec une population clinique, on ne relève pas d'effet d'apprentissage dû à l'expérience d'une première passation. Les coefficients de stabilité sont inférieurs à .70 pour les quatre indices (voir Tableau 5, p. 162). Le coefficient de stabilité est de .70 pour le QIT. Les coefficients de stabilité des subtests varient de .28 (Symboles) à .62 (Cubes). Il est à noter qu'aucune correction n'a été appliquée sur les coefficients de corrélation, alors qu'on peut supposer une certaine homogénéité des performances par rapport à l'échantillon de standardisation. La réduction de l'étendue des scores possibles dans cet échantillon peut conduire à une sous-estimation des coefficients de fidélité. Pour l'analyse des performances au niveau individuel, Lander examine le pourcentage d'enfants présentant des performances comprises dans un intervalle construit au moyen de l'erreur type de mesure (ETM). En vertu des propriétés de la distribution normale, 68 % des performances sont théoriquement incluses dans un intervalle de ±1

ETM, 95 % des performances sont théoriquement incluses dans un intervalle de ±2 ETM, et 99 % des performances sont théoriquement incluses dans un intervalle de ±3 ETM. Communément, on utilise dans la littérature l'intervalle de ±2 ETM. Pour l'adaptation américaine, le Tableau 4.3 dans le *WISC-IV Technical and Interpretive Manual* renseigne sur les valeurs des ETM de chaque subtest et indice (Wechsler, 2003, p. 38). Connaissant la valeur de l'ETM de l'indice, on peut définir les bornes d'un intervalle. Par exemple, prenons un enfant qui a la première passation obtient un QIT de 100. Sachant que l'ETM du QIT est de 2.68 points pour l'adaptation américaine, l'intervalle à ±2 ETM donne ±5.36 points. Si à la seconde passation, l'enfant obtient un QIT compris dans l'intervalle [94.64 ; 105.36], alors ses performances sont considérées comme stables. Lander présente les résultats du pourcentage d'enfants dont les performances sont comprises dans les intervalles de ±1 ETM, ±2 ETM et ±3 ETM entre les deux passations (voir Tableau 6, p. 164). Nous pouvons voir que le pourcentage d'enfants présentant des performances incluses dans l'intervalle de ±1 ETM entre les deux passations varie de 52 % (IMT) à 60 % (QIT et ICV). Pour l'intervalle de ±2 ETM, 70 % (IMT), 73 % (IRP et IVT) et 78 % (QIT et ICV) d'enfants présentent des performances variant à l'intérieur de cet intervalle entre le test et le retest. Enfin, le pourcentage d'enfants présentant des performances comprises dans l'intervalle le plus large de ±3 ETM entre les deux passations varie de 85 % (QIT) à 91 % (ICV). Aucun indice n'approche les proportions théoriquement attendues de 68 %, 95 % et 99 % pour les intervalles de ±1 ETM, ±2 ETM et ±3 ETM respectivement. Ces résultats montrent qu'il y a un pourcentage plus important d'individus qui voient leur performance varier au-delà de ce qui est théoriquement attendu à cause de l'erreur de mesure. Lander présente également les résultats d'une évaluation de la stabilité catégorielle pour le QIT. En fonction de l'étendue des QI Totaux dans l'échantillon, cinq catégories sont définies : extrêmement faible (≤ 69), limite (70-79), moyen faible (80-89), dans la moyenne (90-109) et moyen fort (110-119). Dans les 131 enfants de l'échantillon, les résultats montrent que 57 % (soit 75 enfants) sont restés dans la même catégorie descriptive à la première et à la seconde passation. Parmi les 43 % qui changent de catégorie (soit 56 enfants), l'écrasante majorité a soit descendu soit monté d'une catégorie. Moins de 1 % des enfants changent de deux catégories. Pour la catégorie dans la moyenne (QIT entre 90-109), il y a 41 enfants à la première passation. À la seconde passation, 25 enfants (soit 61 %) sont restés dans cette catégorie, 15 enfants (soit 36.6 %) sont descendus dans la catégorie des performances moyennes faibles (80-89), et 1 enfant (soit 2.4 %) est descendu de deux catégories dans les performances limites (70-79). Pour la catégorie moyen faible (QIT entre 80-89), il y a 45 enfants à la première passation. À la seconde

passation, 30 enfants (soit 66.7 %) sont restés dans cette catégorie, 6 enfants (soit 13.3 %) ont monté dans la catégorie dans la moyenne, 8 enfants (soit 17.8 %) sont descendus dans la catégorie des performances limites, et 1 enfant (soit 2.2 %) est descendu de deux catégories dans les performances extrêmement faibles. Pour la catégorie limite (QIT entre 70-79), il y a 39 enfants à la première passation. À la seconde passation, 19 enfants (soit 48.7 %) sont restés dans cette catégorie, 14 enfants (soit 35.9 %) sont montés dans la catégorie moyen faible, 5 enfants (soit 12.8 %) sont descendus dans la catégorie des performances extrêmement faibles, et 1 enfant (soit 2.6 %) est monté de deux catégories dans les performances dans la moyenne. Pour la catégorie extrêmement faible (QIT <69), il y a 5 enfants à la première passation. À la seconde passation, 1 enfant (soit 20 %) est resté dans cette catégorie, 4 enfants (soit 80 %) sont montés dans la catégorie limite.

Dans les travaux de Watkins et Smith (2013), l'échantillon comprend 344 enfants (66 % de garçons) présentant des troubles des apprentissages, des retards mentaux ou des troubles émotionnels. L'âge moyen est de 8.74 ans (écart type = 1.57 an) à la première passation, et de 11.6 ans (écart type = 1.69 an) à la seconde passation. L'intervalle test-retest moyen est de 2.84 ans (écart type = 0.75 an). Les performances moyennes des indices varient de 84.27 (IMT) à 95.55 (IRP) lors de la première passation, et de 88.10 (IMT) à 95.92 (IRP) lors de la seconde passation. Pour comparer les moyennes entre les deux passations, des *t*-tests pour échantillons appariés sont réalisés. La comparaison de moyennes entre les deux passations ne montre aucune différence statistiquement significative pour les indices. Seuls les subtests Cubes, Similitudes et Code diffèrent de manière statistiquement significative entre la première et la seconde passation. Cependant, la taille d'effet associée aux différences de moyenne est négligeable à petite. Ces résultats rejoignent ceux de Lander (2010). Si nous regardons les coefficients de stabilité, des différences sont à relever par rapport à l'étude de Lander (2010). En effet, Watkins et Smith trouvent des coefficients de stabilité corrigés supérieurs à .70 pour tous les indices, sauf l'IVT (voir Tableau 5, p. 162). Le QI Total présente le coefficient de stabilité le plus élevé (r = .84). À l'instar des autres recherches, les coefficients de stabilité des subtests se révèlent moins élevés que ceux des indices auxquels ils contribuent (voir Tableau 5, p. 162). Au niveau individuel, Watkins et Smith observent qu'entre les deux passations, plus de 70 % des enfants présentent des performances comprises entre ±9 points pour le QIT et l'ICV ; et respectivement 61 %, 63 % et 56 % pour l'IRP, l'IMT et l'IVT (voir Tableau 6, p. 164).

La troisième étude est conduite par Bartoi et al. (2015). L'échantillon est constitué de 51 enfants âgés de 8 à 16 ans (69 % de garçons) qui sont référés pour une évaluation psychoéducationnelle. Au terme de la passation initiale, plusieurs diagnostics sont posés (TDA/H, trouble des apprentissages, retard mental, trouble anxieux, etc.). Il est à noter qu'à plus ou moins fort degré, tous les enfants de cette étude présentent des problèmes d'attention. L'intervalle test-retest moyen est de 1.84 an (écart type = 0.50 an). Les performances moyennes des indices varient de 90.10 (IVT) à 98.08 (ICV) lors de la première passation, et de 89.31 (IVT) à 98.10 (ICV) lors de la seconde passation. Aucun *t*-tests n'est rapporté, néanmoins les auteurs déclarent qu'il n'y a pas de changement significatif dans les scores au cours du temps. Les coefficients de stabilité non corrigés sont autour de .80 pour tous les indices, excepté pour l'IMT et l'IVT (voir Tableau 5, p. 162). L'analyse des variations individuelles montre qu'entre les deux passations, 78.4 % des enfants ont des performances comprises dans un intervalle de ±9 points pour le QIT (voir Tableau 6, p. 164). De même 68.6 %, 56.9 %, 54.9 et 54.9 % des enfants ont des différences de performances entre les deux passations inférieures ou égales à 9 points pour l'ICV, l'IRP, l'IMT et l'IVT (voir Tableau 6, p. 164). Ces résultats au niveau intra-individuel sont proches de ceux de Watkins et Smith (2013) et indiquent des variations importantes dans les performances d'un individu d'une passation à l'autre. Pour le QIT, quatre enfants sur cinq voient leurs performances augmenter ou diminuer de plus de 9 points entre deux passations.

Tableau 5

Coefficients test-retest, différences de moyennes et d de Cohen pour trois études sur la stabilité à long terme du WISC-IV

	Lander (2010) (2.89 ans, *N* = 131)			Watkins & Smith (2013) (2.84 ans, *N* = 344)				Bartoi et al. (2015) (1.84 an, *N* = 51)	
	r_{12}	Δ*M*	*d*	r_{12}	r_c	Δ*M*	*d*	r_{12}	Δ*M*
Subtest									
CUB	.62	-0.40	-0.16	.70	.72	-0.50*	-0.18	.78	-0.14
SIM	.48	0.57	0.26	.58	.63	0.40*	0.15	.80	-0.29
MCH	.46	-0.14	-0.05	.60	.65	-0.17	-0.07	.63	0.10
IDC	.44	0.38	0.13	.46	.43	0.58	0.19	.54	0.04
COD	.46	-0.90	-0.40	.52	.50	-0.90*	-0.30	.52	-0.18
VOC	.56	-0.49	-0.22	.69	.73	-0.15	-0.06	.81	-0.16
SLC	.31	0.08	0.01	.48	.51	0.18	0.06	.35	-0.14
MAT	.48	-0.25	-0.10	.63	.64	0.12	0.04	.74	-0.45
COM	.55	0.19	0.23	.48	.53	0.09	0.03	.49	0.50
SYM	.28	0.05	0.02	.54	.51	0.25	0.08	.62	0.04

	Lander (2010) (2.89 ans, *N* = 131)			Watkins & Smith (2013) (2.84 ans, *N* = 344)				Bartoi et al. (2015) (1.84 an, *N* = 51)	
	r_{12}	Δ*M*	*d*	r_{12}	r_c	Δ*M*	*d*	r_{12}	Δ*M*
Composite									
ICV	.65	0.64	0.06	.72	.78	0.55	0.04	.81	0.05
IRP	.62	-0.30	-0.02	.76	.76	0.37	0.03	.79	-1.16
IMT	.54	-0.45	-0.04	.66	.70	-0.17	-0.01	.60	-0.06
IVT	.52	-2.14*	-0.18	.65	.65	-1.90	-0.13	.58	-0.79
QIT	.70	-0.55	-0.06	.82	.84	-0.12	-0.01	.86	-0.60

Note. r_{12} = coefficient de stabilité non corrigé ; r_c = coefficient de stabilité corrigé ; Δ *M* = différence de moyennes ; *d* = taille de la différence ; CUB = Cubes ; SIM = Similitudes ; MCH = Mémoire des chiffres ; IDC = Identification de concepts ; COD = Code ; VOC = Vocabulaire ; SLC = Séquence lettres-chiffres ; MAT = Matrices ; COM = Compréhension ; SYM = Symboles ; ICV = Indice de Compréhension Verbale ; IRP = Indice de Raisonnement Perceptif ; IMT = Indice de Mémoire de Travail ; IVT = Indice de Vitesse de Traitement ; QIT= QI Total.

*$p < .05$.

Dans l'ensemble des études à court et à long terme que nous venons de présenter, nous pouvons relever que le coefficient de corrélation du QI Total est le plus élevé et qu'il exprime généralement une stabilité différentielle suffisante pour des décisions au niveau du groupe (c.-à-d. ≥ .70). Toutefois, si on se tient au critère de .90 pour les décisions au niveau de l'individu, il n'est atteint dans aucune des recherches. En outre, on peut relever le focus sur les coefficients de stabilité (stabilité différentielle) comme indicateur privilégié de la stabilité des scores. En effet, les résultats au niveau intra-individuel sont moins fréquemment rapportés. Pour notre part, il s'agit pourtant du niveau le plus important dans une utilisation clinique d'un test d'intelligence tel que le WISC-IV. Les décisions à l'aide du WISC-IV sont sur des cas individuels, il est nécessaire de mieux documenter sur différents groupes d'individus (cliniques et non cliniques) afin que le clinicien puisse évaluer les limites des résultats pour tel individu. Les résultats des recherches présentées soulignent bien le fait que les informations sur la stabilité absolue ou différentielle ne renseignent pas sur la stabilité au niveau intra-individuel, et inversement. Ces trois types d'évaluation de la stabilité devraient toujours être évalués de concert afin d'éclairer au mieux sur l'utilisation et l'interprétation des scores d'un test.

Tableau 6

Pourcentage de différences individuelles sur les indices du WISC-IV incluses dans un intervalle de points entre les deux passations

	Ryan et al. (2010)	Lander (2010)			Watkins & Smith (2013)	Bartoi et al. (2015)	
	±5 points	±1 ETM	±2 ETM	±3 ETM	±9 points	±5 points	±9 points
ICV	-	61.0	78.0	91.0	71.0	41.2	68.6
IRP	-	57.0	73.0	88.0	61.0	27.2	56.9
IMT	-	52.0	70.0	86.0	63.0	25.5	54.9
IVT	-	57.0	73.0	88.0	56.0	29.4	54.9
QIT	58.1	61.0	78.0	85.0	75.0	51.0	78.4

Note. ETM = erreur type de mesure ; ICV = Indice de Compréhension Verbale ; IRP = Indice de Raisonnement Perceptif ; IMT = Indice de Mémoire de Travail ; IVT = Indice de Vitesse de Traitement ; QIT= QI Total.

4. PROBLÉMATIQUE

Dans ce chapitre, nous présentons les objectifs et les hypothèses de recherche du présent travail. Deux études ont été menées à partir des données récoltées. La première porte sur les items du WISC-IV et plus précisément l'éventualité d'un biais d'item. Il s'agit d'une étude qui est suscitée par un intérêt personnel à la fois pour la question de l'équité dans l'évaluation psychologique et pour l'application des modèles de réponse à l'item sur les donnés du WISC-IV. La seconde étude porte sur le sujet principal de la thèse : la stabilité des scores du WISC-IV. Cette seconde étude répond à une lacune de résultats, notamment avec des données de l'adaptation en français du WISC-IV. Pourtant, les psychologues pratiquant l'évaluation psychologique sont en demande sur les prédictions qu'ils peuvent se permettre avec des instruments tels que les Échelles de Wechsler, surtout avec la population enfant.

4.1. FONCTIONNEMENT DIFFÉRENTIEL DES ITEMS DU WISC-IV

Selon la demande d'un bilan psychologique, le psychologue peut recourir à des tests cognitifs – tels que les populaires Échelles de Wechsler – pour aider aux diagnostics et orienter les interventions. Étant donné les enjeux inhérents à de telles décisions, il est important de s'assurer que les différences observées sur les scores d'un test cognitif traduisent bel et bien des différences interindividuelles dans le fonctionnement intellectuel des individus, et non des différences liées à d'autres caractéristiques (p. ex., ethnie, âge, sexe, milieu socio-économique). Dans l'évaluation de la validité de l'interprétation des scores, la détection des biais répond à la préoccupation d'équité dans l'évaluation psychologique. Pour l'adaptation française du WISC-IV (Wechsler, 2005b), les enfants français et belges sont le public visé. Le repérage d'éventuels items biaisés est réalisé avec une version préexpérimentation du WISC-IV sur un échantillon de 220 enfants issus de régions françaises et un échantillon de 125 enfants belges. À l'issue de la comparaison entre ces deux échantillons, seuls les items possédant les qualités métriques requises sont sélectionnés pour la version définitive. À noter qu'aucune analyse n'a donc été réalisée pour vérifier l'équité pour une population suisse francophone. Il s'agit d'un manquement que nous ne pouvons

hélas pas remédier. N'ayant pas accès aux données de l'échantillon de standardisation du WISC-IV, nous ne pouvons pas réaliser des analyses de comparaisons entre notre échantillon d'enfants suisses francophones et les enfants français de l'échantillon de standardisation.

L'objectif de notre première étude est d'évaluer si les items des subtests du WISC-IV se comportent de la même manière pour tous les individus qui ont la même habileté sur le trait latent évalué. L'échantillon étant constitué d'enfants suisses francophones, la détection d'un fonctionnement différentiel des items s'est portée sur les variables âge, sexe et statut socio-économique. L'impact du statut socio-économique combiné des deux parents sur les performances intellectuelles de l'enfant est bien connu, en revanche, l'influence de chaque parent est rarement examinée de manière distincte. Nos analyses examineront la part de variance expliquée par la profession de chaque parent. Par ailleurs, des recherches montrent que différents niveaux d'habileté s'observent selon le sexe, notamment dans les tâches visuospatiales et verbales. Généralement, les garçons obtiennent des scores plus élevés en rotations mentales (Voyer, Voyer, & Bryden, 1995), tandis que les filles montrent de meilleures performances sur les tâches verbales (Hyde & Linn, 1988). Dans la première étude sur le fonctionnement différentiel des items du WISC-IV, nous examinerons ce qu'il en est des items biaisés pour un échantillon d'enfants suisses francophones. Plus précisément, nous déterminerons les proportions de variances expliquées par l'âge, le sexe et le statut socio-économique des parents, ainsi que le fonctionnement différentiel des items.

Compte tenu des analyses menées par les concepteurs du WISC-IV, nous posons comme hypothèse sur le fonctionnement différentiel du WISC-IV qu'il n'y a pas d'items biaisés pour l'âge, le sexe ou le statut socio-économique.

4.2. STABILITÉ À LONG TERME DU WISC-IV

L'utilité d'un test réside dans sa propension à mettre en lumière des différences interindividuelles sur la propriété mentale évaluée. Les différences entre les individus ne peuvent être interprétées comme des différences sur la propriété mentale que si le test présent des qualités psychométriques (homogénéité, sensibilité, fidélité, validité). En tant que reflet de la précision, de la consistance et de la stabilité des scores, la fidélité est l'une des propriétés importantes à évaluer dans un test. En effet, si les résultats d'un

test traduisent le fonctionnement psychologique de l'individu, ceux-ci doivent être suffisamment fidèles et reproductibles dans le temps afin d'arriver à des interprétations comparables d'une passation à l'autre. La stabilité des scores sera évaluée aussi bien au niveau interindividuel qu'intra-individuelle. Le niveau intra-individuel est souvent moins présenté dans les études longitudinales sur la stabilité des scores d'un test. Pourtant, il est particulièrement pertinent avec la pratique clinique des tests dans une évaluation psychologique. Notre seconde étude explore donc la question de la stabilité des scores du WISC-IV sous plusieurs angles : (1) la stabilité absolue, (2) la stabilité différentielle, (3) la stabilité intra-individuelle absolue, (4) la stabilité catégorielle et (5) la stabilité des forces et faiblesses.

Pour la stabilité absolue, il s'agit de tester la différence de moyennes entre la première et la seconde passation pour l'échantillon total. Nous déterminons ainsi, au niveau du groupe et par comparaisons de moyennes (*t*-tests pour échantillons appariés), si les moyennes de la première et la seconde passation sont équivalentes. Les études longitudinales sur les tests cognitifs montrent que les tâches impliquant des habiletés de compréhension-connaissance présentent une meilleure stabilité des scores que les tâches impliquant du raisonnement fluide et de la résolution de problème (Calamia et al., 2012; Dikmen et al., 1999; Schwartzman et al., 1987). Sur des intervalles à court terme (3 à 6 mois), les gains à la seconde passation tendent à être plus importants pour les épreuves simples de vitesse de traitement que pour les épreuves verbales de vocabulaire ou de culture générale (Calamia et al., 2012; Estevis et al., 2012). Dans les études sur le WISC-IV avec un intervalle inférieur à une année (court terme), des effets d'apprentissage s'observent pour les échantillons d'enfants tout-venant. Les effets d'apprentissage sont plus prononcés pour l'IRP et l'IVT que pour l'IMT et l'ICV (Ryan et al., 2010; Wechsler, 2005b). En outre, les gains dus à l'effet d'apprentissage sont plus importants pour les enfants âgés de 6-7 ans à la passation initiale et ensuite les gains diminuent avec l'âge à la première passation (Flanagan & Kaufman, 2009). Avec des délais test-retest courts, les enfants tout-venant avec les meilleures performances bénéficient davantage d'une seconde passation du WISC-IV que les enfants les moins performants (Ryan et al., 2010). Dans leurs travaux sur le WISC-III administré à deux reprises à un échantillon clinique, Canivez et Watkins (1999, 2001) constatent qu'on n'observe plus, ou de très faible effet d'apprentissage lorsque les deux passations sont séparées par un intervalle supérieur à une année. Avec le WISC-IV et des intervalles supérieurs à 1 an (long terme), les études sur des échantillons cliniques ne montrent pas de différences de moyennes significatives dues à un effet

d'apprentissage pour les indices entre les deux passations (Lander, 2010; Watkins & Smith, 2013). Lander (2010) relève une différence de moyennes pour l'IVT, qui diminue significativement à la seconde passation. Cependant, la taille d'effet est négligeable. À notre connaissance, il n'y aucune étude sur la stabilité absolue à long terme d'un groupe d'enfants tout-venant pour l'adaptation en français du WISC-IV.

Compte tenu des résultats des différentes études, nous posons comme hypothèses sur la stabilité absolue à long terme des scores du WISC-IV qu'on peut relever des effets d'apprentissage au-delà d'un an, mais que ceux-ci sont négligeables à petits. Les effets d'apprentissage s'observent de façon plus prononcée sur les scores des indices IRP et IVT que sur ceux de l'ICV et de l'IMT. Sur la lecture CHC, Gf, Gv et Gs présentent plus d'effets d'apprentissage que Gc et Gwm.

Pour la stabilité différentielle, il s'agit de calculer le coefficient de corrélation test-retest. Nous déterminons ainsi si l'ordre des individus est similaire entre les deux passations. La stabilité différentielle est souvent référée pour l'évaluation de la stabilité des scores d'un test. Sur les études de stabilité à court terme des scores du WISC-IV dans un échantillon non clinique, les coefficients test-retest varient autour de .80 pour les indices, voire .90 pour le QI Total (Ryan et al., 2010; Wechsler, 2003, 2005b). Sur des études de stabilité à long terme du WISC-IV, les échantillons sont cliniques et américains. Les coefficients de stabilité varient autour de .60 - .70. L'IVT et l'IMT présentent des coefficients non seulement moins élevés que l'ICV et l'IRP, mais également inférieurs à .70 (Bartoi et al., 2015; Lander, 2010; Watkins & Smith, 2013). Quant au QI Total, il présente le coefficient de stabilité le plus élevé de tous les indices (Bartoi et al., 2015; Lander, 2010; Watkins & Smith, 2013). Les études montrent que les coefficients de stabilité des subtests sont moins élevés que ceux des notes composites auxquelles ils contribuent (Bartoi et al., 2015; Lander, 2010; Ryan et al., 2010; Watkins & Smith, 2013; Wechsler, 2003, 2005b).

Compte tenu des résultats des différentes études, nous posons comme hypothèses sur la stabilité différentielle à long terme des scores du WISC-IV que les coefficients de stabilités des indices IMT, IVT et ICC sont plus faibles que ceux de l'ICV, l'IRP et l'IAG. Le coefficient de stabilité du QIT est plus élevé que ceux des quatre indices qui le composent et doit s'élever autour de .80. Sur la lecture CHC, Gc, Gf, Gv présentent un coefficient de fidélité plus élevé que Gwm et Gs.

Pour la stabilité intra-individuelle absolue, il s'agit d'évaluer les différences des performances individuelles entre la première et la seconde passation. Nous déterminons ainsi le pourcentage d'enfants qui présentent des performances stables

entre les deux passations. Nous entendons par stables des performances pour les deux passations comprises dans un intervalle défini. Nous utilisons l'intervalle de deux erreurs types de mesure (±2 ETM), souvent choisi dans la littérature. Pour les scores WISC-IV, l'évaluation de la stabilité intra-individuelle à long terme n'est réalisée que sur des échantillons cliniques d'enfants américains. Le QIT présente le pourcentage le plus élevé (autour de 75 %) d'enfants qui gardent des performances stables (soit à ±2 ETM soit à ± 9 points) entre les deux passations (Bartoi et al., 2015; Lander, 2010; Watkins & Smith, 2013).

Compte tenu des résultats des différentes études, nous posons comme hypothèses sur la stabilité intra-individuelle absolue à long terme des scores du WISC-IV que le QIT doit présenter des performances stables (à ±2 ETM) pour plus de 70 % des enfants. Plus sensibles à un effet d'apprentissage, les indices IRP et IVT doivent présenter les moins de performances stables (à ±2 ETM) entre les deux passations.

Généralement, le psychologue communique les résultats au WISC-IV en présentant les forces et les faiblesses du sujet non seulement par rapport à son groupe de référence (comparaison normative), mais également par rapport à lui-même (comparaison ipsative). En effet, les domaines cognitifs identifiés comme force et/ou faiblesse de l'enfant par rapport aux autres enfants de son âge (forces et faiblesses normatives) et par rapport à lui-même (forces et faiblesses personnelles) vont permettre au psychologue d'élaborer des pistes d'intervention. Le niveau ipsatif montre que malgré des performances faibles par rapport aux autres enfants de son groupe d'âge, le sujet peut néanmoins posséder des atouts personnels sur lesquels on peut s'appuyer pour la prise en charge. Une lecture généralement utilisée dans la clinique est celle des catégories de description qualitative qui aide le psychologue à donner du sens à l'interprétation d'un score numérique qui peut parfois être peu parlant pour situer la performance. Étant donné l'importance pour la clinique, nous explorons donc également la stabilité des forces et des faiblesses personnelles. La stabilité des forces et des faiblesses normatives est renseignée dans les résultats de la stabilité catégorielle.

Pour la stabilité catégorielle, il s'agit de voir quel pourcentage d'enfants reste dans la même catégorie descriptive d'une passation à l'autre. Trois classifications sont comparées. La classification traditionnelle en sept catégories : très faible (≤ 69), limite (70-79), moyen faible (80-89), moyen (90-109), moyen fort (110-119), supérieur (120-129), et très supérieur (≥ 130). La classification des performances en trois catégories : faible (≤ 84), dans la moyenne (85-115), et élevé (≥ 116). Cette classification correspond à la lecture normative des faiblesse, moyenne et force de l'enfant par rapport à son

groupe d'âge. Enfin, la classification en cinq catégories : extrémité inférieure (≤ 69), moyen faible (70-84), dans la moyenne (85-115), moyen fort (116-130), et extrémité supérieure (≥ 131). Lander (2010) présente les résultats d'une évaluation de la stabilité catégorielle pour le QI Total. Cinq catégories sont définies : extrêmement faible (≤ 69), limite (70-79), moyen faible (80-89), dans la moyenne (90-109) et moyen fort (110-119). Dans un échantillon de 131 enfants ayant des difficultés d'apprentissage, les résultats montrent que 57 % des enfants sont restés dans la même catégorie descriptive à la première et à la seconde passation. Parmi les 56 enfants (soit 43 %) qui changent de catégorie, l'écrasante majorité a soit descendu soit monté d'une catégorie.

Sur la base d'un nombre limité d'études, nous posons comme hypothèse sur la stabilité catégorielle à long terme des scores du WISC-IV que les enfants qui changent de catégorie descriptive, vont monter ou descendre d'une seule catégorie à la seconde passation.

Pour la stabilité des forces et des faiblesses personnelles, il s'agit d'abord de déterminer si un indice est une force, une moyenne ou une faiblesse personnelle. Pour cela, on calcule à partir de la moyenne des quatre indices (ICV, IRP, IMT et IVT) un niveau de performance moyen pour chaque enfant (indice moyen). Ensuite, chaque indice est comparé à l'indice moyen de l'enfant pour voir s'il est éloigné au-delà d'un certain seuil de l'indice moyen. Un indice qui dévie vers une performance inférieure à l'indice moyen est une faiblesse personnelle. S'il dévie vers une performance supérieure à l'indice moyen, on parle de force personnelle. Enfin, s'il ne dévie pas au-delà du seuil, il s'agit d'une moyenne personnelle.

Nous ne pouvons pas nous appuyer sur les résultats de précédentes études, à notre connaissance aucune n'a porté sur cette question. Nous posons une hypothèse de stabilité des forces et des faiblesses personnelles à long terme des scores du WISC-IV. En effet, en l'absence d'intervention, les forces et les faiblesses identifiées au sein d'un profil du WISC-IV sont supposées être relativement stables dans le temps.

MÉTHODE

5. RÉCOLTE DE DONNÉES

Deux études ont été menées à partir des données récoltées. La première étude porte sur le fonctionnement différentiel des items du WISC-IV (Étude 1). L'évaluation d'éventuels biais dans les items du WISC-IV est préalable à toutes analyses sur ses scores. La seconde étude porte sur la stabilité des scores standards et CHC du WISC-IV (Étude 2).

Ce chapitre commencera par la description de l'échantillon constitué pour les deux études (voir section 5.1). Il s'en suivra une présentation de la procédure de récolte des données dans les écoles genevoises et de l'instrument au cœur de ce travail (voir sections 5.2 et 5.3).

5.1. ÉCHANTILLON

L'échantillon du présent travail est issu de deux recherches FNS : *Analysis of the French WISC-IV structure according to the Cattell-Horn-Carroll narrow ability classification*[21] et *Long-term stability of the WISC-IV: standard and CHC composite scores*[22]. Les protocoles sont recueillis de fin janvier 2008 à juin 2014 dans 24 écoles publiques du canton de Genève. La première passation des enfants (phase Test) se déroule de fin janvier 2008 à mai 2013. Parallèlement, la seconde passation (phase Retest) débute en octobre 2010 pour se terminer en juin 2014.

Dans le canton de Genève, les enfants passent normalement leur scolarité dans le même établissement durant tout le cycle élémentaire et primaire (4 – 12 ans). Pour faciliter la localisation des enfants à la phase Retest et alléger les démarches administratives d'autorisation de recherche dans les écoles, nous restreignons l'étendue des âges dans notre échantillon à 7 – 12 ans.

Quant aux autres critères d'inclusion, ce sont la langue et le parcours scolaire. Étant donné l'importante teneur en verbal du test et l'évaluation d'un indice de compréhension verbale, l'adaptation française du WISC-IV s'adresse aux enfants qui parlent couramment français. Pour éliminer d'éventuel biais lié à un redoublement ou un saut de classe (p. ex., des apprentissages scolaires en retard ou plus en avance pour

[21] Requête no. 118248, requérant principal : Thierry Lecerf, co-requérants : Nicolas Favez et Jérôme Rossier.

[22] Requête no. 135406, requérant principal : Thierry Lecerf, co-requérants : Nicolas Favez et Jérôme Rossier.

leur âge), nous ne recrutons que des enfants dans la bonne année scolaire pour leur âge. Ainsi sont inclus uniquement des enfants francophones qui n'ont ni doublé ni sauté une ou plusieurs classes.

L'accès aux enfants dans les écoles n'est pas aisé. Dans une démarche longue en attente de réponse, nous devons obtenir successivement les accords du département de l'instruction publique (DIP), du directeur de l'école, de l'enseignant et enfin des parents. Chaque semestre, la DIP renouvelle une autorisation de contacter les directeurs ; nous demandons plus d'une trentaine d'écoles publiques durant la phase de récolte de données. Au final, il y a 24 écoles dans lesquelles le directeur et au moins un enseignant acceptent de participer à la recherche. Nous n'avons pas d'information sur le nombre de parents qui sont informés de notre recherche. La majorité des enseignants nous retournent uniquement les talons-réponse positifs. La difficulté de recruter un large échantillon et d'accéder aux enfants rendent inapplicables un contrôle des variables sexe des enfants ou milieu sociodémographique. Néanmoins, la sélection d'écoles dans différents quartiers du canton de Genève nous permet d'obtenir un échantillon relativement représentatif de la population d'enfants genevois.

5.1.1. ÉCHANTILLON ÉTUDE 1

L'échantillon de l'étude 1 se compose de 483 enfants « tout-venant » (c.-à-d. non consultants) de 7 à 12 ans. Les enfants se répartissent en 230 garçons et en 253 filles, âgés en moyenne de 9 ans et 0 mois (écart type de 1 an et 3 mois) lors de la passation (voir Tableau 7, p. 175). L'enfant le plus jeune a 7 ans et 0 mois. L'enfant le plus âgé a 12 ans et 6 mois.

Les enfants participants à l'étude forment un échantillon relativement représentatif de la population d'enfants genevois au niveau de la variable sexe. Selon le département de l'instruction publique (DIP) du Canton de Genève[23], il y a 50 % de garçons et 50 % de filles dans les écoles publiques genevoises contre 48 % de garçons et 52 % de filles dans notre échantillon. Quant aux répartitions des catégories socioprofessionnelles des parents, il y a dans notre échantillon 32 % de parents appartenant à la catégorie « Cadres supérieurs et dirigeants », 45 % appartenant à la catégorie « Petits indépendants, employés, et cadres intermédiaires » et, enfin, 22 %

[23]Les statistiques sont sur la base de la situation de l'année scolaire 2014-2015 (https://www.ge.ch/recherche-education/statistiques/annuaire.asp).

appartenant à la catégorie « Ouvriers, divers et sans indication ». Selon les statistiques du DIP, ces pourcentages sont respectivement de 20 %, 45 % et 35 %. L'échantillon comporte donc une surreprésentation d'enfants dont les parents appartiennent à la catégorie « Cadres supérieurs et dirigeants » et une sous-représentation pour la catégorie « Ouvriers, divers et sans indication ».

Tableau 7

Effectifs d'enfants selon l'âge et le sexe (N = 483)

	Âge						Âge moyen	Total
	7	8	9	10	11	12		
Garçons	26	66	64	34	33	7	9.01 (1.32)	230
Filles	24	70	77	38	35	9	9.07 (1.30)	253
Total	50	136	141	72	68	16	9.04 (1.31)	483

5.1.2. ÉCHANTILLON ÉTUDE 2

L'échantillon de l'étude 2 comprend 277 enfants non consultants âgés de 7 à 12 ans. Il s'agit d'un sous-échantillon des enfants présentés dans l'étude 1 sur le fonctionnement différentiel des items. Les enfants se répartissent en 132 garçons et en 145 filles, âgés en moyenne de 8 ans et 10 mois (écart type de 9 mois) à la première passation et de 10 ans et 7 mois (écart type de 1 an et 1 mois) à la seconde passation (voir Tableau 8, p. 176).

Tous les enfants dont nous avons les données du test et du retest sont inclus dans l'échantillon (aucune mortalité expérimentale). Au sein de l'échantillon Test de 483 enfants, 206 enfants ne sont pas revus en seconde passation. La raison est la clôture de la récolte des données ; l'échantillon de 250 enfants exigé par les experts du FNS est dépassé et les trois ans impartis pour la phase Retest se sont écoulés. Parmi les 206 enfants, la plupart n'ont simplement pas été recontactés. Parmi les enfants que nous essayons de recontacter, un petit nombre n'est pas joignable à la suite d'un déménagement (dans un autre canton ou à l'étranger) et six sont perdus suite au refus des parents de laisser participer leur enfant une seconde fois. Au final, il y a un très faible taux de refus pour une seconde passation auprès des parents.

À nouveau, la proportion de garçons et de filles est proche de celle de la population des enfants genevois (48 % de garçons dans notre échantillon contre 50 % dans la population d'élèves genevois). Quant à la répartition selon les catégories

socioprofessionnelles des parents, cet échantillon présente une surreprésentation de la catégorie « Cadres supérieurs et dirigeants » (32 % contre 20 %), une bonne représentation de la catégorie « Petits indépendants, employés et cadres intermédiaires » (45 % contre 45 %) et une sous-représentation de la catégorie « Ouvriers, divers et sans indication » (23 % contre 35 %).

Tableau 8

Effectifs des enfants selon l'âge et le sexe au Test et au Retest (N = 277)

	Test						Retest						Total
	7	8	9	10	11	Âge moyen (ET)	8	9	10	11	12	Âge moyen (ET)	
Garçons	22	56	43	10	1	8.79 (0.86)	17	20	33	50	12	10.58 (1.17)	132
Filles	16	55	64	10	0	8.93 (0.81)	15	16	42	60	12	10.69 (1.06)	145
Total	38	111	107	20	1	8.87 (0.82)	32	36	75	110	24	10.64 (1.11)	277

5.2. PROCÉDURE

Chaque enfant est vu en passation individuelle lors de 2 séances d'environ 40 minutes durant les heures scolaires. Les locaux mis à disposition se trouvent dans l'enceinte de l'école de l'enfant. Les dix subtests principaux du WISC-IV (Cubes, Similitudes, Mémoire des chiffres, Identification de concepts, Code, Vocabulaire, Séquence Lettres-Chiffres, Matrices, Compréhension et Symboles) sont administrés, ainsi que le subtest optionnel Complètement d'images. Comme ils ne contribuent pas au calcul des scores étudiés, les autres subtests optionnels (Information, Raisonnement verbal, Arithmétique et Barrage) ne sont pas administrés.

Un questionnaire est transmis aux parents et permet de recueillir diverses informations sociodémographiques (voir Annexe B). Les professions des parents sont classées en 10 catégories, allant de la catégorie 1 « Directeurs, cadres de direction et gérants » à la catégorie 10 « Sans emploi, au foyer » (voir Annexe C).

5.2.1. CONSIDÉRATIONS ÉTHIQUES

L'autorisation de conduire cette recherche et de contacter les écoles a été donnée par la sous-commission « Recherche dans les écoles » et la commission d'éthique de la Faculté de psychologie et des sciences de l'éducation de l'Université de Genève (FPSE), par le Département de l'Instruction Publique du Canton de Genève (DIP)

et par les représentants légaux des enfants. Une feuille de consentement éclairée est transmise aux parents par l'intermédiaire de l'enseignant-e de leur enfant. Nous insistons auprès des enseignants-es sur le fait qu'un retour d'information n'est pas possible sur le plan individuel et que nous assurons la confidentialité des données de chaque enfant. En effet, les données sont conservées par nos soins et sont traitées anonymement. Étant donné qu'il s'agit d'une étude longitudinale, nous conservons néanmoins les données personnelles des enfants (séparément de la base de données). La fiche de correspondance entre le numéro d'identification et l'identité de l'enfant est protégée et uniquement consultable par les expérimentateurs de la recherche. Comme requis par la DIP, les passations se déroulent toujours en présence de deux expérimentateurs-trices. Avant de commencer la passation, l'enfant est informé du déroulement et du genre d'activités à réaliser. Il peut décider à tout moment de se retirer de la recherche.

5.2.2. PROCÉDURE ÉTUDE 1

La procédure de l'étude 1 suit ce qui est décrit plus haut. Un changement de test est introduit au cours de la recherche *Analysis of the French WISC-IV structure according to the Cattell-Horn-Carroll narrow ability classification.* Il est décidé de remplacer le subtest Triangle de la Batterie pour l'Examen Psychologique de l'Enfant (KABC ; Kaufman & Kaufman, 1993) par le subtest Complètement d'images du WISC-IV dont le score s'additionne à celui du subtest Cubes dans le calcul du facteur Gv. Ainsi, Complètement d'images est administré à 464 enfants. Les 19 enfants à qui ils manquent Complètement d'images avaient passé l'épreuve des Triangles du KABC.

5.2.3. PROCÉDURE ÉTUDE 2

Pour constituer l'échantillon de l'étude 2, nous reprenons contact avec les écoles ayant participé à la phase Test. Les enfants dont les parents acceptent de donner leur accord pour une seconde passation sont revus selon le même setting de passation que décrit plus haut.

Dans notre recherche, nous tenons un délai d'au moins une année entre les deux passations. Hormis ce seuil minimal, la planification des passations Retest ne peut pas être contrôlée. Rappelons la variabilité imprévisible dans temps de réponse des

directeurs et des enseignants pour qui notre recherche n'est pas une priorité dans leur agenda respectif. Comme nous voyons les enfants dans les heures d'école, nous devons privilégier la coopération avec l'école, et plus particulièrement l'enseignant, ce qui ne nous permet pas d'imposer des dates. Dans les limites du temps imparti pour la récolte de protocoles suffisants, nous attendons les délais test-retest le plus long possible. Ainsi, les intervalles test-retest dans notre étude varient de 1 an et 0 mois à 3 ans et 3 mois (moyenne test-retest = 1 an et 9 mois ; écart type = 6 mois). Dans l'Annexe D sont décrits les fréquences, les pourcentages et les pourcentages cumulés des intervalles test-retest.

Étant donné que le subtest optionnel Complètement d'images n'a pas été administré durant les premières passations, les scores impliquant ce subtest présentent 29 données manquantes. Ainsi, les analyses de Complètement d'images et de Gv portent sur un échantillon de 248 enfants

5.3. INSTRUMENT

L'instrument clinique de notre recherche est la 4^{e} édition de l'Échelle d'Intelligence pour Enfants et Adolescents (WISC-IV). La valeur clinique des échelles de Wechsler ainsi que la passation individuelle, ludique et conviviale font la popularité des échelles de Wechsler dans les services de psychologie. Le WISC-IV a été standardisé sur un échantillon de 1'103 enfants âgés de 6 à 16 ans 11 mois considéré comme représentatif de la population française (recensement général de la population de 1999).

Pour rappel, la batterie se compose de 10 subtests obligatoires et de 5 subtests optionnels. En plus de fournir un QI Total (QIT) comme mesure de l'efficience cognitive générale de l'enfant, le WISC-IV mesure également un Indice de Compréhension Verbale (ICV), un Indice de Raisonnement Perceptif (IRP), un Indice de Mémoire de Travail (IMT) et un Indice de Vitesse de Traitement (IVT). Seuls les scores totaux des subtests obligatoires entrent dans le calcul des notes composites des différents indices et du QI Total. Dans les cas où les résultats sur un subtest principal seraient invalides, la substitution par un subtest supplémentaire est autorisée (voir les règles de substitution dans le *Manuel d'administration et de cotation*, Wechsler, 2005a). Aucune substitution n'a été effectuée sur nos données.

5.3.1. DÉMARCHE DE COTATION DES SCORES

La passation des dix subtests principaux permet de calculer le QI Total et les indices standards (inclus IAG et ICC). Toutes nos analyses sont réalisées sur des notes standardisées. Elles permettent d'interpréter les performances de l'enfant en les situant par rapport aux performances d'autres enfants du même groupe d'âge.

(1) Avant de pouvoir interpréter les scores, on commence par la cotation du protocole WISC-IV. Les points obtenus aux différents items de chaque subtest sont additionnés pour obtenir des scores bruts totaux pour chaque subtest.

(2) À l'aide de la table de conversion correspondant à l'âge chronologique de l'enfant (c.-à-d. son âge exact en année, mois et jours), les notes brutes totales de chaque subtest sont converties en notes standards (variant de 1 à 19, moyenne de 10 et d'écart type de 3). Une note standard de 10 à un subtest signifie que les performances de l'enfant sur ce subtest se trouvent dans la moyenne de son groupe d'âge ; il y a donc 50 % des enfants du même groupe d'âge qui ont des performances supérieures à lui, et inversement. On parle alors de performance « moyenne normative » (MoN). Si un enfant obtient une note standard de 7, ses performances le situent à un écart type en dessous de la moyenne. Il y a donc 84 % des enfants du même groupe d'âge qui ont des performances supérieures à lui, et 16 % qui ont des performances inférieures à lui. On parle alors de « faiblesse normative » (FaN). Si un enfant obtient une note standard de 15, ses performances le situent à un écart type en dessus de la moyenne. Il y a donc 16 % des enfants du même groupe d'âge qui ont des performances supérieures à lui, et 84 % qui ont des performances inférieures à lui. On parle alors de « force normative » (FoN).

(3) La somme des notes standards des subtests évaluant un même indice donne une note intermédiaire à convertir à l'aide des tables de conversion des sommes des notes standards en indice (Wechsler, 2005a). Cette conversion transforme la note intermédiaire de chaque indice et du QI Total en une note composite de type *QI* (c.-à-d. moyenne = 100 et d'écart type = 15). La distribution des QI est normalisée de sorte que nous retrouvons 68 % de la population générale entre +/- 1 écart type par rapport à la moyenne de 100 (voir Annexe E). Ainsi, des QI reflétant des performances moyennes (entre 85 et 115) s'observent théoriquement chez 68 % de la population.

5.3.1.1. Calcul indices standards

Selon ses concepteurs, le WISC-IV permet de calculer un QIT, quatre indices (IVT, IRP, IMT et IVT) et deux indices globaux (IAG et ICC). L'Indice de Compréhension Verbale s'obtient avec les notes des subtests Similitudes, Vocabulaire et Compréhension. L'Indice de Raisonnement Perceptif s'obtient avec les notes des subtests Cubes, Identification de concepts et Matrices. L'Indice de Mémoire de Travail s'obtient avec les notes des subtests Mémoire des chiffres et Séquence Lettres-Chiffres. L'Indice de Vitesse de Traitement s'obtient avec les notes des subtests Code et Symboles. Le QI Total s'obtient avec les notes des quatre indices ou avec les notes des dix subtests. L'Indice d'Aptitude Générale s'obtient avec les notes des trois subtests de l'ICV et de l'IRP, tandis que l'Indice de Compétence Cognitive s'obtient avec les notes des subtests de l'IMT et l'IVT. Comme les études sur l'IAG et l'ICC sont apparues postérieurement à la publication de la batterie, ces deux indices ne sont pas intégrés dans les manuels du WISC-IV. Des normes francophones pour ces indices ont été développées et sont disponibles dans les ouvrages de Grégoire (2009) et de Turon-Lagot (2012, pp. 233–235) ainsi que dans l'article de Lecerf et al. (Lecerf et al., 2011).

5.3.1.2. Calcul indices CHC

Plusieurs travaux ont démontré l'adéquation du modèle CHC avec les données du WISC-IV (p. ex., Keith et al., 2006; Lecerf, Rossier, et al., 2010). Sur des données suisses-romandes, les résultats du précédent FNS *« Analysis of the French WISC-IV structure according to Cattell-Horn-Carroll (CHC) narrow ability classification »* a en particulier mis en évidence 5 facteurs CHC (Gc, Gf, Gwm, Gs et Gv). L'Intelligence cristallisée s'obtient avec les notes des subtests Similitudes et Compréhension. L'Intelligence fluide s'obtient avec les notes des subtest Matrices et Identification de concepts. La Mémoire de travail s'obtient avec les notes des subtests Mémoire des chiffres et Séquence Lettres-Chiffres. La Vitesse de traitement s'obtient avec les notes des subtests Code et Symboles. Le Traitement visuel s'obtient avec les notes des subtests Cubes et Complètement d'images.

Pour la conversion en notes QI des facteurs CHC, nous nous sommes référée aux tables de conversion fournies par Lecerf et al. (Lecerf et al., 2012).

5.3.2. ÉPREUVES

Nous allons détailler les onze tâches cognitives qui ont été proposées aux enfants de notre recherche. L'ordre de présentation suit l'ordre standard d'administration. N'ayant pas été administrés, les quatre subtests optionnels Barrage, Information, Arithmétique et Raisonnement verbal ne sont pas décrits. Le facteur CHC supposé être évalué par le subtest sera également indiqué.

Pour chaque subtest, des règles de départ, de retour et d'arrêt sont précisées. Ces règles ont été établies pour réduire la durée de passation et prévenir un effet de lassitude (Wechsler, 2005a). La règle de départ stipule à quel item commencer selon l'âge du sujet. Les items de départ ont un faible degré de difficulté pour les individus de la tranche d'âge correspondant. Étant donné que les items sont ordonnés en difficulté croissante, la réussite aux deux items de départ implique la réussite aux items précédents plus faciles. Ces derniers ne sont pas administrés, mais leurs points sont accordés. Si le sujet n'a pas obtenu une note parfaite (c.-à-d. maximum de points) à l'un des deux premiers items administrés, la règle de retour s'applique et stipule de revenir sur les items précédents en ordre inverse jusqu'à l'obtention de deux notes parfaites consécutives. Conçue pour éviter la fatigue ou le découragement, la règle d'arrêt stipule l'arrêt du subtest après un certain nombre d'items échoués consécutivement. La difficulté des items étant croissante, la réussite à un item qui survient après le nombre d'échecs consécutifs établi par la règle d'arrêt est supposée relever du hasard.

5.3.2.1. Cubes

Dans le subtest Cubes (CUB), l'enfant doit utiliser des cubes bicolores pour reproduire une configuration en un temps déterminé à partir d'un modèle présenté devant lui ou dans le livret de stimuli (voir Figure 33, p. 182). Ce subtest comporte 14 items de difficulté croissante. Pour l'item 1, l'enfant dispose de deux cubes. Puis pour les items 2 à 10 de quatre cubes, et enfin les quatre derniers items sont à réaliser à l'aide de neuf cubes.

Règle de départ : à l'item 1 pour les 6 – 7 ans et à l'item 3 pour les 8 – 16 ans.

Règle d'arrêt : après 3 notes 0 consécutives.

Cotation : les trois premiers items comportent deux essais possibles et sont cotés 0, 1 ou 2 points. Les items 4 à 8 sont cotés 0 ou 4 points. Puis les items suivants

sont au bénéfice d'une bonification de temps et sont cotés 0, 4, 5, 6 ou 7 points en fonction de la rapidité d'exécution et si la construction est réalisée correctement dans le temps imparti. Le score brut maximum est de 68 points, ou de 50 points sans bonification de temps. La possibilité de calculer un score sans bonification de temps permet de tenir compte de la pression du chronométrage sur les stratégies de résolution de la tâche.

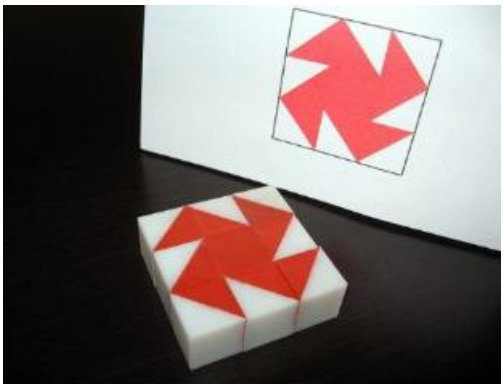


Figure 33. Illustration de l'épreuve Cubes.

Cubes est un subtest obligatoire de l'IRP. Selon le manuel du WISC-IV, il fait principalement appel à la capacité à analyser et à synthétiser des stimuli visuels abstraits ainsi qu'à la coordination visuomotrice (Wechsler, 2005b). Il évalue l'analyse et le raisonnement visuospatial, l'habileté de coordination visuomotrice et la pression du temps. Selon le modèle CHC, ce subtest est un bon indicateur de l'aptitude étendue Traitement visuel Gv et plus précisément, l'aptitude restreinte *Visualisation* VZ.

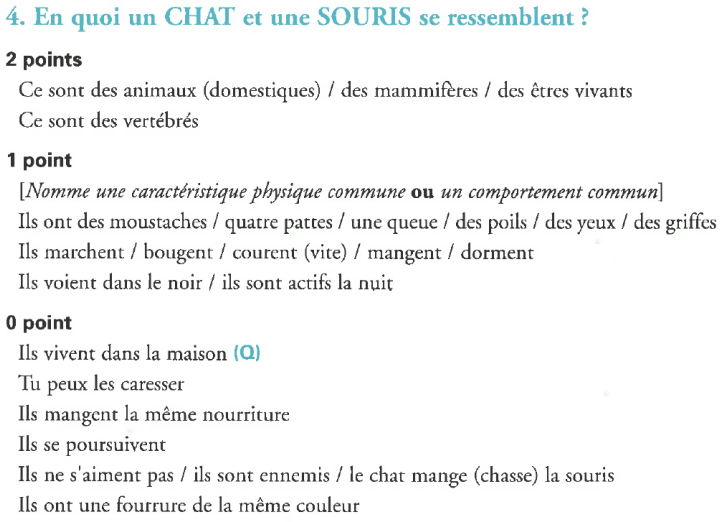
5.3.2.2. Similitudes

Dans le subtest Similitudes (SIM), le psychologue lit oralement des paires de mots à l'enfant. Pour chaque paire, l'enfant doit trouver la similitude soit entre deux objets soit entre deux concepts. Par exemple à l'item 4, on demande à l'enfant : *En quoi un CHAT et une SOURIE se ressemblent* ? (voir Figure 34, p. 183). Ce subtest comporte 23 items de difficulté croissante avec des paires de mots concrets (p. ex., item 3 : *chemise – chaussure*) et des paires de mots abstraits (p. ex., item 12 : *colère – joie*).

Règle de départ : après l'item d'exemple, départ à l'item 1 pour les 6 – 8 ans, à l'item 3 pour les 9 – 11 ans et à l'item 5 pour les 12 – 16 ans.

Règle d'arrêt : après 5 notes 0 consécutives.

Cotation : Les deux premiers items sont des items d'apprentissage, et sont cotés 0 ou 1 point. Les 21 items suivants sont cotés 0, 1 ou 2 points. La note parfaite est attribuée aux réponses correspondant à une catégorisation générale pertinente et avec un degré d'abstraction, tandis que les réponses plus concrètes et qui font référence à des caractéristiques communes à la paire de mots sont cotés 1 point. Le score brut maximum est de 44 points.



4. En quoi un CHAT et une SOURIS se ressemblent ?

2 points

Ce sont des animaux (domestiques) / des mammifères / des êtres vivants
Ce sont des vertébrés

1 point

[*Nomme une caractéristique physique commune* **ou** *un comportement commun*]
Ils ont des moustaches / quatre pattes / une queue / des poils / des yeux / des griffes
Ils marchent / bougent / courent (vite) / mangent / dorment
Ils voient dans le noir / ils sont actifs la nuit

0 point

Ils vivent dans la maison (Q)
Tu peux les caresser
Ils mangent la même nourriture
Ils se poursuivent
Ils ne s'aiment pas / ils sont ennemis / le chat mange (chasse) la souris
Ils ont une fourrure de la même couleur

Figure 34. Illustration de l'épreuve Similitudes.

Similitudes est un subtest obligatoire de l'ICV. Selon le manuel du WISC-IV, il fait principalement appel aux capacités de penser les concepts, ainsi qu'aux capacités de formation de concepts et de catégories hiérarchisées (Wechsler, 2005b). Il évalue donc la capacité d'abstraction verbale de conceptualisation et de catégorisation. Selon le modèle CHC, ce subtest est un bon indicateur l'aptitude étendue Intelligence cristallisée Gc et plus précisément, l'aptitude restreinte Connaissances lexicales VL.

5.3.2.3. Mémoire des chiffres

Dans le subtest Mémoire des chiffres (MCH), le psychologue énonce au rythme d'un chiffre par seconde une suite de chiffres (de plus en plus longue) à l'enfant, qui doit les répéter dans le même ordre, ou dans l'ordre inverse (voir Figure 35, p. 184). Par exemple à l'item 3 en ordre direct, on énonce à l'enfant : *3 – 4 – 1 – 7*. L'enfant devra

répéter cette suite dans le même ordre. Pour la partie en ordre inverse, à l'item 2, on énonce à l'enfant : *3 – 5*. L'enfant devra répéter cette suite dans l'ordre inverse, soit 5 – 3. Ce subtest comporte 8 items pour la partie en ordre direct et 8 items pour la partie en ordre inverse.

Règle de départ : à l'item 1 pour les 6 – 16 ans.

Règle d'arrêt : après 2 notes 0 consécutives aux deux essais d'un même item.

Cotation : chaque item comporte deux essais. Les items sont cotés 0, 1 ou 2 points selon le nombre d'essais réussis. Le score brut maximum est de 32 points.

Item	Essai
1. Essai 1	2 – 9
Essai 2	4 – 6

Figure 35. Illustration de l'épreuve Mémoire des chiffres en ordre direct.

Mémoire des chiffres est un subtest obligatoire de l'IMT. Selon le manuel du WISC-IV, il fait principalement appel à la mémoire auditive à court terme, à des capacités de séquençage, à l'attention et à la concentration (Wechsler, 2005b). Selon le modèle CHC, ce subtest est un bon indicateur de l'aptitude étendue Mémoire de travail Gwm et plus précisément, l'aptitude restreinte Empan mnésique MS (ordre direct) ou Mémoire de travail MW (ordre inverse).

5.3.2.4. Identification de concepts

Dans le subtest Identification de concepts (IDC), l'enfant doit identifier les images (une par rangée) qui s'associent autour d'un concept commun parmi deux, puis trois rangées d'images (voir Figure 36, p. 185). Ce subtest comporte 28 items au total.

Règle de départ : après deux items d'exemple, départ à l'item 1 pour les 6 – 8 ans et à l'item 5 pour les 9 – 11 ans et à l'item 7 pour les 12 – 16 ans.

Règle d'arrêt : après 5 notes 0 consécutives.

Cotation : les items qui sont cotés 0 ou 1 point. Le score brut maximum est de 28 points. D'abord sur deux rangées, puis à partir de l'item 13, les images sont sur trois rangées.

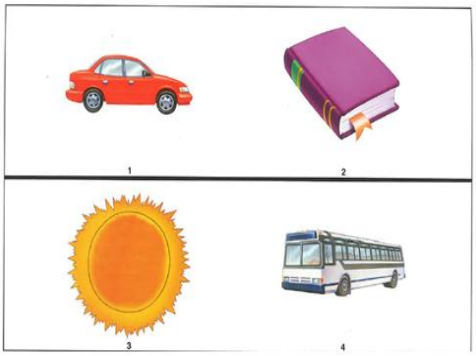


Figure 36. Illustration de l'épreuve Identification de concepts.

Identification d'images est un subtest obligatoire de l'IRP. Selon le manuel du WISC-IV, il fait principalement appel aux aptitudes de raisonnement catégoriel et de raisonnement abstrait (Wechsler, 2005b). Il évalue le raisonnement inductif, les capacités perceptives visuelles, la (re)connaissance des objets, l'inhibition et la flexibilité mentale. Selon le modèle CHC, ce subtest est un bon indicateur de l'aptitude étendue Intelligence fluide Gf et plus précisément, l'aptitude restreinte Raisonnement séquentiel général RG.

5.3.2.5. Code

Dans le subtest Code (COD), des chiffres sont appariés à des symboles. En dessinant le symbole correspondant au chiffre, l'enfant doit compléter le plus de cases en un temps défini de 2 minutes (voir Figure 37, p. 186). Deux versions sont à disposition, le Code A pour les 6-7 ans et le Code B pour les 8-16 ans.

Règle de départ : après des items d'entraînement, départ à l'item 1 pour tous.

Règle d'arrêt : après 120 secondes ou si tous les items sont complétés avant le temps imparti.

Cotation : pour le Code A, il y a des points de bonification du temps si l'enfant termine avant les 2 minutes. Chaque signe correct donne un point. Le score brut maximum est de 65 points pour le Code A et de 119 points pour le Code B.

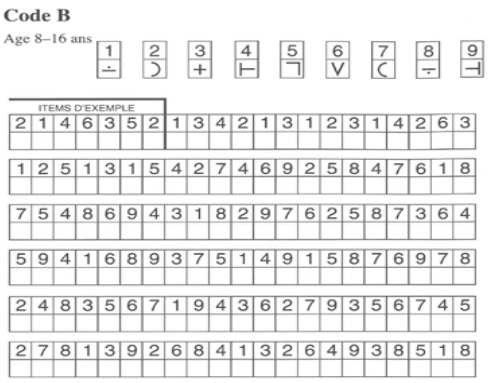


Figure 37. Illustration du feuillet de passation de l'épreuve Code B.

Code est un subtest obligatoire de l'IMT. Selon le manuel du WISC-IV, il fait principalement appel à la vitesse de traitement, à la capacité d'apprentissage, à la perception visuelle, à la coordination visuomotrice, à la capacité de balayage visuel, à l'attention et à la motivation (Wechsler, 2005b). Il est influencé par la mémoire à court terme visuelle. Selon le modèle CHC, ce subtest est un bon indicateur de l'aptitude étendue Vitesse de traitement Gs et plus précisément, l'aptitude restreinte *Rate-of-test-taking* R9.

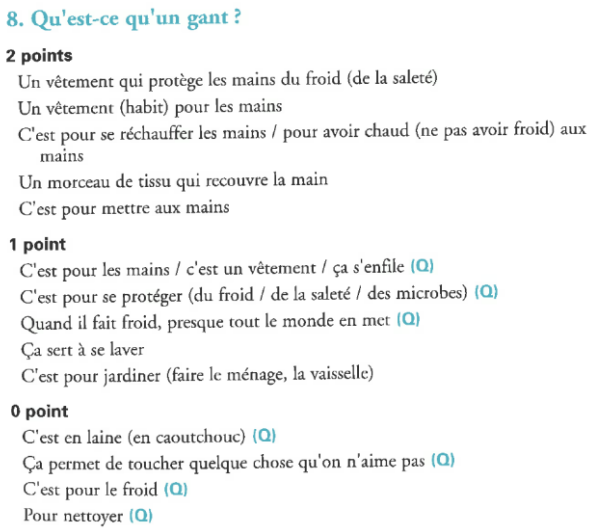
5.3.2.6. Vocabulaire

Dans le subtest Vocabulaire (VOC), le psychologue montre des images que l'enfant doit nommer ou lit oralement des mots que l'enfant doit définir. Par exemple à l'item 8, on demande à l'enfant : *qu'est-ce qu'un gant* ? (voir Figure 38, p. 187). Ce subtest comporte 36 items de difficulté croissante.

Règle de départ : à l'item 5 pour les 6 – 8 ans et à l'item 7 pour les 9 – 11 ans et à l'item 9 pour les 12 – 16 ans.

Règle d'arrêt : après 5 notes 0 consécutives.

Cotation : les 4 premiers items sont des images à nommer et sont cotés 0 ou 1 point. Les 32 items verbaux sont cotés 0, 1 ou 2 points. La note parfaite est attribuée s'il y a une bonne compréhension du mot (p. ex., bon synonyme, caractéristiques fondamentales). La note 1 est attribuée pour les réponses correctes, mais dont le contenu est pauvre et peu élaboré. Le score brut maximum est de 68 points.



8. Qu'est-ce qu'un gant ?

2 points

Un vêtement qui protège les mains du froid (de la saleté)
Un vêtement (habit) pour les mains
C'est pour se réchauffer les mains / pour avoir chaud (ne pas avoir froid) aux mains
Un morceau de tissu qui recouvre la main
C'est pour mettre aux mains

1 point

C'est pour les mains / c'est un vêtement / ça s'enfile (Q)
C'est pour se protéger (du froid / de la saleté / des microbes) (Q)
Quand il fait froid, presque tout le monde en met (Q)
Ça sert à se laver
C'est pour jardiner (faire le ménage, la vaisselle)

0 point

C'est en laine (en caoutchouc) (Q)
Ça permet de toucher quelque chose qu'on n'aime pas (Q)
C'est pour le froid (Q)
Pour nettoyer (Q)

Figure 38. Illustration de l'épreuve Vocabulaire.

Vocabulaire est un subtest obligatoire de l'ICV. Selon le manuel du WISC-IV, il fait principalement appel aux connaissances du lexique, aux capacités de formation de concepts verbaux, à la capacité d'apprendre, à la mémoire à long terme et au niveau de développement du langage (Wechsler, 2005b). Il évalue les connaissances sur les concepts verbaux et la capacité d'élaboration du langage oral. Selon le modèle CHC, ce subtest est un bon indicateur de l'aptitude étendue Intelligence cristallisée Gc et plus précisément, l'aptitude restreinte Connaissances lexicales VL.

5.3.2.7. Séquence Lettres-Chiffres

Dans le subtest Séquence Lettres-Chiffres (SLC), le psychologue énonce une suite mélangée de chiffres et de lettres (de plus en plus longue) à l'enfant, qui doit restituer en premier les chiffres, en ordre croissant, puis les lettres dans l'ordre alphabétique (voir Figure 39, p. 188). Par exemple à l'item 6, on énonce à l'enfant : *D – 8 – M – 1*. L'enfant devra redonner 1 – 8 – D – M. Ce subtest comporte 10 items avec trois essais différents chacun.

Règle de départ : pour les enfants de 6 – 7 ans, on leur demande de compter jusqu'à trois et de réciter l'alphabet jusqu'à C. Pour les 8 – 16 ans, on les exerce sur deux exemples, puis on démarre à l'item 1.

Règle d'arrêt : après 3 notes 0 consécutives aux trois essais d'un même item.

Cotation : les items sont cotés 0, 1, 2 ou 3 points selon le nombre d'essais réussis. Comme la difficulté de la tâche demeure la même, le point est également accordé si l'enfant donne d'abord les lettres dans l'ordre alphabétique, puis les chiffres dans l'ordre croissant. Le score brut maximum est de 30 points.

Item	Essai	Réponses correctes
3.	1. **B – 1 – 2**	1 – 2 – B B – 1 – 2
	2. **1 – 3 – C**	1 – 3 – C C – 1 – 3
	3. **2 – A – 3**	2 – 3 – A A – 2 – 3

Figure 39. Illustration de l'épreuve Séquence Lettres-Chiffres.

Séquence Lettres-Chiffres est un subtest obligatoire de l'IMT. Selon le manuel du WISC-IV, il fait principalement appel au séquençage, à la capacité à manipuler mentalement, à l'attention, à la mémoire auditive à court terme, aux représentations visuospatiales et à la vitesse de traitement (Wechsler, 2005b). Selon le modèle CHC, ce subtest est un bon indicateur de l'aptitude étendue Mémoire de travail Gwm et plus précisément, l'aptitude restreinte Mémoire de travail MW.

5.3.2.8. Matrices

Dans le subtest Matrices (MAT), l'enfant doit choisir la partie manquante parmi cinq propositions dans une matrice incomplète (voir Figure 40, p. 189). Ce subtest comporte 35 items.

Règle de départ : après trois items d'exemple, départ à l'item 4 pour les de 6 – 8 ans, à l'item 7 pour les 9 – 11 ans et à l'item 11 pour les 12 – 16 ans.

Règle d'arrêt : après 4 notes 0 consécutives ou 4 notes 0 à cinq items consécutifs.

Cotation : les items sont cotés 0 ou 1 point. Le score brut maximum est de 35 points.

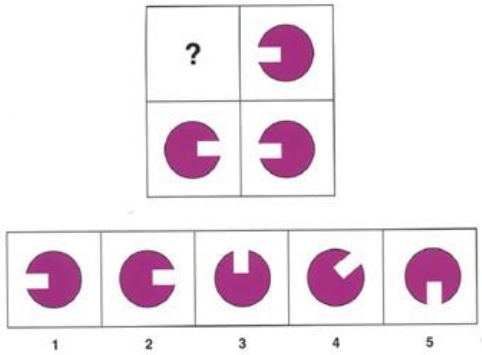


Figure 40. Illustration de l'épreuve Matrices.

Matrices est un subtest obligatoire de l'IRP. Selon le manuel du WISC-IV, il fait principalement appel aux capacités de traitement de l'information visuelle ainsi que de raisonnement abstrait (Wechsler, 2005b). Il évalue l'intelligence fluide et visuospatiale, et dans une certaine mesure la mémoire de travail et le raisonnement déductif. Selon le modèle CHC, ce subtest est un bon indicateur de l'aptitude étendue Intelligence fluide Gf, et plus précisément l'aptitude restreinte Induction I.

5.3.2.9. Compréhension

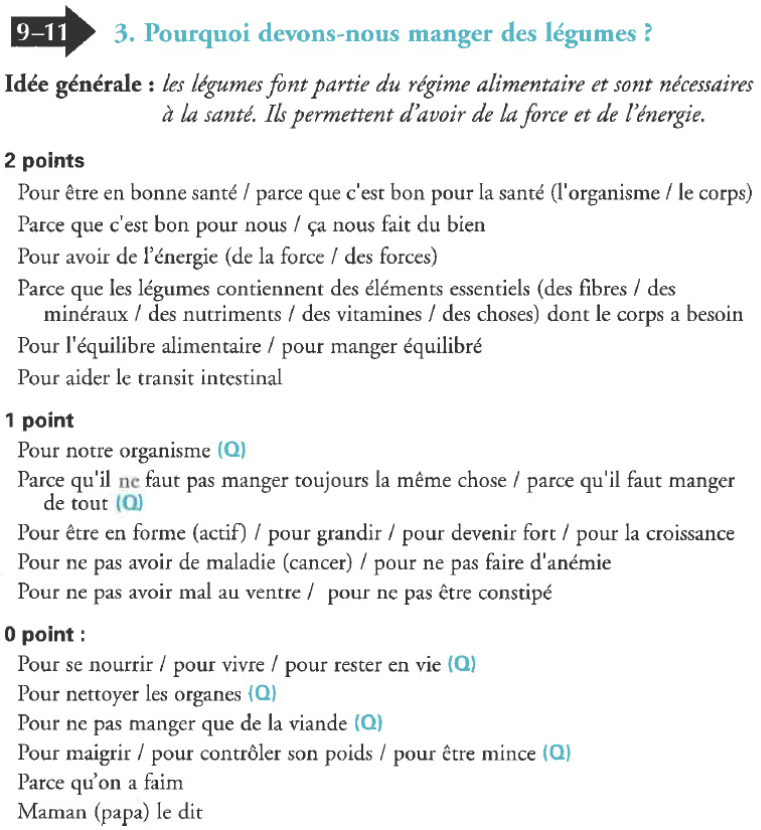
Dans le subtest Compréhension (COM), l'enfant répond à des questions relatives à des situations de la vie quotidienne, sociale ou interpersonnelle. Par exemple à l'item 3, on demande à l'enfant : *pourquoi devons-nous manger des légumes ?* (voir Figure 41, p. 190). Ce subtest comporte 21 items.

Règle de départ : pour les enfants de 6 – 7 ans, on leur demande de compter jusqu'à trois et de réciter l'alphabet jusqu'à C. Pour les 8 – 16 ans, on les exerce sur deux exemples, puis on démarre à l'item 1.

Règle d'arrêt : après 4 notes 0 consécutives.

Cotation : les items sont cotés 0, 1 ou 2 points. La note parfaite est attribuée aux réponses de qualité qui indique une bonne compréhension de l'idée générale recherchée. La note 1 est attribuée des réponses qui vont dans le sens de l'idée générale, mais pas suffisamment élaborée. Le score brut maximum est de 42 points.

Compréhension est un subtest obligatoire de l'ICV. Selon le manuel du WISC-IV, il fait principalement appel à la connaissance des conventions, à l'aptitude à utiliser des expériences passées ainsi qu'à la capacité à expliquer des problèmes de la vie quotidienne, sociale et interpersonnelle (Wechsler, 2005b). Il évalue les connaissances sociales et générales de l'enfant. Cependant, il ne permet pas d'évaluer si l'enfant saurait appliquer adéquatement ces connaissances en situation réelle. Selon le modèle CHC, ce subtest est un bon indicateur de l'aptitude étendue Intelligence cristallisée Gc et plus précisément, l'aptitude restreinte Information verbale générale KO.



9–11 3. Pourquoi devons-nous manger des légumes ?

Idée générale : *les légumes font partie du régime alimentaire et sont nécessaires à la santé. Ils permettent d'avoir de la force et de l'énergie.*

2 points

Pour être en bonne santé / parce que c'est bon pour la santé (l'organisme / le corps)
Parce que c'est bon pour nous / ça nous fait du bien
Pour avoir de l'énergie (de la force / des forces)
Parce que les légumes contiennent des éléments essentiels (des fibres / des minéraux / des nutriments / des vitamines / des choses) dont le corps a besoin
Pour l'équilibre alimentaire / pour manger équilibré
Pour aider le transit intestinal

1 point

Pour notre organisme (Q)
Parce qu'il ne faut pas manger toujours la même chose / parce qu'il faut manger de tout (Q)
Pour être en forme (actif) / pour grandir / pour devenir fort / pour la croissance
Pour ne pas avoir de maladie (cancer) / pour ne pas faire d'anémie
Pour ne pas avoir mal au ventre / pour ne pas être constipé

0 point :

Pour se nourrir / pour vivre / pour rester en vie (Q)
Pour nettoyer les organes (Q)
Pour ne pas manger que de la viande (Q)
Pour maigrir / pour contrôler son poids / pour être mince (Q)
Parce qu'on a faim
Maman (papa) le dit

Figure 41. Illustration de l'épreuve Compréhension.

5.3.2.10. Symboles

Dans le subtest Symboles (SYM), l'enfant doit repérer dans une série de 5 symboles, si OUI ou NON, l'un des 2 symboles isolés en début de série s'y retrouve (voir Figure 42, p. 191). S'il retrouve l'un des deux symboles isolés dans la série de symboles isolés, il coche OUI et, s'il ne retrouve aucun des symboles isolés, il coche NON. L'enfant

dispose d'un temps défini de 2 minutes. Deux versions sont à disposition, le Symboles A pour les 6-7 ans et le Symboles B pour les 8-16 ans.

Règle de départ : après des items d'entraînement, départ à l'item 1 pour tous.

Règle d'arrêt : après 120 secondes.

Cotation : chaque réponse correcte donne un point. Les éventuelles erreurs sont soustraites du total des points. Le score brut maximum est de 45 points pour le Symboles A et de 60 points pour le Symboles B.

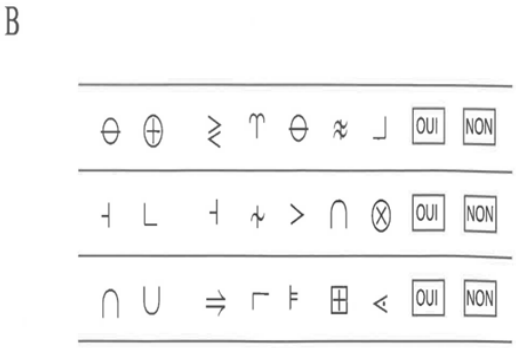


Figure 42. Illustration du feuillet de passation de l'épreuve Symboles B.

Symboles est un subtest obligatoire de l'IMT. Selon le manuel du WISC-IV, il fait principalement appel à la vitesse de traitement, à la coordination visuomotrice, à la discrimination visuelle et à la concentration (Wechsler, 2005b). Il met en jeu des capacités de traitement perceptif visuel, de discrimination visuelle et de rapidité cognitive sur une tâche simple. Selon le modèle CHC, ce subtest est un bon indicateur de l'aptitude étendue Vitesse de traitement Gs, et plus précisément, l'aptitude restreinte Vitesse perceptive P.

5.3.2.11. Complètement d'images

Dans le subtest Complètement d'images (CIM), le psychologue montre une image à l'enfant qui doit, dans un temps limité (20 secondes maximum), nommer ou pointer la partie importante manquante. Dans l'exemple de la Figure 43 (p. 192), on attend que le sujet relève qu'il manque une oreille au renard. Ce subtest comporte 38 items.

Règle de départ : après un item d'exemple, départ à l'item 1 pour les 6 – 8 ans, à l'item 5 pour les 9 – 11 ans et à l'item 10 pour les 12 – 16 ans.

Règle d'arrêt : après 6 notes 0 consécutives.

Cotation : chaque réponse correcte et dans le temps imparti donne un point. Le score brut maximum est de 38 points.

Figure 43. Illustration de l'épreuve Complétement d'images.

Complètement d'images est un subtest optionnel de l'IRP. Selon le manuel du WISC-IV, il fait principalement appel à la perception et à l'organisation visuelle, ainsi qu'à la concentration et à la reconnaissance visuelle des détails essentiels des objets (Wechsler, 2005b). Il évalue les capacités de discrimination, de reconnaissance visuelle et d'identification, la représentation visuelle en mémoire à long terme et la flexibilité mentale. Selon le modèle CHC, ce subtest est un bon indicateur de l'aptitude étendue Traitement visuel Gv et plus précisément, l'aptitude restreinte Flexibilité de fermeture CF.

6. ANALYSES DE DONNÉES

Dans ce chapitre, nous décrirons les étapes des analyses réalisées sur les données de l'échantillon des 483 enfants (étude 1) et de l'échantillon des 277 enfants (étude 2). Chaque étude s'inscrit dans un cadre de mesure différente à savoir la théorie de réponse à l'item pour l'étude 1 et la théorie classique des tests pour l'étude 2.

6.1. ÉTUDE 1 : FONCTIONNEMENT DIFFÉRENTIEL DES ITEMS

Dans l'étude du fonctionnement différentiel des items, l'objectif est de détecter la présence ou non d'items biaisés pour un groupe d'individus selon leur appartenance à un sexe ou à un milieu socio-économique. Un item biaisé compromet la validité de l'interprétation du score du test. Comme la réussite ou l'échec de l'item n'est pas expliqué par la propriété mentale évaluée par le test, un item biaisé ne contribue pas à estimer le niveau du sujet sur ladite propriété mentale (ou le trait latent). Dans le cas d'un biais, le comportement de l'item (probabilité de réussir ou d'échec) varie d'un groupe d'individus à l'autre (p. ex., entre les garçons et les filles), et non en fonction de l'habileté du sujet sur le trait latent évalué par le test. Les analyses de cette première étude portent sur les items des différents subtests du WISC-IV. Nous allons exposer plus en détail la démarche dans ce qui suit.

Les données des protocoles WISC-IV sont rentrées item par item pour chacun des onze subtests administrés. Cependant, les deux épreuves de l'Indice de Mémoire de Travail (IMT) et les deux épreuves de l'Indice de Vitesse de Traitement (IVT) n'ont pas été analysées. Pour la mémoire, il est théoriquement peu justifié de supposer la possibilité d'un biais à l'encontre d'un groupe sur les items proposés par les subtests de l'IMT. Pour les subtests de l'IVT, il n'est pas possible d'ajuster un modèle de réponses à l'item. S'agissant de tâches simples à réaliser le plus vite possible, tous les items de Code ou de Symboles sont réussis par tous. C'est le manque de temps ou des erreurs d'inattention qui empêchent de les compléter. Tous les items doivent donc avoir la même difficulté et la même discrimination qu'ils soient au début ou à la fin du subtest. Comme nous n'avons pas administré tous les items des subtests de l'IVT, il est aberrant d'estimer des paramètres d'item (difficulté, discrimination) sur nos données. En effet, les estimations de la difficulté et de la discrimination seront faussées à cause des items non complétés à cause de la fin du temps imparti. Ainsi, seuls 7 subtests sont a priori

adéquats pour une modélisation par un modèle de mesure tiré des modèles de réponses à l'item, à savoir Cubes, Similitudes, Identification de concepts, Vocabulaire, Matrices, Compréhension et Complètement d'images.

Le choix du modèle de réponse à l'item pour la modalisation est fonction des caractéristiques des données à disposition. Les subtests du WISC-IV sont supposés évaluer une dimension. De plus, les items des subtests analysés n'ont pas été construits pour avoir une discrimination identique. Comme le modèle à un paramètre de Rasch contraint le paramètre de discrimination à une valeur identique pour tous les items, il n'est pas le plus adéquat pour les données des subtests du WISC-IV. Toutefois, du fait de sa simplicité, le modèle de Rash a été utilisé par les concepteurs du WISC-IV. Pour notre part, nous optons pour un modèle unidimensionnel de réponse à l'item à deux paramètres librement estimés, soit le modèle unidimensionnel logistique à deux paramètres (2 PL) défini par Birnbaum (1968, 1969). Le modèle de Birnbaum est fréquemment utilisé pour le cas de variables à échelle catégorielle et à cotation dichotomiques. À partir du modèle 2 PL de Birnbaum, Samejima (1970, 1997) propose une généralisation pour le cas des items à cotation polytomique, soit le modèle gradué de Samejima. Le modèle gradué est adapté même pour un nombre variable de catégories parmi les items d'un même subtest. Par exemple, dans l'épreuve Cubes, les points obtenus varient soit de 0, 1 ou 2 points, soit 0 ou 4 points, soit de 0, 4, 5, 6 ou 7 points selon les items. Le modèle gradué de Samejima est fréquemment utilisé pour le cas de variables à échelle ordinale et à cotation polytomique. Ainsi, nous appliquons le modèle 2 PL aux items à cotation dichotomique des subtests Identification de concepts, Matrices et Complètement d'images, tandis que le modèle gradué de Samejima est appliqué aux items à cotation polytomique des subtests Cubes, Similitudes, Vocabulaire et Compréhension.

Le choix de l'estimateur pour l'estimation des paramètres des modèles s'est porté sur l'estimateur par défaut du logiciel M*plus* 7.2 de Muthén & Muthén avec lequel les analyses sont réalisées. Pour les données catégorielles, les concepteurs de M*plus* recommandent un estimateur dérivé de la méthode des moindres carrés pondérés (*weighted least square ; WLS*) qu'ils abrègent WLSMV. Pour les données catégorielles, l'estimateur WLSMV est plus robuste que l'estimateur de la méthode du maximum de vraisemblance (*maximum likelihood* ; *ML*). Pour une comparaison plus approfondie et plus technique entre les estimateurs WLSMV et ML, nous renvoyons à la lecture de T. A. Brown (2015), Muthén, du Toit et Spisic (1997) ou Yu (2002).

Les estimations des paramètres d'item (difficulté et discrimination) et de l'échelle d'habileté du trait latent θ sont facilitées si les données s'ajustent bien au modèle unidimensionnel postulé pour l'application des modèles 2 PL et gradué de Samejima. Généralement, l'ajustement des données à un modèle faiblit lorsqu'il y a des items de variance nulle ou très faible, car ces items-ci ne contribuent pas à donner de l'information ni pour l'estimation de leurs paramètres (p. ex., difficulté, discrimination) ni pour l'estimation de l'échelle du trait latent θ. Un item sans variance est réussi ou échoué par tous les individus quel que soit leur habileté sur le trait latent θ. L'ajustement des données à un modèle faiblit également lorsqu'il y a un certain nombre de données contradictoires. Ces dernières rendent instables les estimations entre la probabilité de réussir l'item pour un individu possédant un certain degré d'habileté et les paramètres de l'item. Sur un échantillon très important, l'estimation d'un modèle tend à être moins sensible aux données incohérentes et à mieux se stabiliser puisqu'il y a alors assez d'informations pour aider à l'estimation. La grande limitation à l'utilisation des MRI vient de la nécessite d'un très large échantillon et cela d'autant plus qu'il y a de paramètres librement estimés.

Les analyses réalisées par le logiciel M*plus* 7.2 de Muthén & Muthén permettent de tester le fonctionnement différentiel des items (FDI) selon l'approche des modèles Multiples Indicateurs et Multiples Causes (MIMIC : *multiple-indicators multiple-causes*). Dans cette approche, une variable latente est évaluée par plusieurs indicateurs et est expliquée par plusieurs autres variables (Woods, 2009 pour en savoir plus sur l'approche). Nous rappelons qu'un FDI apparaît lorsque les individus issus de différents groupes (p. ex., ethnique, sexe) avec une même habileté sur le trait latent n'ont pas la même probabilité de réussir un même item. La notion de fonctionnement différentiel des items (ou biais) est liée à la notion d'invariance. Si des individus issus de différents groupes avec une même habileté sur le trait latent ont la même probabilité de réussir un même item, alors on suppose l'invariance (ou l'équivalence) du score et des comparaisons entre les individus sont alors possibles. On relève deux types de fonctionnement différentiel des items : (1) le FDI uniforme qui survient uniformément tout au long de l'échelle du trait latent et (2) le FDI non uniforme qui ne survient pas uniformément tout au long de l'échelle du trait latent. Par exemple, un FDI peut apparaître entre le groupe des filles et des garçons seulement pour de individus de très faible ou de très élevée habileté sur le trait latent (FDI non uniforme). La version que nous avons de M*plus* 7.2 ne permet que la détection d'un FDI uniforme.

6.2. ÉTUDE 2 : STABILITÉ DES SCORES DU WISC-IV

Dans les publications, la stabilité des scores de tests d'intelligence est évaluée principalement au niveau interindividuel, et plus précisément par le biais d'un coefficient de fidélité. Or, les différences de moyennes entre les passations renseignent utilement sur le niveau de changement interindividuel et mettent en évidence d'éventuels effets d'apprentissage. De plus, des analyses au niveau intra-individuel sont essentielles pour fournir des pistes pertinentes aux cliniciens, qui, dans leur pratique, s'intéressent à l'évaluation d'un individu singulier. L'une des finalités de l'évaluation de l'intelligence d'un enfant est non seulement de faire un bilan sur une situation présente, mais également de pouvoir faire des prédictions quant à ses futures performances au même test et plus largement, extrapoler sur sa réussite scolaire, sociale ou professionnelle. C'est donc bien le niveau intra-individuel qui importe aux praticiens et aux clients. Afin d'estimer dans quelle mesure l'interprétation des scores du WISC-IV permet des prédictions, il est nécessaire d'évaluer la stabilité à long terme des scores aussi bien au niveau inter- qu'intra-individuel.

Les analyses sont réalisées pour les indices standards (ICV, IRP, etc.) et pour les indices CHC (Gf, Gc, etc.). Les protocoles récoltés ont donné lieu à des analyses descriptives (moyennes, *d* de Cohen, pourcentages, etc.), des comparaisons de moyennes (*t*-test pour échantillons appariés) et des calculs de corrélations.

6.2.1. STABILITÉ SUR LE PLAN INTERINDIVIDUEL

Deux analyses sont réalisées au niveau du groupe : les comparaisons de moyennes entre les deux passations (stabilité absolue) et les corrélations entre les performances aux deux passations (stabilité différentielle).

6.2.1.1. Stabilité absolue

La stabilité absolue examine la différence de moyennes entre le test et le retest. Au niveau du groupe et par comparaisons de moyennes (*t*-tests pour groupes appariés), nous déterminons si les moyennes aux deux passations sont équivalentes. Si les résultats du *t*-test montrent une différence statistiquement significative entre les moyennes, on peut exclure qu'elle soit due au hasard. Toutefois, une différence

significative sur le plan statistique peut être insignifiante sur le plan clinique. La taille de l'effet (d de Cohen) renseigne sur la magnitude ou l'importance de la différence de moyennes observée.

6.2.1.2. Stabilité différentielle

La stabilité différentielle examine si l'ordre des individus est similaire entre le test et le retest. À partir des performances au test et au retest, il s'agit de calculer un coefficient de corrélation test-retest (ou coefficient de stabilité). Outre le coefficient de stabilité non corrigé (r_{12}), nous calculons un coefficient de stabilité corrigé (r_c) selon la formule de Magnusson (1967). La procédure développée par Magnusson applique une correction sur les coefficients de fidélité pour tenir compte des problèmes d'homogénéité (ou d'hétérogénéité) dans un échantillon par rapport à la variabilité qu'on retrouve dans la population générale. Pour rappel, les coefficients de corrélation sont sensibles à l'homogénéité/hétérogénéité des échantillons. Ainsi, dans un échantillon trop homogène, la réduction de l'étendue des différences interindividuelles entraîne une diminution du coefficient de corrélation, tandis qu'à l'inverse, dans un échantillon trop hétérogène, la corrélation tend à être surestimée.

6.2.2. STABILITÉ SUR LE PLAN INTRA-INDIVIDUEL

La stabilité intra-individuelle peut être traitée sous divers aspects. Trois analyses sont réalisées au niveau des individus : le pourcentage d'enfants dont les performances entre les deux passations sont comprises dans un intervalle défini par l'erreur type de mesure (stabilité intra-individuelle absolue), le pourcentage d'enfants qui sont restés dans la même catégorie descriptive entre les deux passations (stabilité catégorielle)le et le pourcentage d'enfants qui présentent les mêmes forces et/ou faiblesses personnelles aux deux passations (stabilité des forces et faiblesses personnelles).

6.2.2.1. Stabilité intra-individuelle absolue

La stabilité intra-individuelle absolue examine le pourcentage d'enfants qui présentent des performances stables entre les deux passations. Les performances sont

considérées comme stables si elles restent à l'intérieur d'un intervalle de confiance correspondant à ±2 erreurs types de mesure pour chaque score (2 ETM étant l'arrondi de 1.96 ETM). Cet intervalle de confiance est régulièrement utilisé dans la littérature (Canivez & Watkins, 2001; Lander, 2010). Comme nous l'avons déjà mentionné, l'erreur type de mesure[24] (ou erreur standard de mesure) représente le degré de dispersion théorique des scores observés d'un individu qui passerait un même test de façon répétée. En effet, « si l'on répétait un grand nombre de fois la mesure . . . on observerait une distribution normale des notes observées (du fait du caractère aléatoire de l'erreur) ayant pour moyenne la note vraie et dont l'écart type serait la distribution des erreurs » (Lautrey, 2006, p. 344). En vertu de la distribution normale, moins de 5 % des enfants devraient théoriquement voir leurs performances varier en dehors des limites de l'intervalle de ±2 ETM. De même, approximativement 95 % des enfants devraient présenter des variations de performances incluses dans cet intervalle.

Nous allons développer les étapes de cette analyse qui concerne les scores des subtests, des indices standards et des facteurs CHC.

a) Il s'agit d'abord de calculer la différence de performances entre le test (T1) et le retest (T2) pour chaque enfant et pour chaque score obtenu (subtests, indices standards et CHC).

Pour tenir compte du phénomène de régression à la moyenne, nous calculons les scores vrais estimés (V_{est}) pour les scores à T1[25]. Prenons l'exemple d'un enfant qui obtient un score de 118 à T1 puis de 123 à T2 pour le QIT. Le coefficient de fidélité du QIT est de .94 (Wechsler, 2005b, p. 30, Tableau 4.1). Le score vrai estimé à T1 est donc $V_{est} = 100 + .94\,(118 - 100) = 116.92$. La différence de scores QIT entre test et retest $= |116.92 - 123| = 6.08$ points.

b) La différence de performances est ensuite comparée à l'intervalle de ±2 ETM pour le score considéré.

Les erreurs types de mesure des subtests et des indices standard sont reprises du *Manuel d'interprétation* du WISC-IV (Wechsler, 2005b, p. 32, Tableau 4.2). Pour IAG, ICC et les indices CHC, les valeurs sont reprises de précédents travaux de notre équipe (Lecerf et al., 2012; Lecerf, Reverte, Coleaux, Favez, & Rossier, 2010). Dans le tableau de l'Annexe F sont rapportés les coefficients de fidélité ainsi que les valeurs correspondant à 1 ETM, 2 ETM et 3 ETM calculés pour chaque score étudié. Si nous poursuivons notre

[24] $ETM = \sigma\sqrt{1 - r_{xx}}$

[25] Voir équation (22) p. 144.

exemple, l'ETM du QIT est de 3.63 points, ce qui définit un intervalle ±2 ETM de -7.26 à +7.26 (2 x 3.63). On considère comme performance stable, si le score observé du QIT lors de la seconde passation se situe dans l'intervalle de 109.66 à 124.18 (116.92 – 7.26 et 116.92 + 7.26). Une autre façon de faire est d'examiner si la différence de score entre les deux passations est supérieure à 2 ETM. Dans notre exemple, il s'agit d'une différence de 6.08 points entre les deux passations ($|116.92 - 123| = 6.08$), qui n'est donc pas supérieure à 7.26 points.

6.2.2.1. Stabilité catégorielle

La stabilité catégorielle examine le pourcentage d'enfants dont les performances se retrouvent dans la même catégorie descriptive au test et au retest. Pour donner un sens plus parlant au score numérique, les cliniciens décrivent les performances de manière qualitative. La stabilité à long terme des catégories qualifiant les performances au WISC-IV est évaluée d'après trois classifications répandues. La classification traditionnelle, proposée dans le *Manuel d'interprétation* (Wechsler, 2005b, p. 87), décrit sept catégories : très faible (≤ 69), limite (70-79), moyen faible (80-89), moyen (90-109), moyen fort (110-119), supérieur (120-129), et très supérieur (≥ 130). Plus répandue auprès des neuropsychologues, nous trouvons la classification des performances en trois catégories : faible (≤ 84), dans la moyenne (85-115), et élevé (≥ 116) ; (Flanagan & Kaufman, 2009). Ce système en trois catégories rejoint la lecture en faiblesse, moyenne et force normatives. Dans un entre-deux, Flanagan et Kaufman (2009, p. 137) recommandent une classification alternative en cinq catégories : extrémité inférieure (≤ 69), moyen faible (70-84), dans la moyenne (85-115), moyen fort (116-130), et extrémité supérieure (≥ 131).

Cette analyse implique les indices standards (IAG et ICC inclus). Les scores obtenus à T1 et à T2 de chaque enfant sont recodés suivant les catégories des trois systèmes de classification. Prenons par exemple, un enfant qui obtient 112 pour l'ICV à la première passation. Dans le système traditionnel en sept catégories, la performance est qualifiée « moyen fort ». Dans la classification en trois catégories, elle est qualifiée « dans la moyenne ». Dans la classification en cinq catégories, elle est également qualifiée « dans la moyenne ». À la seconde passation, si l'enfant obtient 119 à l'ICV. Dans le système en sept catégories, la performance à l'ICV reste dans la catégorie « moyen fort », il s'agit d'une performance stable au niveau catégoriel. Dans le système en trois catégories, la performance à T2 passe dans la catégorie « élevé ». La

performance n'est donc pas stable au niveau catégoriel. Dans la classification en cinq catégories, l'ICV à T2 est qualifiée « moyen fort ». La performance n'est pas stable au niveau catégoriel.

6.2.2.2. Stabilité des forces et des faiblesses personnelles

La stabilité des forces et faiblesses personnelles examine le pourcentage d'enfants qui présentent des forces et/ou faiblesses personnelles stables entre les deux passations. Une force (ou une faiblesse) personnelle « stable » pour un indice donné est une force (ou une faiblesse) personnelle qui est présente dans les performances au test et au retest. L'identification des forces et des faiblesses personnelles suit la démarche de Silverstein (1982), qui a été reprise notamment par Grégoire et Wierzbicki (2007). Nous allons développer les étapes de cette analyse qui concerne les scores des quatre indices standards (ICV, IRP, IMT et IVT).

a) Pour chaque enfant, un indice moyen (IM) est calculé à partir de la moyenne des notes composites aux quatre indices.

$$IM = \frac{ICV + IRP + IMT + IVT}{4}$$

b) On soustrait ensuite chaque indice de l'indice moyen afin d'obtenir un écart à la moyenne.

$$|ICV - IM| ; |IRP - IM| ; |IMT - IM| ; |IVT - IM|$$

La différence indice – indice moyen est exprimée en valeur absolue.

c) Chaque écart à la moyenne est comparé à la taille de la différence requise pour être statistiquement significative au seuil de 5 % calculée sur l'échantillon d'étalonnage français d'après la formule de Davis (1959).

Les différences absolues |ICV – IM| sont comparées à la valeur critique de 11.07 ; les différences absolues |IRP – IM| sont comparées à la valeur critique de 11.44 ; les différences absolues |IMT – IM| sont comparées à la valeur critique de 11.34 ; et les différences absolues |IVT – IM| sont comparées à la valeur critique de 12.56.

Il s'agira d'une force personnelle (FoP), si la différence absolue entre l'indice examiné et l'indice moyen est égale ou supérieure à la valeur critique correspondant à l'indice et si l'indice est supérieur à l'indice moyen. Par exemple, si un enfant a un indice moyen de 82 et une note composite de 95 à l'ICV, la différence absolue |ICV – IM|

équivaut à 13. Cette valeur dépassant la valeur critique de 11.07 et l'ICV étant supérieur à l'IM, on identifie une force personnelle en compréhension verbale. Suivant une logique analogue, il s'agira d'une faiblesse personnelle (FaP) si la différence absolue entre l'indice examiné et l'indice moyen est égale ou supérieure à la valeur critique correspondant à l'indice et si l'indice est inférieur à l'indice moyen. Par exemple, si un enfant a un indice moyen de 124 et une note composite de 108 à l'IMT, la différence absolue |IMT – IM| équivaut à 16. Cette valeur étant supérieure à la valeur critique de 11.34 et l'IMT étant inférieur à l'IM, on identifie une faiblesse personnelle en mémoire de travail. L'identification de forces ou de faiblesses personnelles découle de ce qu'on appelle des indices déviants. Pour finir, si une différence entre l'indice examiné et l'indice moyen n'est pas égale ou supérieure à la valeur critique, cette différence situe alors l'indice dans les limites de la moyenne des performances des quatre indices ; il ne s'agit ni d'une force personnelle ni d'une faiblesse personnelle, mais d'une performance dans la moyenne personnelle (MoP).

RÉSULTATS

7. ÉTUDE 1 : LE FONCTIONNEMENT DIFFÉRENTIEL DES ITEMS DU WISC-IV

Dans ce premier volet consacré à l'étude 1, nous allons présenter les résultats de l'effet direct (β) et les proportions des variances expliquées pour les variables âge, sexe, et statut socio-économique, ainsi que développer les résultats de l'évaluation du fonctionnement différentiel des items (DIF). Pour rappel, les paramètres des items des subtests Identification de concepts, Matrices, et Complètement d'images sont estimés à l'aide du modèle 2 PL de réponses à l'item, tandis que les paramètres des items des subtests Cubes, Similitudes, Vocabulaire et Compréhension sont estimés à l'aide du modèle gradué de Samejima. Quant aux items des subtests des indices IMT et IVT, ils ne peuvent pas être modélisés. Toutes les analyses de cette étude 1 ont été réalisées avec le programme M*plus* 7.2 (Muthén & Muthén).

7.1. ÉVALUATION DE L'UNIDIMENSIONNALITÉ DES ITEMS

Avant de pouvoir appliquer un modèle de réponse à l'item (MRI) sur un subtest, on procède d'abord à l'évaluation statistique de son unidimensionnalité. Dans un modèle unidimensionnel, le trait latent évalué par le test prédit à lui seul la performance sur chacun des items (voir Figure 44, p. 204). Autrement dit, la réussite (ou l'échec) aux différents items d'un subtest devrait uniquement dépendre du niveau du sujet sur la variable latente évaluée par le subtest. Par exemple, pour le subtest Vocabulaire, on suppose que la dimension psychologique évaluée est de la compréhension verbale ; le niveau du sujet sur d'autres propriétés mentales ne devrait pas influencer la réussite (ou l'échec) des items de ce subtest.

Pour évaluer l'unidimensionnalité des items d'un subtest, on spécifie un modèle réflectif avec 1 variable latente qui explique les résultats sur chacune des variables manifestes comme illustré dans la Figure 44 (p. 204). Si nous obtenons un bon ajustement entre le modèle unidimensionnel et nos données, nous acceptons l'hypothèse de l'unidimensionnalité. Les critères pour évaluer l'ajustement des données au modèle unidimensionnel se réfèrent aux indices d'ajustement (*fit indices*) et suivent les recommandations d'experts en la matière (Hooper, Coughlan, & Mullen, 2008; Hu & Bentler, 1999; Yu, 2002). Les deux critères retenus sont : RMSEA < .06 et CFI > .95. Le

critère du chi2 non significatif (p > .05) est difficilement observable sur de grands échantillons, nous ne le prenons pas en compte. Il sera néanmoins rapporté à titre indicatif.

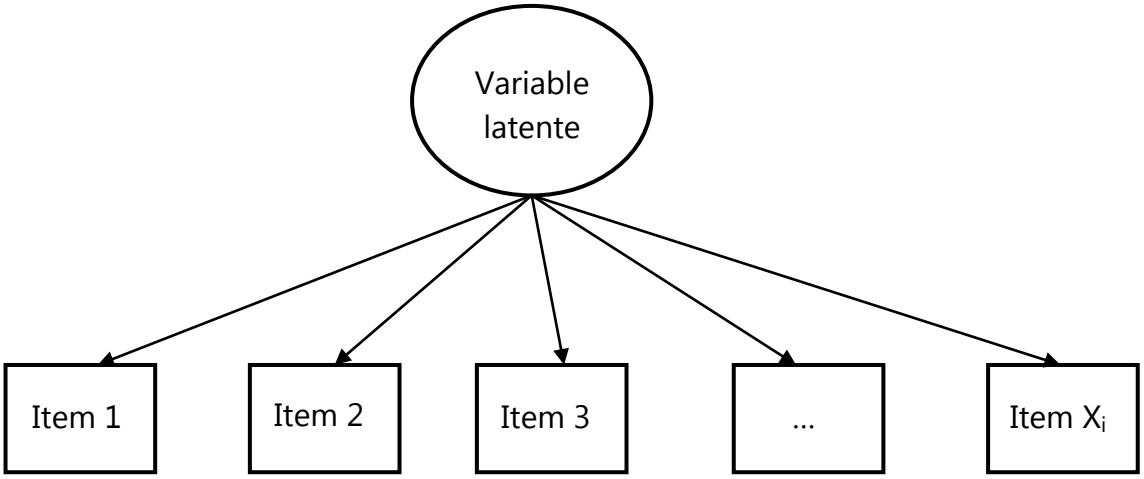


Figure 44. Modèle unidimensionnel.

Deux difficultés sont rencontrées dans cette modélisation. La première concerne les items de variance nulle (réussis ou échoués par tout le monde). Que ce soit dans un modèle 2 PL ou dans un modèle gradué de Samejima, les items de variance nulle empêchent l'estimation d'aboutir et doivent donc être éliminés de l'analyse. La seconde difficulté concerne les items de variance quasi nulle (réussis ou échoués par une large majorité) qui perturbent également l'estimation des paramètres en conduisant à des problèmes de non-respect du postulat d'indépendance locale entre les items. S'il y a des intercorrélations entre paires d'items, il n'y a pas d'indépendance locale. Or, un modèle réflectif postule l'indépendance locale comme condition d'application. Des intercorrélations entre paires d'items signifient que les paires d'items partagent entre eux quelque chose d'autre en commun en plus du trait latent, qui, dans un modèle réflectif, doit pourtant expliquer à lui seul la performance sur tous les items du test. Nous devons donc également éliminer les items de variance quasi nulle qui ne respectent pas le critère d'indépendance locale du modèle réflectif dans lequel s'inscrivent les modèles de réponses à l'item. Les deux difficultés que nous venons de soulever sont d'autant plus susceptibles de survenir quand la taille de l'échantillon n'est pas assez conséquente pour fournir des informations qui stabilisent l'estimation. Par ailleurs, plus le modèle est complexe – c'est-à-dire, plus il y a de paramètres librement estimés –, plus l'estimation du modèle nécessite un échantillon conséquent. Par

exemple, un modèle de réponse à l'item à deux paramètres (modèle 2 PL) est plus complexe qu'un modèle à un paramètre (modèle de Rasch). De même, un modèle de réponse à l'item appliqué à des items à cotation polytomique est plus complexe à modéliser qu'un modèle de réponse à l'item appliqué à des items à cotation dichotomique. En effet, dans le cas d'items à cotation polytomique, l'estimation doit tenir compte de la graduation des points. Par exemple dans le subtest Similitudes, obtenir deux points à un item révèle une habileté supérieure par rapport à l'obtention d'un seul point, et encore plus supérieure par rapport à l'obtention de zéro point. Pour aider l'estimation des paramètres, l'échantillon doit comporter des individus sur tout le continuum de l'échelle d'habileté du trait latent. Ce qui est d'autant plus probable si la taille d'échantillon est très grande. Le Tableau 9 renseigne sur les indices d'ajustement des données au modèle unidimensionnel à 2 PL ou au modèle unidimensionnel gradué de Samejima. Le Tableau 9 précise aussi les items impliqués dans l'analyse après qu'on retire les items de variances nulles ou quasi nulles qui ne satisfassent pas au critère d'indépendance locale. Il va de soi que l'élimination des items problématiques améliore l'ajustement des données au modèle.

Tableau 9

Indices d'ajustement entre les données de chaque subtest et le modèle unidimensionnel

	Items gardés	ddl	Chi2	RMSEA	CFI
Subtest à cotation dichotomique					
Identification de concepts	7 – 28	44	$p < .001$	**.021**	.930
Matrices	7 – 33	54	$p < .001$	**.035**	**.955**
Complètement d'images	3 – 38	72	$p < .001$	**.018**	.935
Subtests à cotation polytomique					
Cubes	8 – 14	32	**.533**	**.000**	**1.000**
Similitudes	3 – 7 et 19-21	53	$p < .001$	**.033**	**.980**
Vocabulaire	7 – 28	66	$p < .050$	**.020**	**.988**
Compréhension	3 – 19	51	$p < .001$	**.032**	**.977**

Les critères d'ajustement (RMSEA < .06 et CFI > .95) montrent un ajustement des donnés au modèle unidimensionnel pour tous les subtests, sauf Identification de concepts et Complètement d'images dont le CFI ne dépasse pas le seuil requis de .95. On ne peut donc pas postuler qu'un seul trait latent est derrière les items d'Identification de concepts et de Complètement d'images. Nous retirons ces deux subtests de la suite des analyses.

7.2. DIFFÉRENCES SELON L'ÂGE, LE SEXE ET LE STATUT SOCIO-ÉCONOMIQUE DES PARENTS

Une fois la condition d'unidimensionnalité satisfaite, nous pouvons introduire les covariés dans le modèle unidimensionnel de base pour l'analyse du fonctionnement différentiel de l'item (DIF). Simultanément à l'analyse du DIF, la proportion de variance expliquée par un covarié est également fournie dans la sortie des résultats du logiciel M*plus*. Pour une meilleure lisibilité des résultats, nous allons dans un premier temps présenter les résultats des effets directs (β) et des variances expliquées en fonction des covariés considérés. Dans un second temps, nous présenterons les résultats de l'analyse du DIF.

Le modèle unidimensionnel 2 PL ou le modèle de Samejima de la Figure 44 (p. 204) est modifié avec l'introduction d'un covarié. Le premier covarié est la variable âge qui est définie en nombre de mois (voir Figure 45). Ce premier modèle estimé permet d'évaluer dans quelle proportion le trait latent évalué par chaque subtest peut être expliqué par l'âge des participants. N'oublions pas que, dans les MRI, nous travaillons uniquement sur les scores bruts obtenus aux items. L'échantillon ayant une étendue d'âges de 7 à 12 ans, il est nécessaire d'estimer l'effet direct de l'âge. En grandissant, les enfants acquièrent toujours plus de connaissances et d'expériences. Pour un trait comme l'intelligence, on suppose une stabilité dans la position des individus par rapport à son groupe de référence, ce qui ne signifie pas une stagnation du niveau absolu d'intelligence de l'individu. En moyenne, plus les enfants sont âgés, plus ils réussissent d'items, et donc obtiennent un score brut plus élevé que des enfants plus jeunes qu'eux. Le score brut n'est donc pas interprétable et ne permet pas de comparaisons lorsque le test s'adresse à une étendue d'âges. C'est pour cela qu'il est converti en score standardisé de moyenne et d'écart type connus afin de permettre les comparaisons de performances entre différents âges et entre différents subtests. Or, dans les MRI, on perd la standardisation des scores étant donné qu'on travaille sur les points obtenus à chaque item. Il est donc important de déterminer l'effet direct de l'âge sur la performance aux items des subtests du WISC-IV.

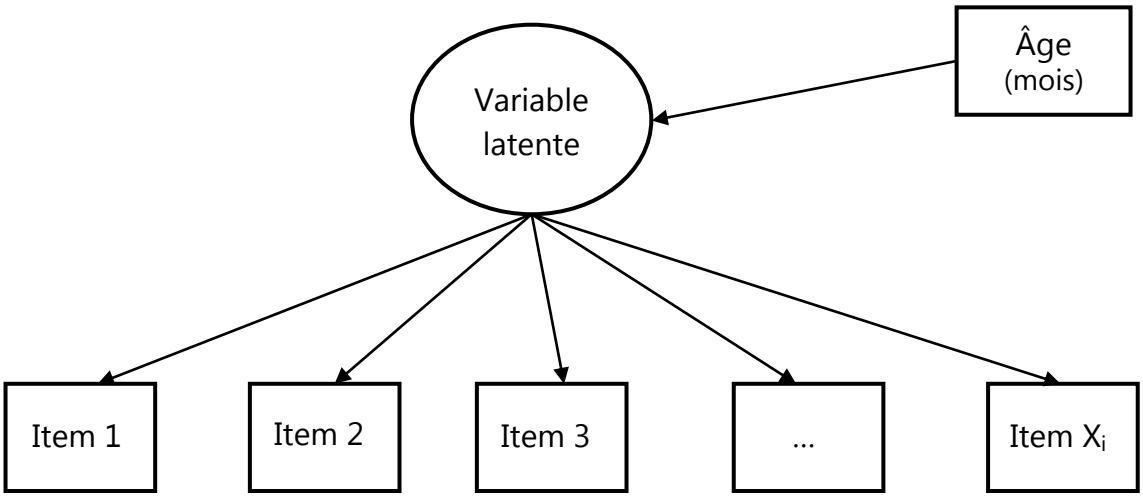


Figure 45. Modèle unidimensionnel avec la variable âge en covarié

Pour tous les subtests, les résultats montrent une influence directe de la variable âge sur le score latent (voir Tableau 10, p. 208). Comme attendu, la variable âge est positivement liée à la performance sur les items (réussite ou échec). Plus un enfant est âgé, plus son score brut est élevé. Les effets directs de l'âge sont plus importants pour les subtests de l'ICV que de l'IRP.

Le Tableau 10 renseigne également sur les proportions de la variance expliquée associées aux effets directs de l'âge. La variance expliquée par l'âge varie de 24.60 % (CUB) à 45.83 % (COM). L'âge explique une part importante de la variance totale pour les subtests qui font appel au verbal. Les subtests de l'ICV mobilisent des connaissances acquises au cours des apprentissages, notamment scolaires. Dans un parcours scolaire ordinaire, les enfants sont plus ou moins uniformément exposés à des apprentissages relativement similaires pour le vocabulaire et la lecture à l'école. En revanche, les habiletés que mobilisent les subtests de l'IRP sont moins rattachées à des activités scolaires, ce qui peut expliquer une moindre influence de l'âge. Étant donné la grande part de variance totale qui est expliquée par la variable âge et pour réintroduire une standardisation des scores, l'âge doit désormais être contrôlé dans tous les modèles estimés. Ainsi, dans la suite des analyses, nous introduisons tour à tour comme covarié supplémentaire les variables sexe de l'enfant, puis le statut socio-économique des parents. À chaque fois, la variable âge est contrôlée en étant également en covarié.

Tableau 10

Effet direct et variance expliquée par l'âge, le sexe et le statut socio-économique

	Âge		Sexe		SES	
	Effet direct (β)	Variance expliquée (%)	Effet direct (β)	Variance expliquée (%)	Effet direct (β)	Variance expliquée (%)
CUB	.496**	24.600	-.120**	1.440	-.186**	3.460
SIM	.631**	39.820	-.023	-	-.284**	8.070
VOC	.667**	44.890	-.002	-	-.289**	8.350
MAT	.534**	28.520	.093*	0.860	-.157**	2.460
COM	.677**	45.830	-.070	-	-.218**	4.750

Note. SES = statut socio-économique ; CUB = Cubes ; SIM = Similitudes ; VOC = Vocabulaire ; MAT = Matrices ; COM = Compréhension.

* $p < .05$, ** $p < .001$.

Dans un deuxième modèle estimé, nous introduisons le covarié sexe, avec l'âge en covarié (voir Figure 46). Les filles sont codées 1 et les garçons sont codés 0. Dans l'épreuve Cubes, il y a une très faible influence de la variable sexe en faveur des garçons (β = -.120 ; voir Tableau 10, p. 208). L'âge contrôlé, le sexe de l'enfant explique 1.44 % de la variance totale pour Cubes. En revanche, une très faible influence de la variable sexe s'observe en faveur des filles dans les Matrices (β = .093). L'âge contrôlé, le sexe de l'enfant explique 0.86 % de la variance totale pour Matrices. Il n'y a pas d'effet du sexe pour aucun des trois subtests de l'ICV.

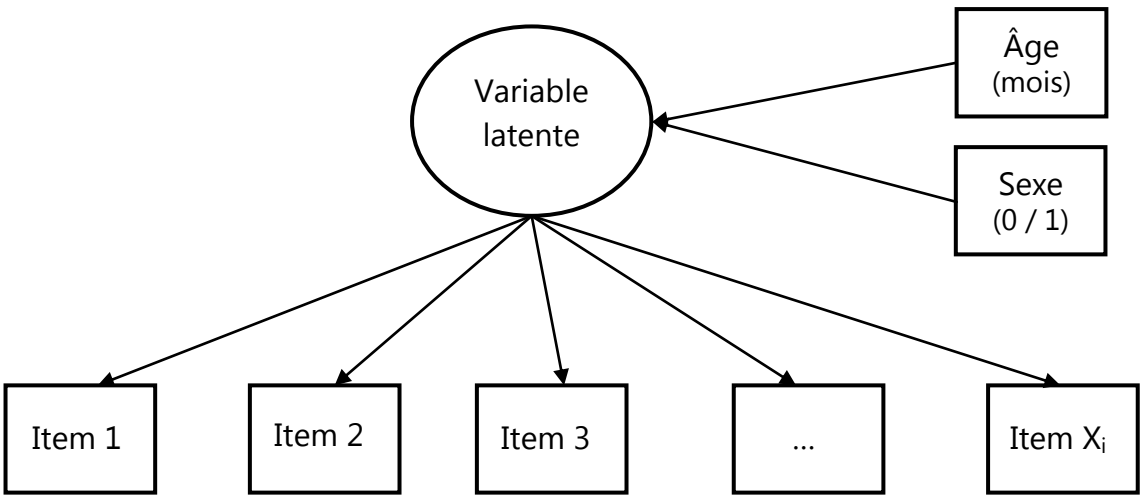


Figure 46. Modèle unidimensionnel avec les variables âge et sexe en covarié

Dans un troisième modèle estimé, nous introduisons le covarié statut économique des parents, avec toujours l'âge en covarié (voir Figure 47, p. 209). La variable statut socio-économique des parents (SES) est construite à partir de deux indicateurs : profession du père et profession de la mère. La situation professionnelle de chaque parent est assignée à une des dix catégories inspirées de la Classification Internationale Type des Professions réactualisée en 2008 (CITP-08 ; voir Annexe C). Dans cette grille, plus la profession demande de longues et de hautes études, plus la valeur du codage de la catégorie est basse. Par exemple, un médecin est catégorisé 2, tandis qu'un maçon est catégorisé 7. Le modèle testé est illustré par la Figure 47 (p. 209).

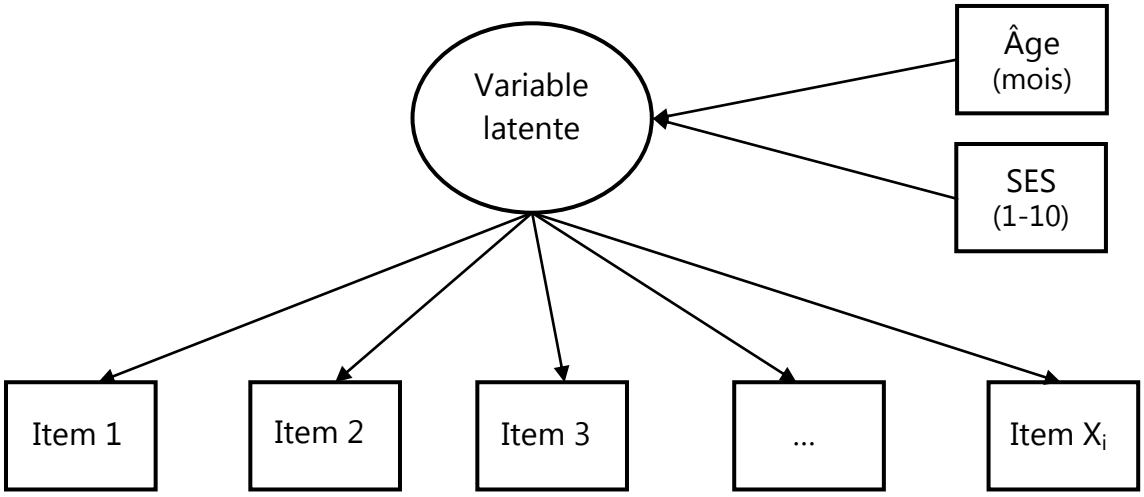


Figure 47. Modèle unidimensionnel avec les variables âge et SES en covarié

Les résultats montrent une influence de la variable SES sur les scores latents pour tous les subtests (voir Tableau 10, p. 208). L'effet direct négatif de la variable SES suggère que les enfants ayant des parents avec une profession socialement valorisée ont des performances plus élevées que ceux ayant des parents avec une profession socialement moins valorisée. Il y a un effet du statut socio-économique pour tous les subtests.

L'effet de l'âge contrôlé, la variance expliquée par la variable SES varie de 2.46 % (MAT) à 8.35 % (VOC). Conformément à la littérature sur le sujet, on peut noter un effet du statut socio-économique plus important sur les performances des subtests qui font appel au verbal.

D'autres analyses nous permettent d'examiner de façon séparée la contribution de la profession de chacun des deux parents. Sur le plan descriptif, l'évaluation de la contribution de la profession de chaque parent montre dans l'ensemble un effet direct différencié entre la profession de la mère et la profession du père sur les performances aux subtests (voir Tableau 11, p. 210). Par exemple, pour Vocabulaire, l'effet direct de la profession de la mère a une influence distincte et statistiquement significative sur la probabilité de réussite sur les items de ce subtest. Plus la mère a une profession socialement valorisée, plus l'enfant obtient une performance élevée à Vocabulaire. Toutefois, l'influence propre à la profession de la mère est très petite (β = -.189). De même, toujours pour Vocabulaire, l'effet direct de la profession du père a une influence distincte et statistiquement significative, même si petite (β = -.169), sur la probabilité de réussite sur les items de ce subtest.

Sachant que la contribution de la profession des parents est différenciée, est-ce la taille de la différence entre les deux contributions est significative ? Pour répondre à cela, le test de Wald (1943) est appliqué. Les résultats du test de Wald ne révèlent aucune taille de différence statistiquement significative entre la contribution de la profession d'un parent par rapport à celle de l'autre parent (voir Tableau 11, p. 210). Ainsi, même si la contribution de la profession de chaque parent influence distinctement sur la performance des items, cette différence d'influence n'est pas assez importante pour avoir une implication clinique. La variance expliquée varie de 1.10 % (MAT) à 3.53 % (VOC) pour la profession des mères, et de 0.81 % (MAT) à 3.31 % (SIM) pour la profession des pères.

Tableau 11

Effets de la profession de chacun des deux parents et test de Wald

	Mère		Père		Test de Wald (*p*)
	Effet direct (β)	Variance expliquée (%)	Effet direct (β)	Variance expliquée (%)	
CUB	-.112*	1.250	-.110*	1.210	.860
SIM	-.188**	3.530	-.182**	3.310	.792
VOC	-.189**	3.570	-.169**	2.860	.960
MAT	-.105*	1.100	-.090*	0.810	.974
COM	-.174**	3.030	-.108*	1.170	.459

Note. CUB = Cubes ; SIM = Similitudes ; VOC = Vocabulaire ; MAT = Matrices ; COM = Compréhension.

* $p < .05$, ** $p < .001$.

Les résultats des analyses sur les effets directs de l'âge, du sexe et du SES confirment la littérature. En effet, il est bien connu que les enfants issus de milieux socio-économiques élevés se distinguent sur les performances verbales. Toutefois, ces différences de groupes ne traduisent pas forcément un biais de l'item dans un sens psychométrique. Pour parler de biais de l'item ou de fonctionnement différentiel de l'item (DIF), il faut que, pour une même habileté sur le trait latent, deux enfants provenant de deux groupes différents n'aient pas la même probabilité de réussir un même item. À la suite, nous passons donc aux résultats de l'analyse du fonctionnement différentiel des items.

7.3. FONCTIONNEMENT DIFFÉRENTIEL DES ITEMS DU WISC-IV

La détection d'un biais d'items consiste à tester un éventuel fonctionnement différentiel des items (FDI). Pour l'équité dans l'évaluation, il est important qu'un item se comporte de la même manière avec tous les individus ayant une même habileté sur le trait latent évalué. Un item qui montre un fonctionnement différentiel au-delà d'un certain seuil est un item biaisé. La présence d'items biaisés dans un test à l'encontre d'un groupe distinct d'individus compromet l'équivalence des scores entre groupes, et donc la validité de l'interprétation du test. Ce n'est plus uniquement des différences sur le trait latent qu'un le test biaisé met en évidence, ce qui rend malaisé une interprétation des scores au test.

Pour la détection d'un FDI avec le logiciel M*plus*, on examine des indices issus de la procédure d'estimation – appelés indices de modifications – qui signalent s'il y a un lien indirect entre un item et le covarié examiné (c.-à-d. âge, sexe ou SES). Nous cherchons à mettre en évidence d'éventuels liens indirects pour les trois modèles estimés suivants : (1) entre les items de chaque subtest et l'âge, (2) entre les items de chaque subtest et âge + sexe, (3) entre les items de chaque subtest et âge + statut socio-économique. Comme nous l'avons précédemment dit, les analyses portent sur les scores bruts aux items et l'échantillon comporte une étendue d'âge de 7 à 12 ans, aussi devons-nous à chaque fois contrôler l'âge. Pour chacun des modèles estimés, l'examen des indices de modification permet la détection d'un fonctionnement différentiel des items (c.-à-d. un effet indirect).

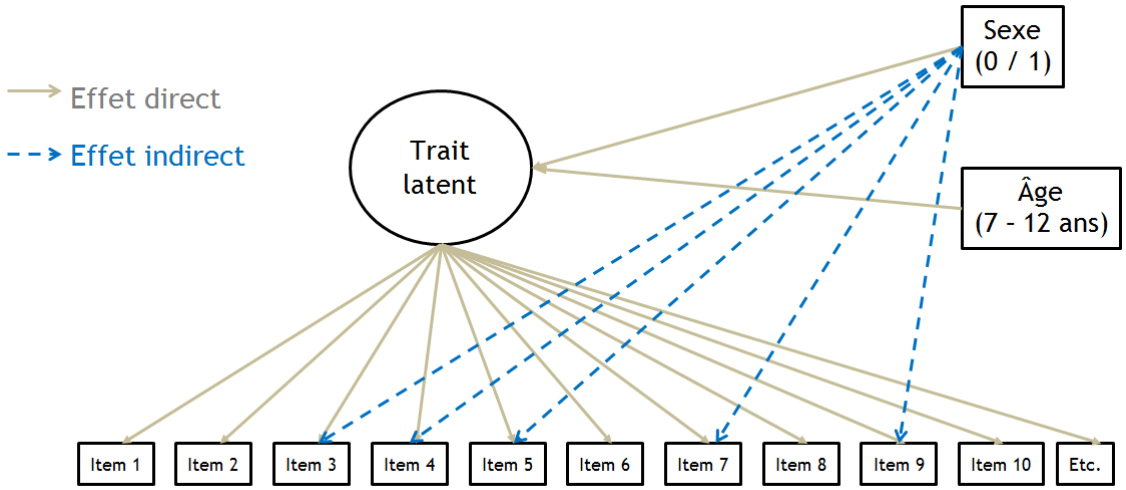


Figure 48. Représentation de la présence d'un fonctionnement différentiel entre les filles et les garçons pour les items 3, 4, 5, 7 et 9 d'un subtest.

Si un lien indirect entre un covarié et un item est suggéré, cela signifie que l'item en question fonctionne différentiellement selon la modalité du covarié examiné. Pour un exemple fictif, référons-nous au schéma de la Figure 48 qui modélise la situation pour un subtest. Cet exemple montre qu'une fois l'effet direct (trait plein) de l'âge et du sexe contrôlé, il y a encore un effet indirect (trait en pointillé) du sexe sur les items 3, 4, 5, 7 et 9. Cet effet indirect signale un FDI. Les filles et les garçons qui ont la même habileté sur le trait latent évalué, n'ont pas la même probabilité de réussir les items 3, 4, 5, 7 et 9 du subtest. L'examen comparatif des courbes caractéristiques de l'item (CCI) des filles et des garçons permet d'interpréter le sens du fonctionnement différentiel pour les items 3, 4, 5, 7 et 9, à savoir si le biais d'item est en faveur des filles ou des garçons. Si les indices de modifications ne signalent pas de lien indirect entre un item et le covarié, l'item se comporte de la même manière pour tous les individus quelle que soit leur groupe d'appartenance. Dans notre exemple inventé, tous les individus (filles et garçons) ont la même probabilité de réussir les items 1, 2, 6, 8 et 10. En l'absence de DIF, on suppose que la variable latente de la dimension évaluée par le subtest explique à elle seule les variations dans la probabilité de réussir l'item.

Nous ne développerons pas sur les aspects techniques du calcul des indices de modifications. De façon conceptuelle, les indices de modifications révèlent l'écart entre les courbes caractéristiques de l'item (CCI) des groupes considérés qui se distinguent sur un covarié. Par exemple, si le covarié est le sexe, l'analyse compare la CCI des filles

et la CCI des garçons pour chaque item d'un subtest. Si l'aire entre la CCI des filles et la CCI des garçons est supérieure à un seuil défini, on considère que l'item présente un fonctionnement différentiel. Dans la situation d'une présence de DIF, cela signifie que pour une même habileté sur le trait latent, il n'y a pas chez les filles et chez les garçons la même probabilité de réussir l'item.

Les résultats de nos analyses vont dans le sens d'une absence de biais, puisque les indices de modifications ne révèlent aucun fonctionnement différentiel de l'item pour les subtests analysés. En effet, une fois contrôlée l'influence directe des variables âge, sexe ou statut socio-économique, il n'y a pas de différence dans les probabilités de réussir l'item pour une même habileté sur le trait latent évalué. Dit autrement, une fois l'influence directe des variables âge, sexe ou statut socio-économique contrôlée, les capacités dans la dimension psychologique évaluée par un subtest prédisent à elles seules la performance (la réussite ou l'échec) sur les items de Cubes, Similitudes, Vocabulaire, Matrices et Compréhension. Il n'y a donc pas d'effets indirects sur les items liés aux variables contrôlées âge, sexe et statut socio-économique.

8. ÉTUDE 2 : LA STABILITÉ DU WISC-IV

Dans ce second volet seront développés les résultats relatifs aux différentes évaluations de la stabilité à long terme pour les scores standards et CHC du WISC-IV. L'évaluation de la stabilité s'est portée aussi bien sur le niveau interindividuel (groupal) que le niveau intra-individuel (individuel). Rappelons que les résultats sont basés sur les données d'un échantillon de 277 enfants tout-venant (c.-à-d. non consultants). Cependant, les analyses sur le subtest Complètement d'images et le facteur Gv comportent 29 données manquantes et sont donc réalisées sur 248 enfants.

8.1. STATISTIQUES DESCRIPTIVES

Pour l'étude de la stabilité à long terme du WISC-IV, les résultats des indices standards et CHC sont rapportés ainsi que ceux des subtests. Tout d'abord, le Tableau 12 (p. 220) renseigne sur les moyennes, les écarts types et les différences de moyennes. Les notes composites et les notes des subtests sont proches des valeurs théoriques (c.-à-d. moyenne de 100 ou 10 et écart type de 15 ou 3). Plus précisément, les moyennes des indices standards varient de 95.18 (IMT) à 104.90 (ICV) pour la première passation, et de 97.20 (IMT) à 107.17 (IVT) pour la seconde passation. En moyenne, pour les deux passations, les enfants de notre échantillon obtiennent leurs performances les plus basses sur l'Indice de Mémoire de Travail. La moyenne du QIT est de 100.81 au test, et de 103.34 au retest. Pour les indices CHC, l'étendue des moyennes au test varie de 95.17 (Gwm) à 105.49 (Gc), et de 97.18 (Gwm) à 106.88 (Gs) au retest. Quant aux subtests, la moyenne des performances s'étend de 9.03 (Séquence Lettres-Chiffres) à 11.13 (Similitudes) pour la première passation, et de 9.38 (Séquence Lettres-Chiffres) à 11.80 (Symboles) pour la seconde passation. Ces moyennes et ces écarts types proches des valeurs théoriques suggèrent que notre échantillon est relativement représentatif des données de standardisation, aussi bien au niveau des performances moyennes que de la variabilité intragroupe.

Dans le Tableau 12 (p. 220) figure la taille des différences entre les deux passations (*d* de Cohen). Cette dernière permet d'estimer l'importance de la différence de performances entre les deux passations. Selon Cohen (1977), elle est considérée

comme négligeable pour une valeur de d < 0.2, petite pour un d entre ≥ 0.2 et < 0.5, modérée pour un d entre ≥ 0.5 et < 0.8 et grande pour un d ≥ 0.8.

8.2. DURÉE DE L'INTERVALLE TEST-RETEST, ÂGE ET DIFFÉRENCE DE PERFORMANCES

Étant donné que nous n'avons pas contrôlé la durée des intervalles test-retest pour appliquer la même longueur de temps à tous les enfants, la variable durée des intervalles test-retest n'est pas une constante qui rendrait impossible le calcul d'une corrélation. Grâce à la variabilité dans les intervalles test-retest au sein des enfants, nous pouvons distinguer l'effet de la durée de l'intervalle (ou l'effet retest) et l'effet de l'âge. Pour relever la part des contributions du retest et de l'âge, des corrélations sont réalisées dont les résultats sont présentés à la suite.

Dans un premier temps, nous examinons la corrélation entre la différence de performances entre T1 et T2 pour un score ($\Delta_{indice} = I_2 - I_1$) et la différence de durée test-retest ($\Delta_T = T2 - T1$). Pour les indices standard, les corrélations sont négatives pour tous les scores et significatives seulement pour le QIT et l'IAG ($\Delta_{QIT} \times \Delta_T = -.16$; $\Delta_{IAG} \times \Delta_T = -.15$, voir Figure 49). Pour ces derniers, les résultats suggèrent une légère tendance à des gains de performances entre T1 et T2 d'autant plus important que l'intervalle test-retest est court. Plus l'intervalle test-retest est long et moins les enfants ont tendance à tirer bénéfice de l'expérience d'une première passation. Toutefois, cette tendance est trop légère pour avoir une implication clinique (r^2 = 2.56 % pour le QIT et r^2 = 2.25 % pour l'IAG).

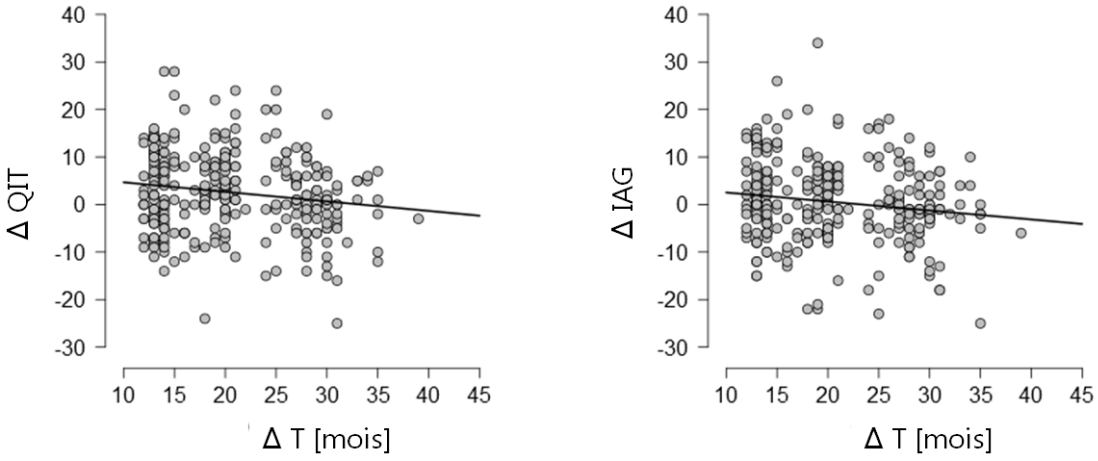


Figure 49. Corrélation entre le delta performances et le delta durée test-retest.

Pour les indices CHC, des corrélations significatives et négatives sont observées pour Gc, Gf et Gv ($\Delta_{Gc} \times \Delta_T = -.12$; $\Delta_{Gf} \times \Delta_T = -.13$, $\Delta_{Gv} \times \Delta_T = -.16$). Quant aux subtests, des corrélations significatives et négatives sont observées pour Cubes, et Similitudes ($\Delta_{CUB} \times \Delta_T = -.12$; $\Delta_{SIM} \times \Delta_T = -.20$). À nouveau, il s'agit de tendances trop faibles pour avoir une implication clinique.

Dans un deuxième temps, nous examinons la corrélation entre la différence de performances entre T1 et T2 pour un score ($\Delta_{indice} = I_2 - I_1$) et l'âge de l'enfant à la première passation (A_1). Les corrélations sont significatives et positives pour les indices IRP, IVT, ICC, QIT et Gs ($\Delta_{IRP} \times A_1 = .13$; $\Delta_{IVT} \times A_1 = .17$, $\Delta_{ICC} \times A_1 = .20$, $\Delta_{QIT} \times A_1 = .19$; $\Delta_{Gs} \times A_1 = .16$, voir Figure 50). Ces résultats suggèrent que les enfants les plus âgés à la passation initiale ont une légère tendance à tirer davantage bénéficie de l'expérience d'une première passation que les plus jeunes à la première passation. Pour les subtests, les corrélations sont significatives et positives pour Code, Matrices et Compréhension ($\Delta_{COD} \times A_1 = .14$; $\Delta_{MAT} \times A_1 = .17$, $\Delta_{COM} \times A_1 = .14$). Il s'agit tous de tendances trop légères pour que nous puissions tirer une conséquence au niveau clinique.

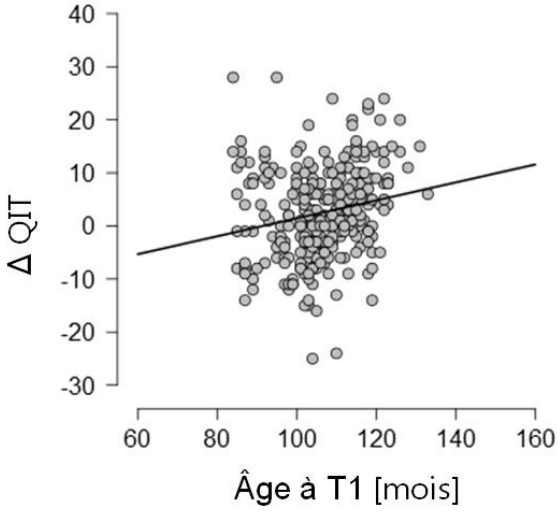


Figure 50. Corrélation entre le delta performances et l'âge à la première passation.

Pour ces deux analyses présentées, il est à préciser que nos données ne permettent pas parfaitement de distinguer entre les effets de l'âge et du retest. En effet, la corrélation entre l'âge des enfants à la première passation et la durée de l'intervalle est significative (r = .26), indiquant que les enfants les plus âgés à T1 ont une légère tendance à avoir un plus long intervalle test-retest. Dans la situation où on aurait pu contrôler les intervalles test-retest, la corrélation aurait été nulle. Pour examiner l'influence de cette faible corrélation (r^2= 6.76 %), nous réalisons des corrélations partielles qui contrôlent pour l'âge ou pour la durée de l'intervalle. Les résultats vont dans le même sens que ceux présentés. Lorsque nous contrôlons pour l'âge, les corrélations partielles entre la différence de performances entre T1 et T2 ($\Delta_{indice} = I_2 - I_1$) et la différence de durée test-retest ($\Delta_T = T2 - T1$) sont significatives et négatives pour les indices ICV, IRP, IVT, QIT, IAG, ICC et Gv (variant de -.22 pour Δ_{FSIQ} x Δ_T à -.13 pour Δ_{VCI} x Δ_T). Lorsque nous contrôlons pour la durée du retest, les corrélations partielles entre la différence de performances entre T1 et T2 ($\Delta_{indice} = I_2 - I_1$) et l'âge de l'enfant à la première passation (A_1) sont significatives et positives pour les indices IRP, IMT, IVT, QIT, IAG, ICC, Gf, Gv et Gs (variant de .12 pour $\Delta_{IMT} \times A_1$ à .25 pour $\Delta_{QIT} \times A_1$). Tous les coefficients sont trop faibles pour une implication au niveau clinique.

Dans un troisième temps, nous examinons la corrélation entre la différence de performances entre T1 et T2 pour un score ($\Delta_{indice} = I_2 - I_1$) et la performance du score à la passation initiale (I_1). Pour tous les scores, les corrélations sont significatives et négatives. Pour les indices standards, les coefficients de corrélation varient de -.32 pour Δ_{ICV} x ICV_1 à -.45 pour Δ_{IRP} x IRP_1. Pour les indices CHC, ils varient de -.31 pour Δ_{Gv} x Gv_1 à -.50 pour Δ_{Gf} x Gf_1. Et pour les subtests, ils varient de -.35 pour Δ_{CUB} x CUB_1 à -.62 pour Δ_{IDC} x IDC_1. Tous ces résultats suggèrent que, plus les enfants obtiennent un score élevé à la première passation, plus la différence de performances entre les deux passations est petite. La Figure 51 (p. 218) illustre le cas du QI Total (r = -.44). Les enfants avec les plus bas QIT à la première passation tendent à tirer davantage bénéfice de cette première expérience de passation que les enfants avec les QIT les plus hauts.

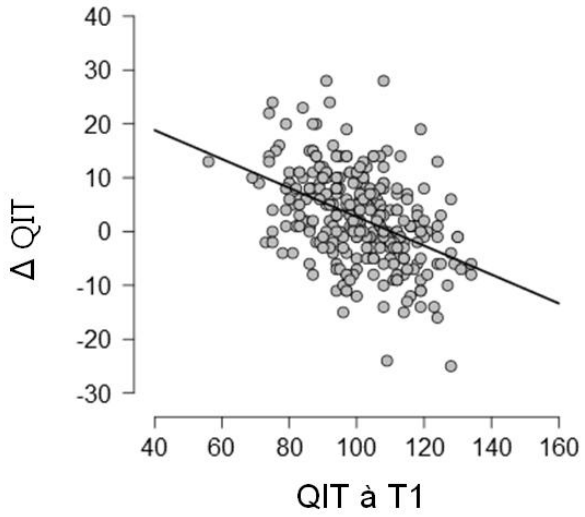


Figure 51. Corrélation entre le delta performances et la performance à la première passation.

Dans un quatrième et dernier temps, nous examinons la corrélation entre la différence de performances pour un indice ($\Delta_{indice\ A} = I_{A2} - I_{A1}$) et la différence de performances pour un autre indice ($\Delta_{indice\ B} = I_{B2} - I_{B1}$). Il s'agit de savoir si le changement entre les deux passations dans un indice est lié au changement dans un autre indice. Pour les pairs d'indices standards, les résultats montrent que seuls les changements dans l'ICV (c.-à-d. Δ_{ICV}) corrèlent significativement avec les changements dans les trois autres indices. Les corrélations sont néanmoins très peu élevées (Δ_{ICV} x Δ_{IRP} = .14 ; Δ_{ICV} x Δ_{IMT} = .16 ; Δ_{ICV} x Δ_{IVT} = .15). Pour les pairs d'indices CHC, il y a des corrélations significatives et positives pour Gc et Gsm (Δ_{Gc} x Δ_{Gsm} = .17), Gc et Gs (Δ_{Gc} x Δ_{Gs} = .16), Gf et Gv (Δ_{Gf} x Δ_{Gv} = .17). Toutes les corrélations montrent des tendances légères qui n'ont pas de conséquences cliniques.

8.3. STABILITÉ ABSOLUE

Afin de déterminer si, au niveau du groupe, les moyennes des différents scores restent stables d'une passation à l'autre, des *t*-tests pour mesures appariées sont réalisés. Pour contrôler le seuil d'erreur α lors de comparaisons multiples, une correction Holm-Bonferroni est appliquée (Aickin & Gensler, 1996; Holm, 1979). En effet, lorsque plusieurs tests statistiques sont réalisés simultanément le risque global d'erreur

de première espèce (ou erreur α) s'accroît à chaque répétition de test[26]. Comme en répétant les tests statistiques, le risque global d'erreur n'est plus au seuil de 5 % – mais bien supérieur –, *in fine* on risque de conclure à tort à un résultat significatif. Moins conservatrice que la correction de Bonferroni qui rejette rarement H_0 lorsque le nombre de comparaisons est grand, la correction Holm-Bonferroni est préférée pour tenir le taux de faux positifs (c.-à-d. conclure à tort à une différence significative) au seuil fixé de 5 %. Après correction, les différences de moyennes associées à une *p*-valeur de < .05 sont indiqués en gras dans le Tableau 12 (p. 220).

Dans le Tableau 12 (p. 220), les comparaisons de moyennes montrent une augmentation significative de la première à la seconde passation pour les indices IMT, IVT, QIT, ICC, Gf, Gwm et Gs, avec des gains allant de +2.01 points (Gwm) à +5.79 points (IVT). Sur l'ensemble des notes composites, les tailles de la différence s'étendent de -0.02 (ICV) à 0.48 (IVT). Les *d* de Cohen associés à une différence de moyennes significative varie de 0.16 à 0.48. Ils indiquent des gains négligeables et sans implication clinique pour les indices IMT, Gf et Gwm. Pour l'IVT, le QIT, l'ICC et Gs, ils montrent une taille de différence petite à modérée. En ce qui concerne les autres indices (ICV, IRP, IAG, Gc et Gv), les moyennes des deux passations ne diffèrent pas sur le plan statistique ; ces indices présentent donc une stabilité absolue à long terme. Autrement dit, en moyenne, les performances des enfants de notre échantillon sont équivalentes lors du test et du retest pour ces indices.

Les subtests Code (+0.82), Matrices (+0.48) et Symboles (+1.09) montrent une augmentation significative, mais sans implication clinique pour Matrices puisque la valeur du *d* de Cohen est faible (*d* = 0.19).

L'augmentation significative des performances moyennes pour certains scores suggère un effet d'apprentissage entre les deux passations. À partir de ces résultats, nous explorons les relations entre la durée de l'intervalle test-retest, l'âge et la différence de performances pour les différents scores.

[26] Pour k comparaisons multiples, le risque global d'erreur est égale à $1 - .95^k$. Par exemple, pour 50 comparaisons multiples, le risque global d'erreur est de 92 %, et non plus 5 %.

Tableau 12

Statistiques descriptives et t-test pour les subtests, les indices standards et les facteurs CHC du WISC-IV

	Test		Retest					
	M	*ET*	*M*	*ET*	*Δ M*	*d*	*ddl*	*|t|*
Subtest								
Cubes	10.97	3.02	10.79	3.07	-0.18	-0.08	276	1.36
Similitudes	11.13	3.35	11.25	2.92	0.12	0.05	276	0.75
Mémoire des chiffres	9.37	2.72	9.68	2.75	0.32	0.13	276	2.11
Identification de concepts	9.38	2.70	9.70	2.37	0.31	0.11	276	1.83
Code	9.77	2.92	10.58	2.56	**0.82***	0.32	276	5.27
Vocabulaire	10.57	2.83	10.61	2.84	0.04	0.02	276	0.31
Séquence Lettres-Chiffres	9.03	3.03	9.38	2.83	0.35	0.11	276	1.86
Matrices	9.29	2.76	9.78	2.54	**0.48***	0.19	276	3.22
Compréhension	10.80	2.70	10.47	2.94	-0.33	-0.13	276	2.11
Symboles	10.71	2.74	11.80	2.90	**1.09***	0.38	276	6.24
Complètement d'images [a]	9.75	2.49	9.74	2.67	-0.01	0.01	247	0.08
Composite								
ICV	104.90	15.23	104.70	15.17	-0.23	-0.02	276	0.40
IRP	99.10	14.54	100.35	13.60	1.26	0.12	276	2.00
IMT	95.18	14.10	97.20	14.33	**2.02***	0.16	276	2.69
IVT	101.38	13.88	107.20	14.03	**5.79***	0.48	276	7.92
QIT	100.80	13.93	103.30	12.78	**2.53***	0.30	276	4.92
IAG	102.60	14.62	103.10	13.43	0.51	0.06	276	0.99
ICC	97.87	13.76	102.70	13.63	**4.87***	0.43	276	7.14
Gc	105.49	14.98	104.92	14.83	-0.57	-0.05	276	0.86
Gf	95.96	13.79	98.04	13.47	**2.08***	0.16	276	2.62
Gv [a]	101.83	14.10	101.32	14.60	-0.51	-0.04	247	0.79
Gwm	95.17	14.12	97.18	14.38	**2.01***	0.16	276	2.68
Gs	101.40	13.61	106.90	13.39	**5.51***	0.47	276	7.79

Note. M = moyenne ; ET = écart type ; Δ *M* = différence de moyennes M2-M1 ; d = d de Cohen ; ddl = degré de liberté ; *t* = *t* de Student ; *p* = *p*-valeur ; ICV = Indice de Compréhension Verbale ; IRP = Indice de Raisonnement Perceptif ; IMT = Indice de Mémoire de Travail ; IVT = Indice de Vitesse de Traitement ; QIT = QI Total ; IAG = Indice d'Aptitude Générale ; ICC = Indice de Compétence Cognitive ; Gc = intelligence cristallisée ; Gf = intelligence fluide ; Gv = Traitement visuel ; Gwm = Mémoire à court terme; Gs = Vitesse de traitement.

[a] *N* = 248.

***** *p* < .05 (avec correction de Holm-Bonferroni).

8.4. DURÉE DE L'INTERVALLE TEST-RETEST ET EFFET D'APPRENTISSAGE

Selon les résultats précédents, un délai d'au moins une année n'est pas suffisant pour éliminer les effets d'apprentissage pour les indices IMT, IVT, QIT, ICC, Gf, Gwm et Gs ainsi que pour les subtests Code, Matrices et Symboles. Nous allons examiner la durée de temps nécessaire pour qu'il n'y ait plus de différences de moyenne statistiquement significative pour ces scores. Pour cela, nous scindons notre échantillon selon la durée du test-retest. Pour avoir des effectifs équilibrés, nous créons trois groupes : (1) intervalles de 12 à 15 mois (N = 90 enfants), (2) intervalles de 16 à 24 mois (N = 92 enfants), (3) intervalles de 25 à 39 mois (N = 95 enfants). Le Tableau 13 décrit les caractéristiques des trois sous-échantillons formés.

Tableau 13

Répartition des intervalles test-retest

	Intervalles 12-15 mois	Intervalles 16-24 mois	Intervalles 25-39 mois
Garçons	43	42	47
Filles	47	50	48
Âge moyen à T1	8.41 (1.03)	9.11 (0.63)	9.07 (0.54)
Âge moyen à T2	9.57 (1.04)	10.78 (0.66)	11.53 (0.49)

Pour comparer les différences de moyennes, des *t*-tests pour échantillons appariés sont réalisés et une correction de Holm-Bonferroni est appliquée sur le seuil α pour le maintenir à 5 %. Les résultats pour les indices sont illustrés dans la Figure 52 (p. 222). Pour des intervalles d'un an à un an et trois mois, il y a des différences de moyenne significatives entre les deux passations pour les indices IVT, QIT, ICC, Gf et Gs. Toutefois, les tailles d'effet sont négligeables à petites, variant de d = 16 (Gf) à d = 0.41 (IVT). Pour l'ensemble des indices étudiés, il y a des différences de moyenne significatives entre les deux passations pour les intervalles d'un an et quatre mois à deux ans. Il est à relever des tailles de différences modérées à grandes pour l'IVT (d = 0.65), l'ICC (d = 0.55) et Gs (d = 0.63). Pour les intervalles de plus de deux ans, on observe encore des différences de moyenne significatives pour l'IVT, l'ICC et Gs, cependant les tailles d'effet sont petites (variant de 0.22 à 0.23).

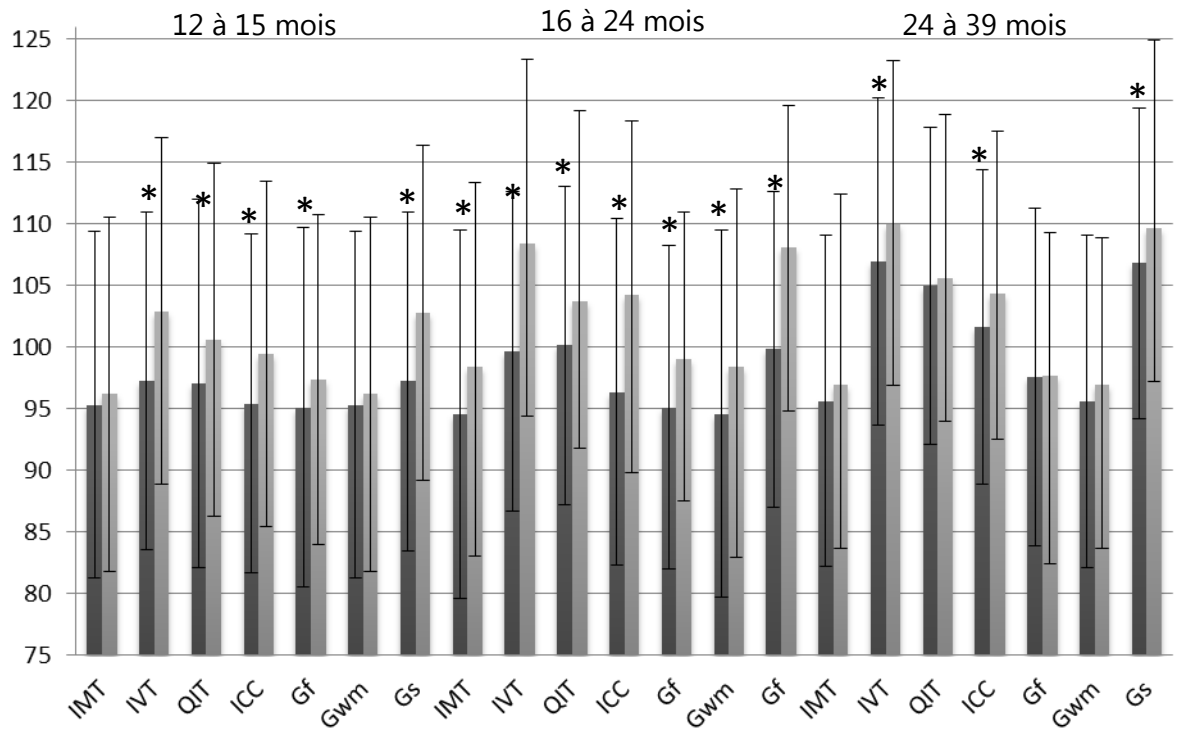


Figure 52. Histogrammes des moyennes entre Test (barre gris foncé) et Retest (barre gris clair).

La Figure 53 (p. 223) illustre les résultats pour les subtests Codes, Matrice et Symboles. Pour des intervalles d'un an à un an et trois mois, il y a des différences de moyenne significatives entre les deux passations pour les deux subtests de l'Indice de Vitesse de Traitement. Si la taille d'effet de Code est négligeable (d = 0.18), en revanche, celle de Symboles est modérée (d = 50). Pour les trois subtests étudiés, il y a des différences de moyenne significatives entre les deux passations pour les intervalles d'un an et quatre mois à deux ans. Les tailles de différences sont modérées à grandes allant de d = 0.39 (Matrices) à d = 0.63 (Code). Pour les intervalles de plus de deux ans, on observe des différences de moyenne significatives uniquement pour Symboles (d = 0.27).

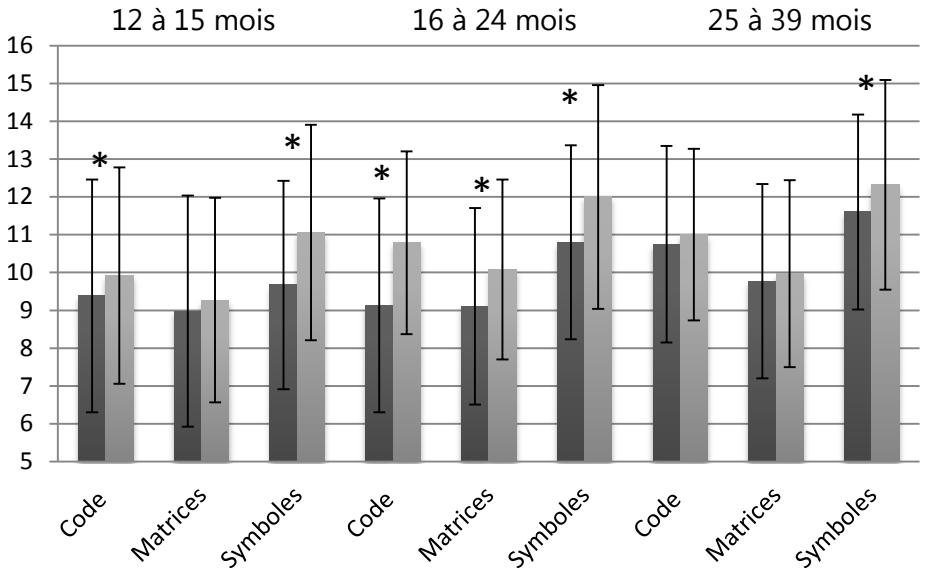


Figure 53. Histogrammes des moyennes entre Test (barre gris foncé) et Retest (barre gris clair) pour Code, Matrices et Symboles.

Nos données indiquent que pour éviter un effet d'apprentissage, l'intervalle entre le test et le retest doit être supérieur à deux ans, voire supérieur à trois ans pour l'IVT, l'ICC, Gs et Symboles.

8.5. STABILITÉ DIFFÉRENTIELLE

Les analyses des différences individuelles au niveau groupal ont également évalué si l'ordre des individus est relativement similaire entre les deux passations : les enfants les plus forts à la première passation tendent-ils à rester les plus forts à la seconde passation ? Et inversement ? Pour cela, des corrélations entre les performances au test et les performances au retest sont calculées pour les différents indices. Le Tableau 14 (p. 224) présente les coefficients de corrélation test-retest non corrigés (r_{12}) ainsi que les coefficients test-retest corrigé (r_c) pour tenir compte de la variabilité de l'échantillon d'étalonnage (Magnusson, 1967). Pour une représentation graphique, dans l'Annexe G se trouvent les figures des nuages de points.

Tableau 14

Coefficients des stabilité pour les subtests, les indices standards et les facteurs CHC du WISC-IV

	r_{12}	r_c
Subtest		
Cubes	.74	.73
Similitudes	.67	.59
Mémoire des chiffres	.58	.65
Identification de concepts	.40	.51
Code	.56	.59
Vocabulaire	.71	.74
Séquence Lettres-Chiffres	.42	.40
Matrices	.56	.63
Compréhension	.59	.67
Symboles	.47	.55
Complètement d'images [a]	.57	.70
Composite		
ICV	.80	.79
IRP	.73	.74
IMT	.61	.66
IVT	.62	.67
QIT	.80	.83
IAG	.82	.83
ICC	.66	.71
Gc	.73	.73
Gf	.53	.60
Gv [a]	.75	.78
Gwm	.61	.66
Gs	.62	.69

Note. r_{12} = coefficient de stabilité non corrigé ; r_c = coefficient de stabilité corrigé ; ICV = Indice de Compréhension Verbale ; IRP = Indice de Raisonnement Perceptif ; IMT = Indice de Mémoire de Travail ; IVT = Indice de Vitesse de Traitement ; QIT = QI Total ; IAG = Indice d'Aptitude Générale ; ICC = Indice de Compétence Cognitive ; Gc = intelligence cristallisée ; Gf = intelligence fluide ; Gv = Traitement visuel ; Gwm = Mémoire à court terme; Gs = Vitesse de traitement.

[a] N = 248.

Les coefficients de stabilité corrigés des indices standards varient de .66 (IMT) à .83 (QIT et IAG), tandis que ceux des indices CHC varient de .60 (Gf) à .78 (Gv). Pour les subtests, on observe que les coefficients de fidélité corrigés varient de .40 (Séquences-Lettres-Chiffres) à .74 (Vocabulaire). Comme les indices ont une étendue

des scores observés plus large que les subtests, les coefficients de corrélation des subtests sont inférieurs à ceux des indices auxquels ils contribuent. De même, toute chose étant égale par ailleurs, les indices globaux (QIT, IAG et ICC) présentent des coefficients de corrélation plus élevés que les indices qui les composent.

Si nous interprétions les coefficients de stabilité en termes de variance vraie et de variance d'erreur, on peut relever que la variance d'erreur explique plus de 30 % de la variance totale des indices IMT, IVT, GF et Gwm. Quant au QIT et à l'IAG qui présentent les coefficients de stabilité les plus élevés, 83 % de la variance totale est expliquée par de la variance vraie et 17 % par de la variance d'erreur. Pour les subtests Identification de concepts et Symboles, seule un peu plus de la moitié de la variance totale est expliquée par de la variance vraie. Pour le subtest Séquence lettres-chiffres, il y a 60 % de la variance totale qui est expliquée par de la variance d'erreur.

8.6. STABILITÉ INTRA-INDIVIDUELLE ABSOLUE

Dans une perspective « diagnostic », il est important d'évaluer la stabilité des performances à long terme au niveau individuel. En effet, il est indispensable de vérifier que les niveaux de performances de chaque enfant restent stables au cours du temps pour les différents indices. La stabilité intra-individuelle est la condition nécessaire et obligatoire pour des prédictions à long terme.

Tout d'abord, l'estimation de cette stabilité intra-individuelle est réalisée en calculant un intervalle de confiance pour chaque enfant et pour chaque indice. Si les performances des enfants entre les deux passations sont comprises dans l'intervalle de ±2 erreurs types de mesure (±2 ETM), alors les performances intra-individuelles sont considérées comme stables. Dans le cas contraire, les performances des enfants sont considérées comme instables. Dans le Tableau de l'Annexe F sont reportés l'étendue des intervalles correspondant à ±1 ETM, ±2 ETM et ±3 ETM pour chaque score. En vertu des propriétés de la courbe normale, nous rappelons que théoriquement, 68 % des enfants devraient présenter des variations de performances dans l'intervalle de ±1 ETM, 95 % devraient présenter des variations de performances dans l'intervalle de ±2 ETM, et 99 % devraient présenter des variations de performances dans l'intervalle de ±3 ETM.

Dans le Tableau 15 (p. 227) sont présentés les pourcentages d'enfants dont les performances au test et au retest sont comprises dans l'intervalle de ±1 ETM, ±2 ETM

et ±3 ETM. À notre connaissance, il n'existe pas de critères pour déterminer si la stabilité intra-individuelle d'un indice est satisfaisante ou non. En l'absence de critères proposés dans la littérature, nous considérons qu'un indice présentant une stabilité intra-individuelle (avec un intervalle de ±2 ETM) pour au moins 80 % des enfants est acceptable. En d'autres termes, les scores qui permettent de positionner 4 enfants sur 5 de manière stable sont considérés comme satisfaisant sur le plan du diagnostic intra-individuel. Toutefois, cela signifie également que la prédiction sera incorrecte pour 1 enfant sur 5.

Tant parmi les indices standard que CHC, aucun ne présente une stabilité intra-individuelle suffisamment acceptable à long terme. Les indices ICV, IRP, IAG, Gc et Gv sont les plus stables avec un peu plus de 70 % des enfants ayant des variations de performances entre les deux passations à l'intérieur de l'intervalle de ±2 ETM (ou ±9.96 points pour l'ICV ; ±10.48 points pour l'IRP ; ±8.28 points pour l'IAG ; ±11.85 points pour Gc ; et ±11.42 points pour Gv).

En ce qui concerne le QI Total, les différences de scores individuels entre les deux passations varient de -25 points à +28 points (voir Annexe G pour les distributions des fréquences cumulées des scores de différences entre les deux passations). Moins de la moitié des enfants (32.9 %) ont des performances variant à l'intérieur de l'intervalle de ±1 ETM. En revanche, 81.2 % des enfants présentent des performances de QIT stables pour l'intervalle de ±3 ETM.

Quant aux subtests, les performances à Cubes, Identification de concepts, Code, Vocabulaire et Compréhension sont stables, avec un peu plus de 80 % des enfants présentant des performances entre les deux passations à l'intérieur de l'intervalle de ±2 ETM.

Tableau 15

Pourcentages d'enfants avec des performances stables entre le Test et le Retest selon un intervalle de ±1 ETM, ±2 ETM et ±3 ETM autour du score vrai estimé à T1

	% d'enfants ayant des performances stables entre les deux passations		
	IC à ±1 ETM	IC à ±2 ETM	IC à ±3 ETM
Subtest			
Cubes	51.6	81.6	91.0
Similitudes	44.0	74.0	92.4
Mémoire des chiffres	43.3	71.5	84.8
Identification de concepts	50.5	81.9	92.8
Code	57.8	84.5	94.9
Vocabulaire	52.0	86.6	96.4
Séquence Lettres-Chiffres	37.9	68.6	85.2
Matrices	37.2	67.9	87.0
Compréhension	57.4	82.7	95.7
Symboles	30.3	62.5	79.8
Complètement d'images [a]	52.4	72.6	85.9
Composite			
ICV	39.4	72.9	90.3
IRP	46.9	70.0	87.7
IMT	35.0	63.5	81.9
IVT	30.0	61.4	85.6
QIT	32.9	59.9	81.2
IAG	43.0	71.5	87.7
ICC	31.8	62.1	80.9
Gc	39.4	67.1	89.2
Gf	39.4	72.6	88.8
Gv [a]	49.2	79.0	91.9
Gwm	32.9	61.0	76.0
Gs	31.0	62.1	87.0

Note. IC = intervalle de confiance ; ETM = erreur type de mesure ; ICV = Indice de Compréhension Verbale ; IRP = Indice de Raisonnement Perceptif ; IMT = Indice de Mémoire de Travail ; IVT = Indice de Vitesse de Traitement ; QIT = QI Total ; IAG = Indice d'Aptitude Générale ; ICC = Indice de Compétence Cognitive ; Gc = intelligence cristallisée ; Gf = intelligence fluide ; Gv = Traitement visuel ; Gwm = Mémoire à court terme; Gs = Vitesse de traitement.

[a] N = 248.

8.7. STABILITÉ CATÉGORIELLE

Dans l'optique de situer et de donner du sens aux notes des indices, les cliniciens recourent à des descriptions qui qualifient les performances du sujet. De par son impression au dos des cahiers de passation du WISC-IV, la lecture en sept catégories – très faible (≤ 69), limite (70-79), moyen faible (80-89), moyen (90-109), moyen fort (110-119), supérieur (120-129), et très supérieur (≥ 130) – est la plus régulièrement utilisée par les praticiens. La lecture en trois catégories – faible (≤ 84), dans la moyenne (85-115), et élevé (≥ 116) – découpe les performances en un écart type en dessous et en un écart type au-dessus de la moyenne de 100 (Flanagan & Kaufman, 2009). Enfin, une classification en cinq catégories – extrémité inférieure (≤ 69), moyen faible (70-84), dans la moyenne (85-115), moyen fort (116-130), et extrémité supérieure (≥ 131) – est proposée par Flanagan et Kaufman (2009).

Dans le Tableau 16 (p. 229) figure un comparatif des pourcentages d'enfants « stables », c'est-à-dire des enfants qui sont dans la même catégorie descriptive à la première et à la seconde passation. Les résultats montrent qu'entre 41.9 % (IVT) et 58.5 % (IAG) des enfants restent dans la même catégorie pour la classification en sept catégories. En ce qui concerne la classification en trois catégories, la stabilité catégorielle se révèle meilleure, puisque plus de 70 % des enfants restent dans la même catégorie lors des deux passations (à l'exception de l'ICC). Quant à la classification en cinq catégories, entre 65 % (ICC) et 76.9 % (IAG) des enfants restent dans la même catégorie entre la première et la seconde passation. Avec la classification en trois catégories, à peu près huit enfants sur dix voient leur IAG et leur QI Total demeurer dans la même catégorie descriptive sur le long terme. Il semble donc qu'on puisse faire des prédictions catégorielles relativement fiables à partir de ces deux indices, si l'on utilise le système en trois catégories.

À partir des résultats sur la stabilité catégorielle présentés ci-dessus, il apparaît que les classifications en trois et en cinq catégories permettent des prédictions relativement fiables. Cependant, la classification en cinq catégories a l'avantage de discriminer un peu plus finement les performances que la classification en trois catégories. Pour cette raison, il nous semble plus pertinent de présenter en détail les résultats de la classification en cinq catégories à la place de ceux de la classification en trois catégories.

Tableau 16

Pourcentages d'enfants « stables » entre les deux passations selon la lecture catégorielle

	Classification en 7 catégories		Classification en 3 catégories		Classification en 5 catégories	
	N	%	N	%	N	%
ICV	142	51.3	197	71.1	187	67.5
IRP	141	50.9	206	74.4	199	71.8
IMT	127	45.8	201	72.6	190	68.6
IVT	116	41.9	202	72.9	190	68.6
QIT	147	53.1	216	78.0	212	76.5
IAG	162	58.5	223	80.5	213	76.9
ICC	124	44.8	186	67.1	180	65.0

Note. ICV = Indice de Compréhension Verbale ; IRP = Indice de Raisonnement Perceptif ; IMT = Indice de Mémoire de Travail ; IVT = Indice de Vitesse de Traitement ; QIT = QI Total ; IAG = Indice d'Aptitude Générale ; ICC = Indice de Compétence Cognitive.

Le Tableau 17 (p. 230) présente la répartition catégorielle des enfants d'une passation à l'autre, selon la classification en cinq catégories. Ainsi, parmi les 187 enfants ayant des performances stables pour l'ICV (c.-à-d. qui sont restés dans la même catégorie soit 1 + 7 + 144 + 28 + 7 = 187 enfants), la grande majorité a des performances catégorielles « moyenne ». En effet, 144 enfants ont des performances catégorisées dans la moyenne lors des deux passations, ce qui représente 77 % des enfants ayant des performances stables pour l'ICV. Cette constatation s'applique également aux trois autres indices. Ainsi sont restés dans la moyenne entre les deux passations, 161 enfants sur 199 enfants pour l'IRP (soit 80.9 %). Pour l'IMT, 190 enfants ont des performances stables entre les deux passations, dont 166 qui sont restés dans la catégorie moyenne (soit 87.4 %). Enfin, pour l'IVT, 190 enfants ont des performances stables entre les deux passations, dont 175 qui sont restés dans la catégorie moyenne (soit 92.1 %).

Lorsque les performances des indices changent de catégorie d'une passation à l'autre, il s'agit généralement d'un changement vers **une** catégorie plus haute, ou vers **une** catégorie plus basse. Il est rare qu'à la seconde passation, le changement soit de plusieurs catégories au-dessus, ou de plusieurs catégories en dessous.

Tableau 17

Répartition des performances au test et au Retest pour les quatre indices selon la classification en cinq catégories

	Retest					Total
	Extrémité inférieure	faible	Dans la moyenne	forte	Extrémité supérieure	
Test						
ICV						
Extrémité inférieure	**1**	1	0	0	0	2
faible	1	**7**	18	0	0	26
Dans la moyenne	0	14	**144**	22	1	181
forte	0	0	25	**28**	4	57
Extrémité supérieure	0	0	0	4	**7**	11
IRP						
Extrémité inférieure	**3**	3	3	0	0	9
faible	2	**15**	18	0	0	35
Dans la moyenne	0	14	**161**	20	1	196
forte	0	0	15	**19**	0	34
Extrémité supérieure	0	0	0	2	**1**	3
IMT						
Extrémité inférieure	**2**	6	2	0	0	10
faible	3	**14**	25	1	0	43
Dans la moyenne	2	21	**166**	17	1	207
forte	0	0	6	**8**	1	15
Extrémité supérieure	0	0	1	1	**0**	2
IVT						
Extrémité inférieure	**0**	1	2	0	0	3
faible	0	**2**	16	0	0	18
Dans la moyenne	0	7	**175**	25	12	219
forte	0	0	12	**10**	7	29
Extrémité supérieure	0	0	1	4	**3**	8

Note : ICV = Indice de Compréhension Verbale ; IRP = Indice de Raisonnement Perceptif ; IMT = Indice de Mémoire de Travail ; IVT = Indice de Vitesse de Traitement.

En gras sont indiquées le nombre de performances stables sur le plan catégoriel entre la première et la seconde passation.

Dans un focus sur le sous-échantillon d'enfants avec les performances à un écart type en dessous de la moyenne (< 85), nous examinons les changements de catégories pour les indices globaux (QIT, IAG et ICC). L'intérêt de ce sous-échantillon est qu'il correspond au groupe d'enfants qui à l'issue d'une première passation au WISC-IV sont décrits « en difficultés » par rapport aux performances de leur groupe de référence. Tout d'abord, le Tableau 18 renseigne sur les caractéristiques des sous-échantillons d'enfants avec un QIT inférieur à 85 (*N* = 35 enfants), un IAG inférieur à 85 (*N* = 30 enfants) ou un ICC inférieur à 85 (*N* = 47 enfants).

Tableau 18

Statistiques descriptives des sous-échantillons d'enfants avec les performances au Test à un écart type en dessous de la moyenne (< 85) pour le QI Total, l'IAG et l'ICC

	QI Total	Indice d'Aptitude Générale	Indice de Compétence Cognitive
Garçons	20	14	25
Filles	15	16	22
Âge moyen au Test	8.99 (1.05)	9.04 (1.11)	8.75 (0.87)
Âge moyen au Retest	10.44 (1.23)	10.48 (1.28)	10.39 (1.16)
Intervalle test-retest moyen	1.41 (0.36)	1.38 (0.36)	1.60 (0.49)

Dans le Tableau 19 (p. 232) figure la répartition des performances des enfants à la première et à la seconde passation. Pour les enfants dont le score global est inférieur à 85 à la première passation, 13 enfants sur 35 (soit 37.1 %) pour le QIT, 16 enfants sur 30 (soit 53.3 %) pour l'IAG et 16 enfants sur 47 (soit 34 %) pour l'ICC sont restés dans cette catégorie de performance à la seconde passation. Pour les enfants dont le score global est dans la moyenne normative (entre 85 et 115) à la première passation, 86.5 % pour le QIT, 88 % pour l'IAG et 77.2 % pour l'ICC sont restés dans cette même catégorie à la seconde passation. Pour les enfants dont le score global est supérieur à 115 à la première passation, 74.4 % pour le QIT, 67.9 % pour l'IAG et 50 % pour l'ICC sont restés dans cette catégorie de performance à la seconde passation. Les performances des enfants dans les catégories en dessous et au-dessus de la moyenne tendent à revenir vers la moyenne à la seconde passation, tandis que les performances dans la moyenne présentent une stabilité catégorielle pour la plupart des enfants entre les deux passations.

Tableau 19

Répartition des performances au Test et au Retest pour le QI Total, l'IAG et l'ICC

	Retest			
	Indice ≤ 84	85 ≤ Indice ≤ 115	Indice ≥ 116	Total
Test				
QIT ≤ 84	**13**	22	0	35
85 ≤ QIT ≤ 115	4	**173**	23	200
QIT ≥ 116	0	12	**30**	42
IAG ≤ 84	**16**	14	0	30
85 ≤ IAG ≤ 115	8	**168**	14	191
IAG ≥ 116	0	18	**38**	56
ICC ≤ 84	**16**	31	0	47
85 ≤ ICC ≤ 115	12	**156**	34	202
ICC ≥ 116	0	14	**14**	28

Note : QIT = QI Total ; IAG = Indice d'Aptitude Générale ; ICC = Indice de Compétence Cognitive.

En gras sont indiquées le nombre de performances stables sur le plan catégoriel entre la première et la seconde passation.

Le phénomène de régression à la moyenne explique ces résultats, ainsi que l'effet d'apprentissage. En effet, on peut relever que les enfants dans la catégorie des performances supérieures à un écart type au-dessus de la moyenne restent davantage dans cette catégorie que les enfants présentant des performances à un écart type en dessous de la moyenne (p. ex., 74.4 % vs 37.1 % pour le QIT).

8.8. STABILITÉ DES FORCES ET DES FAIBLESSES

La comparaison entre chaque indice et l'indice moyen permet d'identifier les forces et les faiblesses personnelles. Suivant la procédure décrite précédemment (voir 6.2.2.1, p. 199), un indice est décrit comme force personnelle (FoP) ou faiblesse personnelle (FaP) du sujet selon qu'il s'écarte significativement vers le haut ou vers le bas de la moyenne des performances des quatre indices, c'est-à-dire de l'indice moyen (IM).

Pour la première passation, l'indice moyen des 277 enfants s'étend de 67.25 à 124 (moyenne = 100.14 et écart type = 10.04). À la seconde passation, l'indice moyen s'étend de 74 à 126.75 (moyenne = 102,35 et écart type = 9.24). Quant aux différences entre un indice et l'indice moyen, elles varient de façon importante de -36,50 (IMT-IM) à +35,75 (ICV-IM) pour la première passation, et de -33,75 (IMT-IM) à +36,25 (IVT-IM) pour la seconde passation. Si l'on fait la moyenne des différences absolues entre chaque indice et l'indice moyenne, celles sont assez élevées, allant de 7.40 avec un écart type de 5.93 (|IRP-IM|) à 9.07 avec un écart type de 7.15 (|IVT-IM|) au test, et allant de 8.21 avec un écart type de 5.75 (|IRP-IM|) à 10.09 avec un écart-type de 7.75 (|IVT-IM|) au retest. Dans notre échantillon d'enfants, l'IRP est en moyenne l'indice qui s'écarte le moins de l'indice moyen du sujet, tandis que l'IVT est celui qui s'en écarte le plus largement.

Le Tableau 20 renseigne sur le pourcentage de forces personnelles (FoP), de faiblesses personnelles (FaP) et de moyennes personnelles (MoP) pour les différents indices. Le pourcentage d'enfants présentant des performances « moyenne » sur le plan personnel varie de 68.6 % (IMT) à 77.6 % (IRP) à la première passation et de 65.7 % (IMT) à 73.3 % (IRP) pour la seconde passation. Pour 26.4 % des enfants, l'ICV est une force personnelle, tandis que l'IMT est une faiblesse personnelle pour 26 % des enfants au test. Au retest, l'IVT est une force personnelle pour 26.7 % des enfants, alors que l'IMT est une faiblesse personnelle pour 28.9 % des enfants. Entre les deux passations, la répartition des performances en forces, faiblesses et moyennes personnelles est relativement comparable au sein de chaque indice, sauf pour l'IVT. Ce dernier représente une force personnelle pour 15.2 % des enfants au test et pour 26.7 % des enfants au retest.

Tableau 20

Pourcentages de forces et de faiblesses personnelles pour les différents indices

	Test			Retest		
	FaP (%)	MoP (%)	FoP (%)	FaP (%)	MoP (%)	FoP (%)
ICV	4.7	69.0	26.4	11.6	67.5	20.9
IRP	12.3	77.6	10.1	18.4	73.3	8.3
IMT	26.0	68.6	5.4	28.9	65.7	5.4
IVT	8.3	76.5	15.2	6.9	66.4	26.7

Note. FaP = faiblesse personnelle ; MoP = moyenne personnelle ; FoP = force personnelle ; ICV = Indice de Compréhension Verbale ; IRP = Indice de Raisonnement Perceptif ; IMT = Indice de Mémoire de Travail ; IVT = Indice de Vitesse de Traitement.

Passons aux résultats de la stabilité à long terme des forces et des faiblesses personnelles qui sont présentés dans le Tableau 21. Les performances « moyenne personnelle » (MoP) des quatre indices sont relativement stables d'une passation à l'autre. En effet, 152 enfants sur 202 (8+152+42 = 202) se sont maintenus dans la catégorie « moyenne personnelle » pour l'ICV (soit 75.2 %), 171 enfants sur 201 (19+171+11 = 201) pour l'IRP (soit 85.1 %), 146 enfants sur 197 (45+146+6 = 197) pour l'IMT (soit 74.1 %), et enfin 155 enfants sur 191 (7+155+29 = 191) pour l'IVT (soit 81.2 %).

Tableau 21

Stabilité des forces et des faiblesses personnelles entre les deux passations

	Retest			Total
	FaP	MoP	FoP	
Test				
ICV				
FaP	**8**	5	0	13
MoP	23	**152**	16	191
FoP	1	30	**42**	73
IRP				
FaP	**19**	15	0	34
MoP	32	**171**	12	215
FoP	0	17	**11**	28
IMT				
FaP	**45**	27	0	72
MoP	35	**146**	9	190
FoP	0	9	**6**	15
IVT				
FaP	**7**	16	0	23
MoP	12	**155**	45	212
FoP	0	13	**29**	42

Note : FaP = faiblesse personnelle ; MoP = moyenne personnelle ; FoP = force personnelle ; ICV = Indice de Compréhension Verbale ; IRP = Indice de Raisonnement Perceptif ; IMT = Indice de Mémoire de Travail ; IVT = Indice de Vitesse de Traitement.

En gras sont indiqués le nombre d'enfants ayant des forces, des faiblesses ou des moyennes personnelles stables entre la première et la seconde passation.

La stabilité des forces personnelles (FoP) est globalement peu satisfaisante, voire faible pour l'IRP et l'IMT. En effet, pour l'IRP, 11 enfants sur 28 seulement présentent une stabilité de cette « force personnelle » (soit 39.3 %). Pour l'IMT, 6 enfants sur 15 présentent une stabilité de cette « force personnelle » (soit 40 %).

En ce qui concerne la stabilité des faiblesses personnelles (FaP), elle est également peu satisfaisante pour les différents indices, et notamment pour l'IVT. En effet, pour ce dernier indice, seuls 7 enfants sur 23 présentant une faiblesse personnelle lors de la première passation ont présenté une faiblesse personnelle à la seconde passation (soit 30.4 %). En résumé, les enfants qui présentent une force ou faiblesse personnelle à la première passation ont tendance à présenter des performances moyennes sur le plan personnel lors de la seconde passation. Environ la moitié seulement des enfants qui présentaient des forces ou des faiblesses lors de la première passation, ont présenté ces mêmes forces ou faiblesses lors de la seconde passation.

DISCUSSION

9. LE FONCTIONNEMENT DIFFÉRENTIEL DES ITEMS DU WISC-IV

Dans une première étude, nous nous sommes intéressée à la détection de biais d'items pour le WISC-IV. Étant donné que le contenu des items d'un test est plus ou moins imprégné de la culture dans laquelle baignent les concepteurs du test, la transposition d'un test dans une autre langue n'est pas une simple traduction. Il s'agit d'une adaptation qui tient compte des différences culturelles, linguistiques et éducatives d'un pays à l'autre. Un test comme les Échelles de Wechsler prend particulièrement de temps à être adapté, d'une part à cause de l'importance des contenus linguistique et culturel dans certains subtests, et d'autre part, à cause des normes fournies par tranche d'âge de trois mois en trois mois qui demandent un large échantillon d'individus. Dans des subtests verbaux comme Vocabulaire, la traduction du mot à définir d'une langue à l'autre peut modifier le degré de difficulté de l'item selon que le mot traduit renvoie à des représentations conceptuelles différentes en français qu'en anglais. De même, des représentations imagées dans certains items peuvent être plus faciles/difficiles à reconnaître, parce que l'objet ou la situation est plus/moins familier pour les enfants français que pour les enfants américains. La validation de l'adaptation d'un test doit donc s'assurer que les versions du test – originale et adaptée – évaluent les mêmes choses et de manière équivalente. L'équivalence des scores est nécessaire pour les comparaisons des individus sur ce qu'évalue le test. Afin de présenter un niveau de difficulté équivalente, il est fréquent que des items soient modifiés, déplacés dans leur ordre d'administration ou remplacés lors de l'adaptation d'un test dans une autre langue. Par exemple, l'adaptation en français du subtest Vocabulaire du WISC-IV comporte 15 items sur 36 qui ont été changés par rapport à la version américaine.

La comparaison des groupes ne se limite pas à des différences entre populations parlant différentes langues, mais également au sein d'une même communauté linguistique. En effet, compte tenu du coût d'une adaptation, un test dans une langue donnée est souvent utilisé pour différentes populations d'une même communauté linguistique. Par exemple, l'adaptation en français du WISC-IV est utilisée pour les enfants francophones de France, de Belgique et de Suisse. Or, étant donné les différences culturelles, éducatives et les spécificités linguistiques entre ces trois pays, il est également important d'examiner l'équivalence des scores entre les individus à qui on administre une même adaptation d'un test. Cette comparaison a été examinée avec

une version préexpériementale de l'adaptation en français du WISC-IV. En effet, sur la version préexpérimentale du WISC-IV, une étude sur deux populations d'enfants est réalisée qui comprend 220 enfants français et 125 enfants belges de même âge et de même niveau scolaire. À l'issue des analyses, les items qui ne présentent pas la même probabilité de réussite auprès des enfants français et belges, sont supprimés. Ainsi, la version définitive du WISC-IV, est sans items biaisés à l'encontre des enfants belges francophones. En revanche, aucune comparaison n'a pas été réalisée avec un groupe d'enfants suisses francophones à qui pourtant le test s'adresse également. Il s'agit d'une lacune à laquelle nous n'avons pas pu remédier, car nous n'avons pas accès à l'échantillon de standardisation de l'adaptation en français pour la comparer avec notre échantillon d'enfants suisses francophones.

Dans notre étude sur le fonctionnement différentiel des items du WISC-IV, les analyses portent uniquement sur des enfants suisses francophones. Plus précisément, les données que nous avons récoltées constituent un échantillon de 483 enfants suisses francophones âgés de 7 à 12 ans. Les dix subtests obligatoires du WISC-IV ainsi que le subtest optionnel Complémentement d'images ont été administrés. À l'aide des modèles de réponse à l'item (MRI), les paramètres de l'item (p. ex., sa difficulté, sa discrimination) peuvent être estimés et le comportement de chaque item peut être modélisé. La modélisation du comportement d'un item est représentée par une courbe caractéristique de l'item (CCI) qui donne la probabilité de réussir l'item en fonction de l'habileté sur le trait latent évalué. L'approche des MRI propose des outils pour la détection des biais de l'item. En effet, en comparant les courbes caractéristiques de l'item de groupes différents, on peut évaluer l'écart entre les CCI de chaque groupe. Si un item se comporte de la même manière, c'est-à-dire si pour une même habileté sur le trait latent, tous les individus ont la même probabilité de le réussir, alors les CCI des différents groupes considérés sont proches, voire se superposent. Si un item ne se comporte pas de la même manière, c'est-à-dire si, pour une même habileté sur le trait latent, les individus qui se distinguent sur une variable n'ont pas la même probabilité de le réussir, alors les CCI des différents groupes sont distantes l'une de l'autre. On considère alors que l'item présente un fonctionnement différentiel (c.-à-d. un comportement différent selon le groupe considéré). La présence d'un fonctionnement différentiel de l'item (FDI) est un indicateur de biais d'items sur le plan statistique. En effet, nous rappelons que deux critères sont à considérer pour parler d'item biaisé à l'encontre d'un groupe. Il faut que (a) deux individus d'habileté équivalente sur la propriété mentale évaluée par le test, mais issus de deux groupes distincts, n'aient pas

la même probabilité de réussir un même item (fonctionnement différentiel sur le plan statistique) et que (b) la différence de probabilité de réussir un même item dépende d'une autre variable que la propriété mentale évaluée par le test (Bertrand & Blais, 2004). La présence d'un FDI à lui seul n'est pas a proprement parlé suffisant, il faut encore montrer qu'il engendre une iniquité effective dans l'interprétation des scores du test. Pour cela, une réflexion conceptuelle et contextuelle est nécessaire sur le construit évalué. Nous pouvons rappeler l'exemple d'un test de mathématiques dont les items sont des problèmes de maths à résoudre. En langue anglaise, la formulation des problèmes de maths est plus courte et plus directe qu'en langue française. L'adaptation du test en français fera davantage intervenir les compétences en lecture qu'en anglais. Entre un échantillon américain et un échantillon français, des différences de probabilité de réussir un même item peuvent s'observer pour des individus ayant la même habileté en mathématiques (fonctionnement différentiel sur le plan statistique). Cependant, s'agit-il d'un biais ? Cela dépend si dans notre position théorique, on considère que l'habileté de raisonnement quantitatif comprend une certaine habileté en lecture ou non.

Dans notre étude, nous n'avons pas la possibilité de faire des comparaisons entre les enfants français et les enfants suisses francophones. Nos analyses se sont portées sur les variables âge, sexe et statut socio-économique. Même s'il n'y a pas lieu de penser à un biais d'item en fonction de l'âge des enfants, la variable âge est néanmoins considérée à cause de l'étendue des âges de l'échantillon étudié. Dans les modèles de réponse à l'item (MRI), l'analyse statistique est réalisée sur les points bruts obtenus à chaque item. Comme notre échantillon est composé d'une étendue d'âge allant de 7 à 12 ans, nous avons contrôlé l'âge dans chaque modèle afin de ramener les enfants sur une échelle comparable. En effet, rappelons que les scores bruts ne permettent pas d'interprétation des performances puisque l'étendue des valeurs possibles dépend du nombre d'items du test et des points qui peuvent être obtenus sur chaque item. De plus, la performance d'un sujet à un test prend sens dans la comparaison avec les performances d'autres individus avec qui le sujet partage des caractéristiques communes. Pour un test dont les mêmes items s'adressent à une étendue d'enfants de 6 à 16 ans, on comprend encore mieux que les scores bruts n'ont pas de signification en tant que telle. En effet, on s'attend qu'avec l'âge, les enfants réussissent un nombre plus élevé d'items puisque au cours d'un développement normal, les enfants accumulent de nouveaux apprentissages et de nouvelles connaissances. Dès lors, un score brut donné peut chez un enfant de 7 ans montrer une

performance élevée par rapport à son groupe de référence, alors que le même score brut chez un enfant de 8 ans montre une performance moyenne par rapport à son groupe de référence. Les scores bruts doivent donc être transformés en scores standards de distribution normale avec moyenne et l'écart type connus pour permettre une interprétation des performances entre les (sub)tests et entre les individus de différents âges. Comme dans les analyses avec des modèles de réponses à l'item, les scores bruts sur chaque item sont utilisés, nous devons pour chaque modèle testé contrôler pour l'âge.

Nous avons testé l'influence des différences de sexe suivant les hypothèses que les garçons tendent à obtenir des scores plus élevés en rotations mentales (Cubes), tandis que les filles se distingueraient sur les tâches verbales (Similitudes, Vocabulaire, Compréhension). Quant au statut socio-économique des parents, cette variable est bien connue pour influencer les performances cognitives, notamment sur les épreuves verbales. De nombreuses études montrent que les enfants de milieux socio-économiques élevés obtiennent de meilleures performances que les enfants de milieux socio-économiques bas.

Dans les analyses, les subtests de l'IMT (Mémoire des chiffres et Séquence lettres-chiffres) et de l'IVT (Code et Symboles) n'ont pas été modélisés. Pour l'IMT, il n'y a pas d'hypothèses qui soutiennent l'éventualité de biais d'item dans les habiletés de mémoire à court terme en fonction des variables considérées (c.-à-d. âge, sexe et statut socio-économique). Les items des subtests de l'IMT sont composés de séries de chiffres (Mémoire des chiffres) ou de séries de chiffres et de lettres (Séquence lettres-chiffres) de plus en plus longues. Même si l'âge, le sexe ou le statut socio-économique peuvent avoir une influence directe (par ex. les enfants de milieux socio-économiques élevés obtiennent en moyenne de meilleures performances que les enfants de milieux socio-économique bas), il n'y a pas lieu de supposer une influence indirecte (c.-à-d. qu'ayant la même habileté, deux individus appartenant à deux groupes différents n'ont pas la même probabilité de réussir l'item), suggérant ainsi un biais d'item dans un échantillon d'enfants suisses francophones. Pour l'IVT, les paramètres d'items ne peuvent pas être estimés correctement. Les items ne sont pas ordonnés par ordre croissant de difficulté, puisqu'ils présentent tous à peu près la même difficulté. La tâche demandée dans ces subtests est très simple ; le but est de la réaliser le plus vite possible. Les items échoués le sont par manque de temps, et non à cause de leur difficulté. La plupart des items des subtests de l'IVT n'ont montré aucune variance ; leur modélisation avec des MRI n'était donc pas possible.

Comme condition d'application, les MRI postulent l'unidimensionnalité des items. En préambule aux analyses sur le fonctionnement différentiel, une analyse qui compare l'ajustement des données au modèle unidimensionnel a montré un ajustement suffisant pour les subtests Cubes, Similitudes, Vocabulaire, Matrices et Compréhension selon les critères des indices d'ajustement RMSEA et CFI. Quant aux subtests Identification de concepts et Complètement d'images, l'ajustement au modèle unidimensionnel n'était pas satisfaisant. Ce résultat pose question sur la validité de l'interprétation des scores sur ces deux subtests, puisqu'il ne semble pas y avoir qu'un seul trait latent derrière les performances de ces derniers dans notre échantillon. Seuls les subtests Cubes, Similitudes, Vocabulaire, Matrices et Compréhension ont pu être modélisés selon les MRI et ont été retenus pour l'étude du fonctionnement différentiel des items.

Au modèle unidimensionnel de base qui est postulé pour les subtests retenus, on a introduit un covarié à tour de rôle. Le premier lien direct et indirect examiné est celui entre la probabilité de réussir l'item en fonction de l'habileté sur le trait latent et l'âge de l'enfant. Comme attendu, un lien direct est montré. Plus un enfant est âgé, plus il réussit d'items. En effet, en grandissant, les enfants augmentent leur apprentissage et leurs connaissances. Lorsqu'on considère l'intelligence comme un trait stable, il s'agit d'une stabilité de la position de l'individu par rapport à son groupe de référence, et non de son niveau absolu d'intelligence. La variance expliquée par l'âge varie de 24.60 % (Cubes) à 45.83 % (Compréhension). Pour les subtests qui font appel à des capacités verbales, il y a une part importante dans la variance totale des scores au test qui est expliqué par l'âge. Nous supposons que les subtests de l'ICV mobilisent des connaissances qui sont directement liées aux activités à l'école. Des enfants d'une même tranche d'âge sont plus ou moins dans un même degré scolaire et ont une base commune d'acquisitions verbales. En revanche, les subtests de l'IRP mobilisent des habiletés moins directement rattachées à des activités scolaires. Les habiletés sur ce qu'évalue l'IRP sont alors plus dépendantes des opportunités d'apprentissage qui s'offrent aux enfants dans leur contexte de vie, ce qui explique une moindre influence de l'âge. Pour l'effet indirect, nous n'avions pas d'hypothèse sur un fonctionnement différentiel lié à l'âge et, en effet, les résultats confirment qu'il n'y a pas de biais d'item liés à cette variable pour tous les items des subtests considérés. Pour un même item, des enfants de différents âges ayant la même habileté sur le trait latent évalué par le subtest ont la même probabilité de le réussir.

Le deuxième covarié introduit était le sexe de l'enfant. Nous regardions d'abord s'il y a un lien direct entre le sexe et la probabilité de réussir l'item en fonction de l'habileté du sujet sur le trait latent. L'âge était également en covarié pour que les scores bruts d'enfants de différents âges soient néanmoins comparables. La variance expliquée par le sexe est faible pour tous les subtests considérés. Pour l'épreuve Cubes, il y a une influence significative mais très faible de la variable sexe en faveur des garçons, qui explique 1.44 % de la variance totale une fois l'effet direct de l'âge contrôlé. La moyenne des performances du groupe des garçons est légèrement plus élevée que la moyenne des performances du groupe des filles sur Cubes. Pour Matrices, il y a une influence significative mais très faible de la variable sexe en faveur des filles, qui explique 0.86 % de la variance totale une fois l'effet direct de l'âge contrôlé. La moyenne des performances du groupe des filles est légèrement plus élevée que la moyenne des performances du groupe des garçons sur Matrices. Néanmoins pour ces deux subtests, aucun FDI n'est détecté. Pour un même item, des enfants d'un sexe différent ayant la même habileté sur le trait évalué par le subtest ont la même probabilité de le réussir. Notre hypothèse selon laquelle les garçons obtiennent des meilleures performances sur les tâches visuospatiales tend à être légèrement confirmée, toutefois, cette légère différence de performances moyennes entre filles et garçons ne conduit pas à un biais d'item. En revanche, l'hypothèse que les filles sont meilleures dans les tâches verbales est infirmée. Dans les subtests de l'ICV, on ne relève aucun lien direct significatif.

Le troisième covarié considéré était le statut socio-économique des parents (SES). La variable SES est construite à partir de la profession des parents. À nouveau, nous avons d'abord examiné le lien direct entre SES et la probabilité de réussir l'item en fonction de l'habileté du sujet sur le trait latent. Pour Cubes, Similitudes, Vocabulaire, Matrices et Compréhension, il y a un effet direct de la variable SES, qui explique de 2.46 % (Matrices) à 8.35 % (Vocabulaire) une fois l'âge contrôlé. Les enfants ayant des parents avec les statuts socio-économiques les plus élevés ont des scores plus élevés que ceux ayant des parents avec les statuts socio-économiques les plus bas. Il y a un effet direct du statut socio-économique pour tous les subtests, et en particulier pour les subtests où les habiletés verbales sont fortement mobilisées. Toutefois, aucun FDI n'est détecté. Pour un même item, des enfants de différents SES ayant la même habileté sur le trait évalué par le subtest ont la même probabilité de le réussir. L'hypothèse selon laquelle, le milieu socio-économique a une influence dans les performances cognitives est confirmée. Les enfants dont les parents ont un SES élevé ont davantage

l'opportunité d'être stimulés intellectuellement par leur environnement. Ce résultat n'est pas surprenant ; de nombreuses études étayent le sujet. Par contre, il y a moins de données sur l'influence différenciée de chaque parent. Au lieu d'en faire une moyenne globale, nous avons également examiné l'influence directe de la contribution de la profession de chacun des deux parents dans les performances de l'enfant. Pour tous les subtests considérés, les résultats montrent un effet direct différencié entre la profession de la mère et la profession du père sur les performances de l'enfant. Si en effet, plus la mère (ou le père) exerce une profession prestigieuse, plus les performances de l'enfant sont élevées, la contribution de chaque parent sur les performances est statistiquement significative. Néanmoins, la taille de la différence entre la contribution de la profession de la mère et la contribution de la profession du père est négligeable et sans implication clinique.

L'absence d'items biaisés est importante pour la validité de l'interprétation d'un test. Aussi, au cours de l'adaptation d'un test, un soin particulier est porté à l'analyse des items et la détection d'éventuels biais. Dès lors, il n'est pas étonnant que nos résultats ne montrent aucun fonctionnement différentiel des items du WISC-IV en fonction des variables que nous avons considérées (âge, sexe et statut socio-économique). En revanche, une comparaison entre les enfants de la standardisation en français du test et les enfants suisses francophones aurait pu montrer certains biais d'items. Comme nous n'avons pas accès aux données de standardisation du WISC-IV, nous n'avons pas pu tester cette comparaison de première importance dans l'évaluation de l'équité d'un test qui s'adresse à différentes populations. D'autant que, d'après des observations, la plupart des enfants de notre échantillon montrent des difficultés pour reconnaitre l'image 5 du subtest Identification de concepts (voir Figure 54, p. 244). Dans Identification de concept, la consigne demande à l'enfant de trouver le concept commun qui relie un objet de chaque rangée.

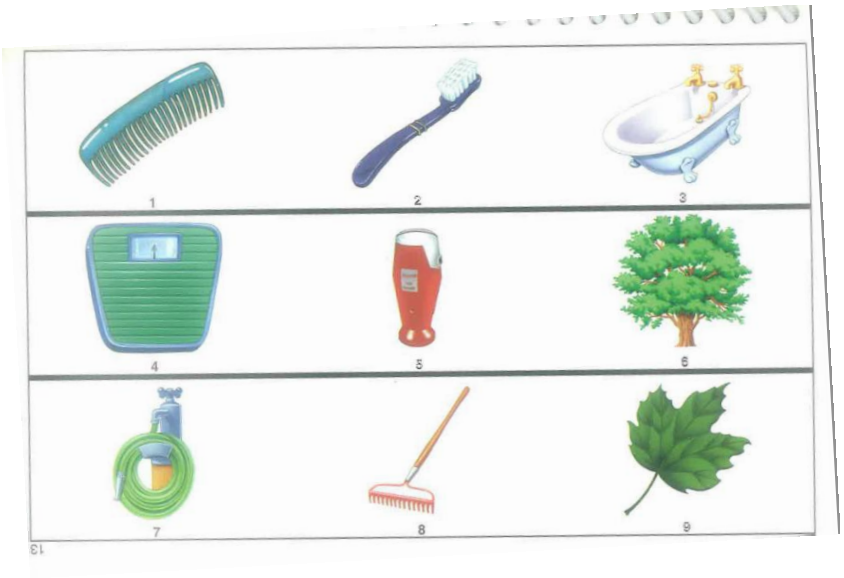


Figure 54. Item 13 du subtest Identification de concepts du WISC-IV.

Pour l'item 13, il y a trois rangées, il y a donc trois objets à choisir en tout. En l'occurrence, les objets à choisir sont la baignoire (image 3), la borne à incendie (image 5) et le tuyau d'arrosage (image 7). Ce sont trois objets qui peuvent contenir de l'eau. Or l'image 5 qui fait partie des bonnes réponses à sélectionner et peu reconnue comme une borne à incendie. Auprès des enfants suisses francophones, l'image 5 peut évoquer une lampe de poche, une bière, un thermos, un rasoir. Il en résulte que l'item 13 est réussi par 64 % des enfants de notre échantillon de 483 enfants, alors que les items 14 et 15 normalement plus difficiles sont réussis par 80 % des enfants de notre échantillon. À noter que les enfants qui réussissent l'item 13 sont souvent ceux qui ont demandé à l'examinateur ce que représente l'image 5. Le niveau d'habileté sur ce qu'évalue le subtest explique peu la réussite de l'item 13. On peut raisonnablement se demander si cette représentation d'une borne à incendie n'est pas plus familière aux enfants français qu'aux enfants suisses. De ce fait, la probabilité de réussir l'item est différente pour des enfants français et des enfants suisses ayant pourtant la même habileté sur le trait latent évalué par le subtest.

L'exemple de l'item 13 d'Identification de concepts souligne la pertinence d'étudier le fonctionnement différentiel des items dans la comparaison des groupes de sujets à qui s'adresse un test. Il est dommage qu'il n'y ait pas une volonté des maisons d'édition des tests de mettre les données de standardisation à disposition de la recherche. Nous l'avons mentionné la validation d'un test est processus continu et

enrichi par les données empiriques des études menées. L'utilisation, la compréhension et l'interprétation des scores d'un test gagnent en précision au fil des résultats de recherche.

Cette étude sur la détection d'un fonctionnement différentiel des items est motivée par un intérêt personnel survenu ultérieurement à la récolte des données. En effet, l'échantillon de 483 enfants est constitué à partir la récolte de données d'une précédente recherche sur la structure factorielle du WISC-IV et de notre recherche sur la stabilité des scores du WISC-IV. De ce fait, nous devons relever des limites aux résultats présentés. Une limite majeure est que les items des subtests ont été administrés selon la pratique usuelle de passation du WISC-IV. Les règles de départ, les règles de retour et les règles d'arrêt ont été appliquées. Tous les items de chaque subtest n'ont donc pas été administrés aux 483 enfants de l'échantillon. La règle de départ stipule à quel item démarrer le subtest en fonction de l'âge de l'enfant. Les items du début non administrés sont considérés comme réussis et on accorde les points de ces items. La règle d'arrêt stipule l'arrêt du subtest après un certain nombre d'items échoués consécutivement. Les items de fin non administrés sont considérés comme échoués et on les cote zéro point. Ainsi, dans une passation, tous les items d'un subtest ne sont pas administrés. Or, pour l'estimation des paramètres d'item (difficulté et discrimination) selon les modèles de réponse à l'item, il aurait fallu administrer tous les items à tous les enfants de l'échantillon. Nous avons supposé parfaitement réussis les items du début non administrés et complètement échoué les items de fin non administrés. Ce qui est une approximation qui conduit à réduire la variance sur ces items. Nous l'avons vu avec l'exemple de l'item 13 d'Identification de concepts, l'ordre de difficulté croissante des items du WISC-IV n'est pas tout à fait la même entre l'échantillon d'enfants suisses francophones et l'échantillon de standardisation en français. À cause de l'approximation sur les items non administrés, nous n'avons pas renseigné sur les paramètres d'item (difficulté et discrimination) qui sont ressortis de la modélisation.

Une autre limite majeure est les tailles de l'échantillon des individus et des items qui ne sont pas assez importantes pour conduire une estimation stable d'un modèle de réponse à l'item. Pour les modèles à deux paramètres que nous avons appliqués sur nos données (c.-à-d. le modèle 2 PL et le modèle gradué de Samejima), il faut disposer d'un test avec une longueur d'au moins 30 items et au moins 500 sujets. Hormis Vocabulaire (36 items) et Matrices (35 items), les subtests considérés n'ont pas un échantillon d'items assez important. De plus, à cause de la variance nulle et

d'intercorrélations entre paires d'items de faible variance qui enfreignent le postulat d'indépendance locale, des items ont dû être retirés des analyses, ce qui a encore plus réduit le nombre d'items. La taille de l'échantillon est également insuffisante pour stabiliser les estimations.

Les deux limites que nous avons évoquées résultent principalement du fait que la récolte des données n'a pas été effectuée pour l'étude du fonctionnement différentiel des items. Nos résultats sur le fonctionnement différentiel des items du WISC-IV sont donc exploratoires et demandent un approfondissement lors d'une étude ultérieure construite spécifiquement à cette fin. Néanmoins, ces premiers résultats nous permettent de souligner l'intérêt pour l'évaluation de l'équité dans l'évaluation psychologique.

10. LA STABILITÉ DU WISC-IV

L'étude de la stabilité à long terme des scores standards et CHC du WISC-IV porte sur un échantillon de 277 enfants non consultants qui ont été vus à deux reprises dans un délai test-retest variant d'un an à trois ans. La stabilité des scores est explorée à la fois aux niveaux interindividuel et intra-individuel. Jusqu'à présent, seule une étude sur la stabilité à court terme a été réalisée pour l'adaptation en français du WISC-IV. Pour des informations sur la stabilité à long terme des scores, il faut se tourner sur des études américaines : Lander (2010), Watkins et Smith (2013), et Bartoi et al. (2015). Ces trois études examinent des échantillons d'enfants suivis pour des difficultés d'apprentissage, émotionnelles ou d'attention. Bien qu'il soit intéressant d'étudier des groupes cliniques, il est également important d'avoir des données sur un groupe d'enfants tout-venant. Le groupe d'enfants tout-venant est supposé suivre un parcours scolaire normal en fonction de leurs habiletés cognitives. Ils ne sont pas pris en charge pour une problématique clinique liée aux apprentissages scolaires. Pour des comparaisons entre groupes, on a besoin d'un groupe tout-venant comme référent. Notre recherche avec un groupe non clinique fournit donc des données sur la stabilité du trait intelligence dans la population générale des enfants. De plus, la plupart des précédentes études ne décrivent la stabilité des scores qu'au niveau interindividuel au moyen de coefficients test-retest (stabilité différentielle). Dans son usage clinique, le WISC-IV est un outil qui aide à la prise de décision lors du bilan cognitif d'un enfant singulier. Il nous semble essentiel de fournir des données également au niveau intra-individuel pour aider les cliniciens dans leur prédiction sur des performances futures au WISC-IV.

Dans ce chapitre, nous allons discuter des résultats sur la stabilité à long terme des scores du WISC-IV. Les différents résultats sont résumés et leur implication développée. Au travers de divers questionnements, nous avons évalué la stabilité à long terme sous cinq angles : (1) la stabilité absolue, (2) la stabilité différentielle, (3) la stabilité intra-indviduelle absolue, (4) la stabilité catégorielle et (5) la stabilité des forces et faiblesses personnelles. Nous terminerons la discussion sur les limites de cette étude.

10.1. STABILITÉ ABSOLUE

Notre premier questionnement porte sur la stabilité absolue : est-ce que le niveau moyen des performances diffère entre la première et la seconde passation pour l'échantillon total ? Pour cela, nous déterminons, au niveau du groupe et par comparaisons de moyennes (*t*-tests pour échantillons appariés), si les moyennes de la première et la seconde passation sont équivalentes. Comme l'intelligence est supposé un trait stable sur le plan interindividuel, on ne s'attend pas à des différences de moyennes significatives entre les deux passations. Le WISC-IV est un test étalonné par tranche d'âges. Si l'on suppose la stabilité de l'intelligence, la position d'un individu par rapport à son groupe de référence ne bouge pas. Néanmoins, des recherches montrent généralement un effet d'apprentissage lors d'une procédure test-retest. Les effets d'apprentissage sont d'autant plus prononcés que l'intervalle test-retest est court. De plus, ils sont également plus prononcés sur les épreuves simples de vitesse de traitement que sur les épreuves verbales de vocabulaire ou de culture générale (Calamia et al., 2012; Estevis et al., 2012).

Les résultats des comparaisons de moyennes montrent une augmentation significative entre les deux passations pour l'Indice de Mémoire de Travail (+2.02 points), l'Indice de Vitesse de Traitement (+5.79 points), le QI Total (+2.53), l'Indice de Compétence Cognitive (+4.87 points), Gf (2.08 points), Gwm (+2.01 points) et Gs (5.51 points). Toutefois, pour l'IMT, Gf et Gwm, la taille d'effet est négligeable ($d < 0.2$), indiquant que la différence de moyenne n'est pas assez importante pour avoir une pertinence clinique. En revanche, pour l'IVT, le QIT, l'ICC et Gs, la taille d'effet indique un effet d'apprentissage petit à modéré ($0.2 < d < 0.5$). Les subtests Code (+0.82), Matrices (+0.48) et Symboles (+1.09) montrent une augmentation significative. La taille d'effet suggère un effet d'apprentissage petit à modéré pour Code et Symboles.

L'hypothèse selon laquelle les effets d'apprentissage tendent à diminuer avec la durée de l'intervalle test-retest se voit confirmée dans nos données. En effet, les différences de moyennes sont moins importantes que celles mises en évidence dans les études à court terme. Par exemple, dans les études à court terme, la différence de moyennes entre les deux passations est de +8.3 points pour le QIT, de +3.6 points pour l'IMT et de +12 points pour l'IVT (Wechsler, 2005b). Cette observation déjà relevée sur le WISC-III par Canivez et Watkins (1999, 2001) leur fait déclarer que les effets d'apprentissage sont faibles à négligeables au-delà d'un intervalle retest d'une année. Sur cette base, nous avions fixé le délai test-retest minimum à 1 an. Cependant, nos

résultats montrent une augmentation significative des performances moyennes à la seconde passation pour l'IMT, l'IVT, le QIT, l'ICC, Gf, Gwm et Gs ainsi que pour les subtests Code, Matrices et Symboles. Ce résultat peut s'expliquer par la différence entre un échantillon clinique et un échantillon non consultant. De nombreux travaux sur des tests cognitifs montrent qu'un effet d'apprentissage est à la fois relevé et plus prononcé chez des sujets sains que chez des sujets patients, où il est parfois même absent (Cooper et al., 2004; Duff et al., 2008, 2001). Avec le WISC-IV et des intervalles supérieurs à 1 an (long terme), les études sur des échantillons cliniques ne montrent pas de différences de moyennes significatives dues à un effet d'apprentissage pour les indices entre les deux passations (Lander, 2010; Watkins & Smith, 2013).

Pour notre échantillon d'enfants tout-venant, la question se pose alors de la durée du délai test-retest qui élimine les effets d'apprentissage pour tous les scores. Pour répondre à cela, une analyse postérieure est réalisée dans laquelle nous avons découpé les durées test-retest de manière suivante : intervalles de 12 à 15 mois (N = 90 enfants), intervalles de 16 à 24 mois (N = 92 enfants) et intervalles de 25 à 39 mois (N = 95 enfants). Les résultats montrent que pour l'IMT, le QIT, Gf, Code et Matrices, un délai supérieur à deux ans est requis pour éliminer les effets d'apprentissage, tandis qu'un délai de plus de trois ans est requis pour l'IVT, l'ICC, Gs et Symboles. Les épreuves qui impliquent de la vitesse de traitement présentent des effets d'apprentissage sur une plus longue période de temps.

D'après les études sur le WISC-IV avec un intervalle inférieur à une année (court terme), les effets d'apprentissage s'observent chez les enfants tout-venant de manière plus prononcée pour l'IRP et l'IVT que pour l'IMT et l'ICV (Ryan et al., 2010; Wechsler, 2005b). Ce résultat se retrouve partiellement dans nos données. En effet, les effets d'apprentissage s'observent de façon plus prononcée sur les scores des indices IMT et IVT. On peut formuler l'hypothèse de la « réduction de la nouveauté » serait en partie à l'origine des effets d'apprentissage. Si lors de la première passation, l'absence de feedback sur la réussite ou l'échec d'un item limite l'effet d'apprentissage, particulièrement pour les subtests de l'ICV, les enfants se familiarisent toutefois avec la situation et le matériel de passation. La familiarité avec la situation et le matériel est un des bénéfices d'une première passation. Pour le subtest Matrices de l'IRP, l'augmentation significative des performances moyennes à la seconde passation peut s'expliquer par le souvenir des stratégies de résolution. De même, pour les subtests de l'IVT, le souvenir des stratégies à appliquer peut faciliter l'intégration de la consigne et l'exécution de la tâche. De plus lors de la seconde passation, les enfants sont en

moyenne plus avancés en âge et dans leur scolarité. Habitués à diverses formes d'évaluation à l'école et mieux conscients des enjeux du chronométrage, ils peuvent bénéficier de l'expérience d'une passation initiale et mieux comprendre l'importance de la vitesse d'exécution que les plus jeunes. Pour les subtests de l'IMT, nous avons observé chez les plus jeunes enfants de notre échantillon des difficultés dans la compréhension de la longue consigne du subtest Séquence Lettres-Chiffres. La tâche qui leur est demandée consiste à redonner une séquence en ordonnant d'abord les chiffres du plus petit au plus grand, puis les lettres dans l'ordre alphabétique. Les enfants de 7 ans, en particulier, s'embrouillent et échouent assez vite dès qu'ils doivent ordonner plusieurs chiffres en ordre croissant et plusieurs lettres dans l'ordre alphabétique. En revanche lors de la seconde passation, étant plus âgés, ils réussissent beaucoup mieux à intégrer la longue consigne. Dans l'ensemble, les effets d'apprentissage résultent de l'effet combiné de la familiarité avec les activités et d'une plus grande conscience de la situation de testing. En effet, à la seconde passation, les enfants sont plus âgés et plus avancés dans leur parcours scolaire. Ils apparentent davantage la passation du WISC-IV à une situation d'évaluation scolaire. Pour appuyer notre explication, on peut rappeler que d'après Flangan et Kaufman (2009), les gains dus à l'effet d'apprentissage sont plus importants pour les enfants âgés de 6-7 ans à la passation initiale et qu'ensuite les gains diminuent avec l'âge de la première passation.

Pour mieux comprendre la relation entre durée de l'intervalle test-retest, âge initial et différences de performances aux indices, nous avons réalisé une série de corrélations entre ces trois variables. Nous aimerions en particulier revenir sur un résultat, qui va à l'encontre de ce qui est relevé dans l'étude de Ryan et al. (2010). Avec un délai test-retest moyen de onze mois et un échantillon d'enfants tout-venant, l'étude de Ryan et al. montre que les enfants avec les meilleures performances à la première passation du WISC-IV bénéficient davantage d'une seconde passation que les enfants les moins performants. À al différence de Ryan et al., nous trouvons que ce sont les enfants avec les QIT les plus bas à la passation initiale qui tendent à tirer davantage bénéfice de cette première expérience de passation et qui obtiennent des gains plus importants à la seconde passation. Une explication est le phénomène de régression à la moyenne qui tend à ramener les scores vers la moyenne lors de mesures répétées. Ainsi, les scores en dessous de la moyenne s'élèvent vers la moyenne, tandis que les scores au-dessus de la moyenne descendent vers la moyenne.

Les résultats de l'évaluation de la stabilité absolue rendent attentifs à l'effet d'apprentissage. Dans la pratique de l'évaluation, il peut arriver qu'une réévaluation soit

demandée. Nous recommandons un délai d'au moins un an entre deux administrations d'un WISC-IV pour limiter les effets d'apprentissage. Toutefois, sur des épreuves de vitesse de traitement en particulier, l'effet d'apprentissage peut se manifester malgré le délai d'un an. De façon plus précise, il faut être attentif aux subtests du WISC-IV qui sont sensibles à un effet retest (c.-à-d., Matrices, Code et Symboles). Notre échantillon d'enfants tout-venant montre que la présence d'un effet d'apprentissage est attendue lors d'évaluations ultérieures. Il s'agit d'un indicateur positif, même s'il ne doit pas être interprété comme une augmentation réelle dans le niveau d'habileté.

10.2. STABILITÉ DIFFÉRENTIELLE

Notre deuxième questionnement porte sur la stabilité différentielle : est-ce que les individus se classent dans le même ordre entre la première et la seconde passation ? Pour cela, des corrélations entre les scores à la première et à la seconde passation sont réalisées pour tous les indices et tous les subtests. Dans la plupart des études, la stabilité des scores d'un test est définie par des coefficients de stabilité. L'hypothèse de la stabilité de l'intelligence est d'ailleurs appuyée par des recherches longitudinales qui calculent des coefficients de corrélation entre au moins deux temps de mesure. Il s'agit donc d'une stabilité interindividuelle, c'est-à-dire de la position de l'individu par rapport à son groupe de référence.

Avant de résumer nos résultats, nous allons rappeler la signification d'un coefficient de stabilité. Ce dernier traduit les différences interindividuelles dans les scores d'un test. On peut directement l'interpréter en termes de variance vraie et de variance d'erreur. Par exemple, un coefficient de .70 signifie que 70 % de la variance dans les scores observés sont de la variance vraie, tandis que les 30 % restants sont de la variance d'erreur. Précision que la fidélité ne permet pas de distinguer dans la variance vraie ce qui relève de la variance pertinente due à la propriété mentale censée être évaluée par le test et ce qui relève de la variance non pertinente due à toutes les autres choses (erreur systématique incluse) qu'évalue aussi le test et qui n'est pas la propriété mentale sur laquelle porte l'interprétation du test. La distinction entre la part de variance pertinente et la part de variance non pertinente relève de l'évaluation de la validité.

Un coefficient de fidélité peut également s'interpréter en termes de changement de rang. Si la fidélité des scores est parfaite, le classement des individus

reste dans le même ordre d'une passation à l'autre. Nous avons présenté l'étude de Thorndike et Hagen (1977) qui présentent le pourcentage de changement de rang pour un individu ou dans un groupe en fonction du coefficient de fidélité. Reprenant l'étude de Thorndike et Hagen, Bernier et Pietrulewicz (1997) proposent une représentation graphique (voir Figure 55, p. 252). On peut voir qu'un individu a 35 % de chance de changer de rang avec un test de fidélité à .60. Il a 30 % de chance de changer de rang si la fidélité est de .70, en revanche, un individu a moins de 10 % de chance de changés de rang si la fidélité est à .90.

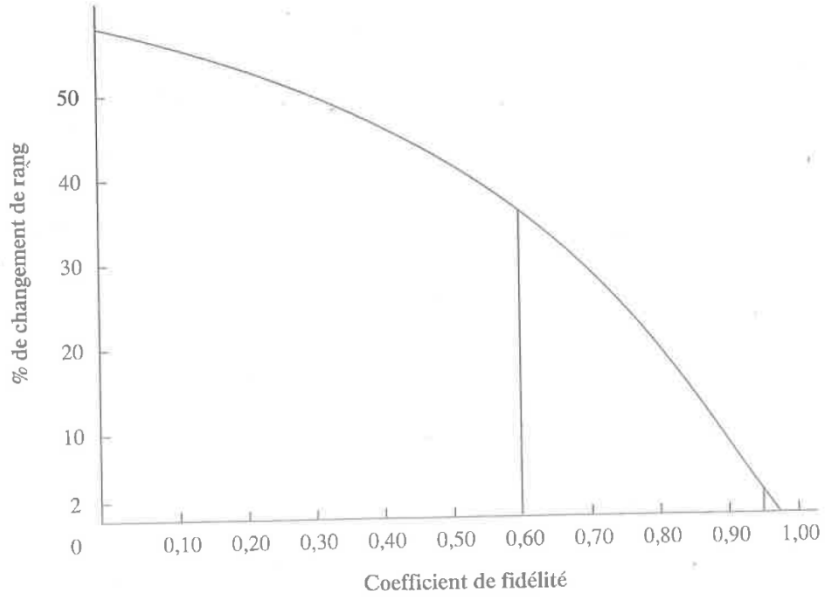


Figure 55. Courbe du changement de rang pour un individu en fonction d'un coefficient de fidélité. Source : Bernier et Pietrulewicz (1997, p. 137)

Avec l'étude de Charter et Feldt (2001), nous avons également présenté l'interprétation du coefficient de fidélité en relation avec les risques d'erreur dans la prise de décision. Ces auteurs montrent que dans le cas d'un test de fidélité parfaite (r = 1), le test permettrait de déceler 10 % de vrais positifs (individus correctement identifiés comme ayant besoin d'une prise en charge clinique), 90 % de vrais négatifs (individus correctement identifiés comme n'ayant pas besoin d'une prise en charge clinique) et, 0 % de faux positifs et de faux négatifs (individus identifiés à tort comme ayant ou n'ayant pas besoin d'une prise en charge clinique). Dans ce cas idéal, le test permettrait donc de prendre 100 % de décisions correctes. À l'issue de l'analyse de plusieurs coefficients de fidélité, Charter et Feldt (2001) concluent qu'une valeur de

fidélité de .98 ou plus est souhaitable pour prendre 90 % de décisions correctes et donc avoir un risque d'erreur de 10 %.

Le rappel de ces différentes manières d'interpréter un coefficient de fidélité nous permet de souligner qu'on ne peut pas s'épargner une réflexion sur la signification d'un coefficient de fidélité et sur le risque d'erreur qu'on tolère. Dans la littérature, on trouve des seuils qui décrètent qu'un coefficient supérieur à .70 représente une fidélité des scores satisfaisante dans le contexte de recherche ou pour une prise de décision au niveau d'un groupe. Pour des décisions dans un contexte clinique, on demande une fidélité des scores de .80, voire .90. Ces seuils, bien qu'utiles pour avoir un ordre d'idée, ne sont pas à appliquer mécaniquement. Selon le contexte et les enjeux de l'évaluation cognitive, le psychologue doit évaluer le risque d'erreur qu'il tolère.

Les résultats des corrélations test-retest montrent que les coefficients les plus faibles pour les scores de l'Indice de Mémoire de Travail, l'Indice de Vitesse de Traitement, Gf, Gwm et Gs. Dans l'ensemble, aucun indice n'atteint le seuil de .90. Néanmoins, le QI Total et l'Indice d'Aptitude Générale ont des coefficients de fidélité supérieurs à .80.

Dans une perspective de recherche ou pour des décisions sur des performances moyennes, le pourcentage de changements de rang est faible pour l'ICV, l'IRP, le QIT, l'IAG, l'ICC, Gc et Gv. En revanche, dans une perspective clinique et pour des décisions sur des scores individuels, il y a un pourcentage de changement de rang proche de 20 % pour le QIT et l'IAG qui sont pourtant les indices présentant les plus hauts coefficients de fidélité.

S'agissant des scores du WISC-IV, leur utilisation principale est clinique et conduit à une prise de décision souvent à fort enjeu pour l'avenir d'un enfant. Hormis pour le QI Total (r = .83) et l'Indice Aptitude Générale (r = .83), nous ne pouvons pas recommander des prédictions sur le long terme pour la plupart des indices du WISC-IV. En particulier pour l'Indice de Mémoire de Travail (r = .66), l'Indice de Vitesse de Traitement (r = .67), Gf (r = .60), Gwm (r = 66) et Gs (r = .69) qui ne présentent pas des différences interindividuelles suffisamment stables sur le long terme. Ceci étant dit, les prédictions sur les performances futures basées sur QIT et l'IAG doivent néanmoins être faites avec prudence.

Les coefficients de fidélité servent dans le calcul d'un intervalle de confiance. Les concepteurs du WISC-IV ont construit leur intervalle de confiance sur la base de

coefficients de fidélité estimés par la méthode du partage. Exception faite des subtests de l'IVT pour qui la méthode est le test-retest avec un intervalle à court terme. Nous avons mentionné que la méthode du partage est une évaluation de la consistance interne d'un test, et non de la fidélité de ses scores. De plus, les coefficients test-retest sont plus faibles que les coefficients de consistance interne qui, eux, ne tiennent pas compte des sources d'erreur liées à la temporalité. De ce fait, les intervalles de confiance proposés par les concepteurs du WISC-IV sont plus étroits que s'ils étaient basés sur des coefficients test-retest à long terme. Au vu de nos résultats sur une étude longitudinale, il nous semble plus adéquat pour les prochaines éditions des Échelles de Wechsler de construire les intervalles de confiances avec des coefficients test-retest à long terme, mais cela paraît peu praticable pour des raisons de coût et de temps.

10.3. STABILITÉ INTRA-INDIVIDUELLE

Notre troisième questionnement porte sur la stabilité intra-individuelle : quelle est la stabilité des scores individuels ? Pour explorer le niveau intra-individuel, nous avons mené trois analyses : l'évaluation de la stabilité intra-individuelle absolue, l'évaluation de la stabilité catégorielle et l'évaluation de la stabilité des forces et faiblesses personnelles. Dans les études sur la stabilité des scores du WISC, les résultats au niveau intra-individuel sont peu développés, voire omis. Pourtant, dans l'utilisation clinique du WISC, les décisions prises sont basées sur les scores individuels d'un individu. Pour pouvoir fonder des prédictions sur les performances futures de l'individu, la stabilité intra-individuelle des scores est requise. De plus, la stabilité des indices au niveau interindividuel n'implique pas une stabilité des indices au niveau intra-individuel (Voelkle, Brose, Schmiedek, & Lindenberger, 2014). En effet, un indice stable sur le plan interindividuel peut être moins stable sur le plan intra-individuel, et inversement. D'ailleurs, les résultats sur la stabilité absolue et la stabilité différentielle ne vont pas non plus de pair. Rappelons que la stabilité absolue concerne les différences de moyenne entre les deux passations, tandis que la stabilité différentielle concerne le classement des individus. Un indice peut présenter une différence de moyenne, notamment due à un effet d'apprentissage, et néanmoins classer les individus dans le même ordre d'une passation à l'autre si, par exemple, tous les individus voient leur score augmenter de cinq points. Dans nos données, c'est le cas du QI Total. Il présente une augmentation de moyenne significative entre les deux passations, toutefois, le classement des individus sur cet indice est relativement stable entre la première et la

seconde passation. Le niveau intra-individuel détermine une différence de performances entre les deux passations à partir de laquelle on considère qu'il y a un changement dans les performances d'un individu. Nous avons choisi trois angles qui nous semblent particulièrement pertinents dans la pratique clinique du WISC-IV.

Tout d'abord, nous avons évalué ce que nous appelons la stabilité intra-individuelle absolue. Pour cela, nous définissons un intervalle de points dans lequel nous considérons qu'il n'y a pas de changement significatif de performances. L'intervalle est construit à l'aide de l'erreur type de mesure des scores du WISC-IV (Annexe F). Pour rappel, dans la théorie classique des tests (TCT), tout score observé à un test (X) résulte du score vrai de l'individu (V) et de l'erreur de mesure (E). Le score vrai est l'espérance mathématique des scores observés. Cela signifie que si on pouvait répéter une infinité de fois la mesure sur un même individu, la moyenne de ses scores observés correspondrait à son score vrai. La TCT postule également que, pour les mesures répétées d'un même test par un même individu, les scores observés de l'individu se distribuent normalement autour de son score vrai. Sur un grand nombre de mesures, l'écart type de cette distribution peut être calculé pour un individu. De même que pour tous les individus d'un échantillon à qui on a fait passer de façon répétée le même test. Une fois calculé l'écart type de la distribution des scores observés pour chaque individu de l'échantillon, on peut calculer la moyenne des écarts types de l'échantillon, et cette moyenne d'écarts types est l'erreur type de mesure (ETM) pour un test. Par construction, 68 % des variations de performances pour un score doivent être incluses dans l'intervalle de ±1 ETM à la seconde passation, 95 % des variations de performances pour un score doivent être incluses dans l'intervalle de ±2 ETM à la seconde passation et 99 % des variations de performances pour un score doivent être incluses dans l'intervalle de ±3 ETM à la seconde passation. Dans la littérature, l'intervalle de ±2 ETM est fréquemment utilisé, c'est donc avec cet intervalle que nous avons considéré qu'un indice présentant une stabilité intra-individuelle pour au moins 80 % des enfants est acceptable. Le critère de 80 % permet des prédictions correctes pour 4 enfants sur 5 et donc, cela signifie également que la prédiction est incorrecte pour 1 enfant sur 5.

Dans l'ensemble, aucun indice ne présente une stabilité intra-individuelle à long terme selon notre critère. Les résultats montrent qu'au moins 70 % des enfants maintiennent des performances dans un intervalle de ±2 ETM pour l'Indice de Compréhension Verbale, l'Indice de Raisonnement Perceptif, l'Indice d'Aptitude Générale, Gf et Gv entre les deux passations.

Comme nous l'avons relevé, les coefficients de fidélité proposés par les concepteurs du WISC-IV sont estimés par la méthode du partage, qui produit des coefficients plus élevés que la méthode test-retest. Par conséquent, les ETM sont plus petits que s'ils sont calculés avec des coefficients test-retest à long terme. Par exemple pour le QI Total, ±2 ETM correspondent à un intervalle de ± 7.26 points avec les données du Manuel que nous avons utilisées. En revanche, si nous avions utilisé les coefficients de fidélité de notre échantillon, ±2 ETM correspondent à un intervalle de ±12.37 points ($ETM = 15\sqrt{1 - .83}$). Dans l'Annexe H, nous pouvons voir que pour le QIT, environ 87 % des enfants de notre échantillon présentent une différence de performances inférieure à 13 points entre les deux passations. Comme les coefficients test-retest avec un intervalle à long terme sont rarement estimés, nous recommandons aux cliniciens de consulter les données sur les distributions des scores de différences pour évaluer les risques de changement de performances.

Dans une deuxième analyse, nous avons évalué la stabilité intra-individuelle sous l'angle de la stabilité des catégories qualitatives. Pour transmettre les résultats du WISC-IV, les cliniciens peuvent décrire les scores de manière qualitative au lieu de numérique. La lecture descriptive fréquemment adoptée est celles en sept catégories : très faible (≤ 69), limite (70-79), moyen faible (80-89), moyen (90-109), moyen fort (110-119), supérieur (120-129), et très supérieur (≥ 130). Moins fin, le système descriptif en trois catégories s'apparente à une lecture normative par rapport à l'écart à la moyenne : faible (≤ 84), dans la moyenne (85-115), et élevé (≥ 116). Dans un entre-deux se trouve la lecture en cinq catégories : extrémité inférieure (≤ 69), moyen faible (70-84), dans la moyenne (85-115), moyen fort (116-130), et extrémité supérieure (≥ 131).

Les résultats montrent que la lecture en 5 catégories propose le meilleur équilibre entre finesse descriptive et stabilité intra-individuelle à long terme. Le QI Total et l'Indice d'Aptitude Générale présentent une stabilité catégorielle chez 76 % des enfants entre les deux passations. L'autre résultat important de cette analyse est le phénomène de régression à la moyenne qui se traduit par une plus grande probabilité des mesures répétées à se rapprocher de la moyenne plutôt qu'à s'en éloigner. En effet, lorsque les performances des indices changent de catégorie d'une passation à l'autre, il s'agit généralement d'un changement en direction de la moyenne de la distribution, c.-à-d. la valeur de 100.

Nous nous sommes intéressée à explorer les changements de catégorie pour les trois indices globaux, à savoir le QIT, l'IAG et l'ICC. Notre intérêt porte plus particulièrement sur le sous-échantillon d'enfants qui à la passation initiale obtiennent

un QIT, un IAG ou un ICC inférieur à 85 (soit une performance à un écart type en dessous de la moyenne). Le sous-échantillon est constitué de 35 enfants pour le QIT, de 30 enfants pour l'IAG et de 47 enfants pour l'ICC. Un indice inférieur à 85 décrit des performances faibles à très faibles. Pour ce groupe d'enfants susceptibles de se voir proposer une intervention pour certaines de leurs difficultés lors d'une évaluation cognitive, quelle prédiction peut-on faire à long terme sur leurs performances futures ? Nos résultats montrent qu'en seconde passation, 22 sur 35 enfants pour le QIT (soit 63 %) changent pour la catégorie des performances dans la moyenne (entre 85 et 115). Pour l'IAG, ce sont 14 sur 30 enfants (soit 47 %) qui changent pour la catégorie des performances dans la moyenne, tandis que, pour l'ICC, ce sont 31 sur 47 enfants (soit 66 %) qui changent pour la catégorie des performances dans la moyenne. Plus de la moitié des enfants voient leur performance remonter vers la moyenne lors d'une seconde passation. En revanche, pour les enfants qui à la première passation obtiennent un QIT, un IAG ou un ICC entre 85 et 115 (performances dans la moyenne), 87 % (QIT), 88 % (IAG) et 77 % (ICC) restent dans des performances entre 85 et 115 à la seconde passation. Pour les enfants ayant des QIT, IAG ou ICC supérieurs à 115 (soit une performance à un écart type en dessus de la moyenne), 71 % (QIT), 68 % (IAG) et 50 % (ICC) gardent des performances supérieures à 115. Nous pouvons relever deux constats à partir de ces résultats. Le premier constat est qu'un phénomène de régression à la moyenne s'observe sur les scores du WISC-IV. Le second constat est que ce phénomène de régression à la moyenne touche davantage les enfants dont les performances initiales sont faibles que ceux dont les performances initiales sont élevées (> 115). Cela s'explique sans doute parce que le phénomène de régression à la moyenne est couplé à l'effet d'apprentissage. Étant donné le phénomène de régression vers la moyenne, une stabilité catégorielle n'est pas supposée pour les enfants avec des performances faibles ou élevées. En revanche, on s'attend à une stabilité des performances dans la moyenne entre deux passations.

Dans une troisième analyse, nous avons évalué la stabilité intra-individuelle sous l'angle de la stabilité des forces et faiblesses personnelles. Dans la stabilité catégorielle, les performances de l'enfant sont comparées aux performances de son groupe (comparaisons normatives). Il est également intéressant de comparer les performances de l'enfant par rapport à lui-même (comparaison ipsative). Pour élaborer des pistes d'intervention, le psychologue s'appuie autant sur les domaines cognitifs de force et/ou de faiblesse de l'enfant par rapport aux autres enfants de son âge (forces et faiblesses normatives) que par rapport à lui-même (forces et faiblesses personnelles). Ces deux

niveaux de comparaison ne vont pas forcément de pair. Par exemple, un enfant peut présenter des performances toutes inférieures par rapport à son groupe d'âge (faiblesse normative), et néanmoins présenter une force personnelle sur tel indice par rapport à la moyenne de ses indices.

Les résultats montre qu'à la première passation, l'Indice de Compréhension Verbale est l'indice le plus souvent en force personnelle (FoP) dans le profil des 277 enfants de notre échantillon, tandis que l'Indice de Mémoire de Travail est le l'indice le plus souvent en faiblesse personnelle (FaP). À la seconde passation, l'Indice d Vitesse de Traitement devient l'indice le plus souvent en force personnelle, tandis que l'Indice de Mémoire de Travail est toujours l'indice le plus souvent en faiblesse personnelle. Si on examine la stabilité des forces et/ou faiblesses personnelles, l'ICV présente la meilleure stabilité à long terme, avec 61 % pour FaP et 56 % pour FoP. Pour l'IVT, l'indice présente une stabilité entre les deux passations chez 69 % d'enfants lorsqu'il est force personnelle. En revanche, lorsqu'il est une faiblesse personnelle, il ne présente une stabilité entre les deux passations que chez 30 % des enfants. Pour tous les quatre indices, les moyennes personnelles présentent une stabilité à long terme entre les deux passations chez environ 80 % des enfants. À nouveau, le phénomène de régression à la moyenne explique en grande partie les résultats. Dans les comparaisons ipsatives, la stabilité des forces et des faiblesses personnelles n'est a priori pas supposée. Par contre, on peut supposer une stabilité pour les performances dans la moyenne personnelle entre deux passations.

Dans la pratique clinique, le Manuel fournit pour chaque valeur d'une note QI un intervalle de confiance à 90% ou à 95%. L'étendue de l'intervalle au niveau de confiance à 95% correspond à peu près à l'étendue de l'intervalle à ±2 ETM que nous avons présenté. Nos résultats montrent dès lors que même si on n'utilise pas le score unique, mais l'intervalle de confiance de 95% autour du score, les prédictions ne sont pas pour autant stables sur le long terme. Pour le QIT qui repose sur la performance à plusieurs subtests, on tombe juste seulement pour 59.9% des enfants. Cela relativise les prédictions que permettent les scores au WISC-IV, mais surtout cela permet de souligner l'importance de recueillir des sources diverses et multiples d'information lors d'une évaluation. Un test seul n'est pas suffisant pour prendre une décision avisée. De plus, les résultats à plusieurs tests ne sont pas non plus suffisants s'ils ne sont pas soutenus par des observations cliniques, des entretiens ou des éléments d'anamnèse. La batterie du WISC-IV fournit une base appréciable de (sub)tests, néanmoins chacun d'eux n'est pas une mesure pure de la propriété mentale. Lorsque des différences de

performances entre des subtests qui contribuent au même indice s'observent, il est pertinent de compléter avec la passation d'autres tests pour affiner les hypothèses. Dans ce but, l'approche du cross-battery (XBA) permet d'intégrer les résultats à différents tests au moyen de la grille de lecture du modèle des aptitudes cognitives de Cattel-Horn-Carroll (CHC). Si l'approche XBA n'est pas encore répandue dans la pratique, nous la recommandons vivement. Les erreurs de mesure sur les tests sont inévitables et difficiles à contrôler. Nous le rappelons, il est essentiel de pouvoir s'appuyer sur des sources d'information diverses et multiples pour l'interprétation des résultats d'un test. Un test psychologique n'est pas à considérer comme un instrument de mesure, mais bien instrument d'évaluation qui repose sur des inférences. À partir de l'observation de réponses/comportements sur une sélection d'items, on calcule un score qui est interprété comme reflétant un niveau d'habileté sur une propriété mentale. Nous n'avons pas directement accès à la propriété mentale et bien souvent nous ignorons à quoi le sujet a fait appel pour réaliser la tâche. Il s'agit au psychologue d'avoir lui-même une réflexion poussée sur la bonne pratique des tests. Les demandes administratives se limitent souvent à une valeur de QIT ou des résultats à un seul test. Pour autant, le psychologue ne doit pas s'en tenir à ce qui lui est demandé, mais savoir réaliser une évaluation qui lui permet de prendre une décision réfléchie pour un individu particulier.

Dans l'ensemble, les résultats sur la stabilité des scores du WISC-IV montrent qu'il est difficile de déterminer si les scores du WISC-IV sont stables ou non. Il faut réfléchir sur le contexte et la finalité de l'utilisation des scores. Dans le court terme, les scores du WISC-IV augmentent à cause de l'effet de l'apprentissage d'une passation à l'autre (stabilité absolue), toutefois, le classement des individus présente un ordre équivalent d'une passation à l'autre (stabilité différentielle). Dans le long terme, l'effet d'apprentissage diminue et il apparaît un phénomène de régression à la moyenne qui tend à ramener les scores vers la moyenne. Sur le plan interindividuel, aucun score du WISC-IV n'atteint un seuil assez élevé pour qu'on puisse baser des décisions individuelles sur un seul score. Même pour le score de QIT qui est souvent utilisé dans des critères administratifs ou diagnostiques. Sur le plan intra-individuel qui intéresse en premier lieu l'utilisation clinique du WISC-IV, nos résultats montrent qu'un pourcentage non négligeable d'enfants (30-40%) voit leurs performances varier d'une passation à l'autre. La restitution d'un score unique ou d'un score associé à un intervalle de confiance devient délicate. Finalement, est-ce vraiment utile aux parents ou à l'enfant d'avoir une valeur numérique si celle-ci est sujette à variation dans le temps ? Nous

sommes tentée de recommander de restituer dans le compte rendu aux parents ou à l'enfant uniquement une description qualitative, et non une valeur numérique. D'ailleurs, la valeur numérique peut focaliser à mauvais escient l'attention si elle est basse ou très élevée. De plus, elle n'est pas toujours bien comprise dans le public, car pour l'interpréter, il faut avoir en tête l'image de la distribution normale et connaître la moyenne et que l'écart type de la distribution des QIs. Étant donné les erreurs de mesure, il nous semble judicieux dans la transmission des résultats de mettre en avant une interprétation des scores qui se détache d'une valeur numérique.

Dans la perspective d'utiliser des catégories descriptives, l'analyse de la stabilité catégorielle montre que la classification en cinq catégories permet une certaine finesse descriptive et présente une certaine stabilité. Cependant, la classification en trois catégories peut aussi être utilisée. Dans l'analyse catégorielle, un des résultats intéressants pour la pratique est la mise en évidence de l'action des effets d'apprentissage et de la régression à la moyenne en fonction du niveau de performance initiale. Pour les enfants ayant des performances à un écart type en dessous de la moyenne à la première passation, ces deux effets agissent dans le même sens et la majorité des enfants voient leur performance remonter vers la moyenne. Leurs performances décrites comme faibles à la première passation remontent dans la catégorie dans la moyenne à la seconde passation. Les enfants ayant des performances dans la moyenne à la première passation restent pour une grande majorité dans cette même catégorie. Les faibles effets d'apprentissage et le faible effet de régression à la moyenne influencent peu dans cette situation. Les enfants ayant des performances à un écart type au-dessus de la moyenne restent pour la majorité dans cette même catégorie à la seconde passation, bien qu'il y ait un certain pourcentage qui voit leurs performances descendre dans la catégorie dans la moyenne. Les effets d'apprentissage et de régression à la moyenne agissent en opposition dans cette situation. Ainsi, selon la performance de l'enfant, on peut prédire une certaine ampleur de l'influence des effets d'apprentissage et de régression à la moyenne.

Dans cette dernière partie de discussion, nous allons relever les limites de notre étude. À l'instar de l'étude sur le fonctionnement différentiel des items, l'échantillon de l'étude de la stabilité des scores soulève également des limites. Ce n'est pas en raison de la taille de l'échantillon qui est adéquate pour les analyses réalisées. La première limite porte sur la restriction de l'étendue des âges aux enfants de 7 à 12 ans, alors que le WISC-IV s'adresse aux enfants de 6 à 16 ans. Cette restriction aux enfants du cycle d'enseignement primaire (CEP) est dictée par des contraintes à la fois d'autorisation de

recherche et de praticabilité. Pour recruter des enfants plus âgés, nous aurions dû demander une autorisation de recherche pour le cycle d'enseignement obligatoire (CEO), ce qui double les démarches administratives. Sur le plan pratique, les enfants changent d'établissement scolaire lors du passage du CEP au CEO, ce qui rend malaisé de les retrouver pour la seconde passation. Pour augmenter la possibilité de voir et de revoir un nombre important d'enfants, nous avons restreint à la tranche d'âge des 7-12 ans. Nos résultats se limitent donc à cette tranche d'âge.

Une deuxième limite également liée à l'échantillon est qu'il s'agit d'enfants provenant uniquement d'écoles dans le canton de Genève. Néanmoins, étant donné une certaine harmonisation dans le système scolaire suisse, on peut supposer que nos résultats peuvent plus largement se généraliser aux enfants suisses romands de 7 à 12 ans. Une troisième limite aussi liée à l'échantillon est que les enfants issus de milieux socio-économiques élevés sont légèrement surreprésentés, tandis que les enfants issus de milieux socio-économiques bas sont légèrement sous-représentés. Étant donné les difficultés pour recruter des enfants en grand nombre, il est difficile de contrôler certaines variables. Toutefois, notre échantillon présente une représentativité entre filles et garçons ainsi que sur le niveau des performances moyennes et sur la variabilité intragroupe (moyennes et écarts types proches des valeurs théoriques de la distribution des QI et des notes standards des subtests).

Une autre variable que nous n'avons pas pu contrôler est le délai test-retest. La passation durant les heures scolaires demande la collaboration des directeurs d'écoles et surtout des enseignants. Nous avons dû nous adapter aux disponibilités proposées en fonction des activités de la classe. De plus entre la première prise de contact par téléphone ou par mail et la venue effective pour les passations, il s'écoule un temps imprévisible. Avant d'accéder aux enfants, nous devons passer par le directeur d'école, puis les enseignants et enfin les parents. Cela rend impossible de tenir une planification contrôlée pour la durée des intervalles test-retest. Nous avons donc échelonné les passations du retest pour qu'il y ait au minimum 12 mois de délai entre les deux passations. Du fait que nous n'avons pas pu contrôler les intervalles test-retest, la corrélation entre l'âge à la première passation et la durée du retest est significative. Elle indique que les enfants les plus âgés à la passation initiale ont une légère tendance à avoir un plus long intervalle test-retest (r = .26). Toutefois, la corrélation est faible (r^2= 6.76 %). De plus, lorsque nous réalisons des corrélations partielles qui contrôlent pour l'âge ou pour la durée de l'intervalle, les résultats vont dans le même sens que ceux présentés.

Un quatrième limite est que notre étude n'est pas construite pour tester les biais liés aux caractéristiques de l'expérimentateur (p. ex., apparence, attitude, sexe). Or, dans une étude avec un échantillon de 2'783 enfants évalués avec le WISC-IV pour des décisions de placement éducative, McDermott, Watkins et Rhoad (2014) relèvent des variations non négligeables – surtout dans les scores du QIT et de l'ICV – liées à l'expérimentateur. En effet, *« nearly all WISC–IV scores conveyed significant and nontrivial amounts of variation that had nothing to do with children's actual individual differences and that the Full Scale IQ and Verbal Comprehension Index scores evidenced quite substantial assessor bias »* (McDermott et al., 2014, p. 207). Si les biais liés à l'examinateur ne sont pas à strictement parlé contrôlés, les six expérimentatrices de notre étude ont été entraînées ensemble à administrer le WISC-IV de manière standardisée. Pour chaque protocole, il y a eu une double cotation. Un consensus entre les cotateurs était requis pour chaque différence de cotation dans les réponses verbales des subtests de l'ICV. Les décisions de cotation sont consignées afin d'appliquer les mêmes décisions si la situation se représente dans un autre protocole.

En dernière limite, nous pouvons discuter de la méthode de test-retest. Cette méthode est la plus commune et plus aisée à mettre en place pour évaluer la fidélité temporelle. Cependant, l'estimation de la fidélité test-retest se réfère à la théorie classique des tests (TCT) qui postule la même erreur type de mesure (ETM) quelle que soit l'habileté de l'individu sur la propriété mentale évaluée (homoscédascité). Or, il s'agit d'une simplification qui n'est pas forcément réaliste en toutes circonstances. Par exemple, il peut arriver que les individus ayant les plus faibles habiletés sur la propriété mentale évaluée soient les plus anxieux dans une situation d'évaluation et donc plus sujette à des erreurs d'inattention. L'alternative que proposent les modèles de réponses à l'item (MRI) avec la courbe d'information contourne le postulat d'homoscédascité. En effet, la courbe d'information montre le pouvoir informatif d'un item ou du test entier en fonction de l'habileté sur le trait latent évalué et des paramètres d'item (degré de difficulté, discrimination et pseudo-chance). L'erreur de mesure n'est pas supposée uniforme. En effet, l'erreur de mesure diminue plus le degré de difficulté de l'item est proche de l'habileté du sujet, plus la discrimination de l'item est élevée et plus le paramètre de pseudo-chance est faible. Si l'approche des MRI permettent de déterminer l'incertitude associée la mesure tout au long de l'échelle du trait latent, le recours à cette approche est néanmoins limité par les contraintes pratiques (très important échantillon de sujets et d'items) et techniques (logiciel avancé de statistiques).

CONCLUSION ET PERSPECTIVES

De nombreuses études longitudinales suggèrent que l'intelligence est un trait stable dans le temps (p. ex., Deary, Pattie, & Starr, 2013; Deary, Whalley, Lemmon, Crawford, & Starr, 2000; Hertzog & Schaie, 1986; McCall, 1977). Ces études suivent des cohortes de l'enfance à un âge adulte avancé. La stabilité est évaluée à travers les corrélations entre les performances à au moins deux temps de passations (T1, T2, T3, etc.). La batterie de tests administrés est généralement constituée pour les besoins de l'étude à partir de tests rapides à passer et faciles à coter (p. ex., les Matrices de Raven, Mémoire des chiffres des échelles de Wechsler, test de fluence verbale). Dans le temps et au niveau de leur groupe de référence, les individus ont tendance à maintenir leur rang tout au long de leur vie. Le caractère stable des différences interindividuelles sur l'intelligence confère une valeur prédictive au QI qui l'opérationnalise. Cependant, l'intelligence recouvre diverses aptitudes cognitives selon la définition qu'on adopte. Les différents tests d'intelligence n'évaluent pas exactement la même chose. De plus, de nombreuses sources d'erreur sont capturées dans le score observé à un test. L'ampleur de l'erreur de mesure dépend des propriétés psychométriques du test. Même si le trait psychologique est stable, l'instrument pour l'évaluer peut proposer des scores peu fidèles dans le temps. Des preuves empiriques de validité de l'interprétation et de fidélité des scores doivent donc être apportées à chaque adaptation d'un test. De plus, dans la pratique de l'évaluation psychologique, le psychologue travaille sur les scores observés d'un individu particulier. La stabilité intra-individuelle est particulièrement importante pour poser des hypothèses sur un individu en particulier. Peu explorée dans les études sur la stabilité des scores de tests tels que les Échelles de Wechsler – avant tout utilisées en clinique –, la stabilité intra-individuelle ne va pas forcément de pair avec la stabilité interindividuelle des scores d'un test.

Notre travail a pour objet d'étude les scores de la 4^{e} édition de l'Échelle d'Intelligence pour Enfants et Adolescents de Wechsler (WISC-IV). La batterie est largement utilisée dans l'évaluation cognitive et contribue à définir les pistes d'interventions psychoéducatives. Deux études distinctes sont réalisées à partir des données récoltées. La première étude explore les items des subtests du WISC-IV et évalue la présence d'un éventuel fonctionnement différentiel des items. Il s'agit d'une contribution empirique à la validité de l'interprétation du WISC-IV pour la population des enfants suisses francophones. La seconde étude évalue la stabilité à long terme des scores du WISC-IV dans un échantillon d'enfants tout-venant. Il s'agit d'une contribution théorique et pratique pour une meilleure utilisation du WISC-IV et des prédictions fondées sur ses scores.

S'agissant d'une étude longitudinale, la récolte des données s'est déroulée sur plus de six ans. En première passation, 483 enfants tout-venant âgés de 7 à 12 ans ont été vus dans différentes écoles du canton de Genève. Issus de l'échantillon des 483 enfants, 277 enfants ont été revus pour une seconde passation après un délai test-retest d'une année au minimum. Six psychologues ont administré de manière standardisée les dix subtests obligatoires et le subtest optionnel Complètement d'images lors de deux séances d'environ 45 minutes. Un échantillon aussi large est rare dans une étude longitudinale avec un test en passation individuelle. Pour le constituer, certaines limites doivent être acceptées. L'âge des enfants est restreint aux 7 – 12 ans et il y a une légère surreprésentation des enfants de statut socio-économique élevé. Toutefois, le critère sexe respecte la représentativité de la population des enfants de Genève. Nous n'avons pas pu contrôler le délai test-retest ; il en résulte que les enfants les plus âgés à la première passation tendent légèrement à avoir les plus longs intervalles test-retest. En dépit de certaines limites, les caractéristiques de la distribution des scores de l'échantillon sont proches des valeurs théoriques des notes QI et des notes standards des subtests, indiquant une représentativité au niveau des performances moyennes et de la variabilité intragroupe.

L'étude du fonctionnement différentiel des items (FDI) est réalisée sur le score brut des items de 483 protocoles. Un FDI est présent lorsque des individus possédant la même habileté sur le trait latent réussissent systématiquement mieux ou moins bien un item selon son groupe d'appartenance (par ex. les garçons réussissent systématiquement mieux que les filles). Seuls les subtests Cubes, Similitudes, Vocabulaire, Matrices et Compréhension sont considérés. Des effets directs sont mis en évidence pour les variables âge, sexe et statut socio-économique. L'influence de la variable âge est attendue, étant donné que les enfants les plus âgés réussissent en moyenne plus d'items que les enfants les plus jeunes. Des influences marginales de la variable sexe sont relevées pour Cubes, où les performances des garçons sont en moyenne légèrement meilleures que celles des filles, tandis que pour Matrices, les performances des filles sont en moyenne légèrement meilleures que celles des garçons. Conformément à de nombreuses recherches sur le sujet, l'influence directe de la variable statut socio-économique montre qu'en moyenne, les enfants issus de milieux socio-économiques élevés présentent de meilleures performances que les enfants issus de milieux socio-économiques bas. Par ailleurs, la proportion de la variance totale expliquée par cette variable est la plus importante pour les subtests verbaux. L'influence directe d'une variable ne conduit pas forcément à un biais d'item, il s'agit

d'également tester l'effet indirect de ces variables. Aucun fonctionnement différentiel des items (effet indirect) n'est détecté sur les items pour les variables étudiées. L'absence d'un fonctionnement différentiel de l'item signifie que pour un même item, la probabilité de réussir est la même pour des individus ayant la même habileté sur le trait latent évalué quel que soit leur sous-groupe d'appartenance.

Le WISC-IV a été adapté pour une population d'enfants francophones de France et de Belgique. Toutefois, il est utilisé plus largement dans la communauté francophone et notamment auprès d'enfants suisses romands. Lors de la phase de préexpérimentation, des comparaisons entre des échantillons d'enfants français et d'enfants belges ont conduit à la modification d'un certain nombre d'items. Dès lors il est raisonnable de supposer également certaines différences entre les enfants français et les enfants suisses. Pourtant, en dépit des différences culturelles, éducatives et des spécificités linguistiques entre la France et la Suisse, aucune étude n'est menée pour la détection d'éventuels biais liés aux items. D'ailleurs, pour la dernière édition en date du WISC – le WISC-V –, il n'est pas fait mention d'études comparatives avec des échantillons d'enfants francophones belges ni suisses. Pourtant, l'équité dans l'évaluation psychologique est au cœur des critiques contre les tests. Pour un test avec des contenus linguistique et culturel, il est nécessaire de mener des recherches sur les biais d'items pour la validité de l'interprétation de tests qui sont souvent utilisés sur une population plus élargie que la population qui a servi à constituer les normes.

L'approche des MRI est très intéressante dans l'évaluation du fonctionnement différentiel des items, cependant, celle-ci demande énormément de ressources pour la mettre en œuvre. Elle s'emploie généralement dans les grandes enquêtes internationales (par ex. PISA).

L'étude de la stabilité des scores du WISC-IV avec un échantillon de 277 enfants tout-venant suisses francophones apporte un autre éclairage que celui des études américaines avec des échantillons cliniques. Dans la plupart des recherches menées sur cette quatrième édition du WISC, l'intérêt s'est concentré sur la fidélité temporelle des notes au niveau interindividuel (stabilité différentielle). Les coefficients de stabilité corrigée dans notre échantillon non consultant sont proches des études américaines (Bartoi et al., 2015; Watkins & Smith, 2013) et permettent ainsi une certaine généralisation des résultats. Le QI Total et l'Indice de Compréhension Verbale sont les deux indices avec le coefficient de stabilité le plus élevé. Les scores supplémentaires de notre étude permettent d'ajouter l'Indice d'Aptitude Générale et Gv dont les coefficients de stabilité corrigés sont également autour de .80.

Dans son utilisation clinique, il est essentiel que l'interprétation et les conclusions sur le fonctionnement intellectuel qui se construisent autour des performances obtenues reposent sur des notes fidèles dans le temps sur le plan inter- et intra-individuel. Nos résultats mettent en évidence deux effets – l'effet d'apprentissage et le phénomène de régression à la moyenne – sur lesquels nous allons revenir pour les lier à la pratique de l'évaluation psychologique.

Lorsqu'un projet psychoéducatif est mis en place, il peut arriver qu'une réévaluation soit demandée pour évaluer le changement. Dans notre échantillon d'enfants non consultants, des effets d'apprentissage persistent au-delà d'un intervalle supérieur à un an pour les performances aux indices de Mémoire de Travail et de Vitesse de Traitement. En revanche, les études américaines avec des échantillons cliniques ne montrent généralement pas de différences de moyennes significatives au-delà d'un an. Lors d'une réévaluation dans un délai d'un an avec le WISC-IV, la présence d'une augmentation de performance, notamment sur l'IVT, indique sans doute un effet d'apprentissage et non un changement réel sur la propriété mentale, cependant, la présence même d'un effet d'apprentissage est un indicateur positif.

Dans nos analyses sur le plan intra-individuel, les résultats montrent un phénomène de régression vers la moyenne qui a une implication dans les prédictions formulées à partir des scores du WISC-IV. En effet, approximativement 80 % des enfants dont les performances sont dans la catégorie de la moyenne normative restent dans cette même catégorie lors d'une passation ultérieure et cela pour tous les indices. En revanche, la plupart des enfants dont les performances sont dans les catégories faible (< 85) ou élevée (> 115) changent pour d'une catégorie (vers le haut ou vers le bas) en direction de la moyenne lors d'une passation ultérieure. En tenant compte du phénomène de régression, on peut prédire un maintien dans une catégorie normative moyenne et un changement de catégorie vers la moyenne. Il est intéressant de relever que les enfants ayant des performances initiales faibles (< 85) sont plus touchés par le phénomène de régression vers la moyenne que les enfants ayant des performances initiales élevés (> 115). Cela s'explique sans doute par l'effet d'apprentissage qui tend à augmenter les performances lors de la seconde passation. Des résultats comparables se retrouvent dans l'analyse de la stabilité des forces et faiblesses personnelles.

Le phénomène de régression à la moyenne est pris en compte dans les tables d'intervalle de confiance calculé autour du score proposé dans le manuel du WISC-IV. Compte tenu de la fidélité imparfaite des scores à cause de l'erreur de mesure, on recommande de donner les scores avec leur intervalle de confiance (à un niveau de

90 % ou 95 %). En intégrant la régression vers la moyenne dans leur calcul, ces intervalles sont asymétriques pour les scores qui s'éloignent de la moyenne. Toutefois, ils sont calculés avec des coefficients de fidélité plus élevés que ceux estimés avec la méthode test-retest. Avec la méthode du partage ou du test-retest à court terme, la fidélité des scores du WISC-IV est surestimée par rapport à une méthode qui tient compte de la source d'erreur temporelle. Les intervalles de confiance sont utiles pour instaurer l'idée d'une marge d'erreur à appliquer sur les scores. Cependant, pour des prédictions, nous recommandons de se référer aux distributions des scores de différences entre deux passations avec un délai à long terme. Dans une recherche de Lecerf, Kieng et Geistlich (2016) sont proposées des valeurs seuils pour une différence qui indique un changement significatif de performances sur le plan clinique. Ces valeurs seuils sont calculées en tenant compte de l'effet d'apprentissage et de la régression à la moyenne. Par exemple, pour le QIT, on peut parler de déclin significatif pour des différences égales ou inférieures à 12 points lors d'une passation ultérieure et d'une progression significative pour des différences supérieures ou égales à 17 points lors d'une passation ultérieure. Si on tient compte de la régression à la moyenne et des effets d'apprentissage, on s'attend donc à une certaine augmentation des scores entre deux passations. Pour disposer des données empiriques sur les fréquences d'apparition des différences, le type d'étude longitudinale que nous avons menée est nécessaire.

La publication en décembre 2016 de l'adaptation en français du WISC-V érige au premier plan le besoin de nouvelles recherches pour explorer la stabilité de ses scores. Des modifications dans les subtests et la structure conduisent à une interprétation d'un QI Total basée sur sept subtests et de cinq indices basés chacun sur deux subtests. L'ajout d'un cinquième indice découle du partage de l'Indice de Raisonnement Perceptif en deux indices : un indice visuospatial et un indice de raisonnement fluide. Nos données sur les indices CHC peuvent donner des pistes sur la stabilité des scores. Les scores aux épreuves de compréhension et de connaissances verbales (Gc) ainsi qu'aux épreuves de visuo-spatiales (Gv) présentent les meilleures fidélités temporelles, tandis les scores aux épreuves de raisonnement fluide (Gf), de mémoire de travail (Gwm) et de vitesse de traitement (Gs) sont les moins stables. Toutefois, comme les appariements de subtests sous un facteur commun ne sont plus les mêmes que ceux du WISC-IV, de nouvelles études sur la structure factorielle de l'adaptation en français du WISC-V et sur la stabilité des scores à long terme doivent être réalisées.

RÉFÉRENCES

Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2005). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice*, *19*, 16–26. http://doi.org/10.1111/j.1745-3992.2000.tb00034.x

Abell, N., Springer, D. W., & Kamata, A. (2009). *Developing and validating rapid assessment instruments*. New York, NY: Oxford University Press.

Ackerman, P. L., & Lohman, D. F. (2006). Individual Differences in Cognitive Functions. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 139–161). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Aickin, M., & Gensler, H. (1996). Adjusting for multiple testing when reporting research results: The Bonferroni vs Holm methods. *American Journal of Public Health*, *86*, 726–728. http://doi.org/10.2105/AJPH.86.5.726

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Long Grove, IL: Waveland Press, Inc.

Anastasi, A. (1994). *Introduction à la psychométrie*. Montréal: Guérin universitaire.

Baltes, P. B., Reese, W., & Nesselroade, J. R. (1997). *Life-span developmental psychology: Introduction to research methods*. Monterey, CA: Brooks/Cole.

Bartoi, M., Issner, J. B., Hetterscheidt, L., January, A. M., Kuentzel, J. G., & Barnett, D. (2015). Attention problems and stability of WISC-IV scores among clinically referred children. *Applied Neuropsychology: Child*, *4*, 133–140. http://doi.org/10.1080/21622965.2013.811075

Bauman, E. E. (1991). Stability of WISC-R scores in children with learning difficulties. *Psychology in the Schools*, *28*, 95–100. http://doi.org/10.1002/1520-6807(199104)28:2<95::AID-PITS2310280203>3.0.CO;2-9

Bernaud, J.-L. (2014). *Méthodes de tests et questionnaires en psychologie*. Paris: Dunod.

Bernier, J.-J., & Pietrulewicz, B. (1997). *La psychométrie: traité de mesure appliquée*. Québec: gaëtan morin.

Bertrand, R., & Blais, J.-G. (2004). *Modèles de mesure: L'apport de la théorie des réponses aux items*. Québec: PUQ.

Binet, A. (1910). Nouvelles recherches sur la mesure du niveau intellectuel chez les enfants d'école. *L'année Psychologique*, *17*, 145–201. http://doi.org/10.3406/psy.1910.7275

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's

ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test* (Addison-We). Reading, MA.

Birnbaum, A. (1969). Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology*, *6*, 258–276. http://doi.org/10.1016/0022-2496(69)90005-4

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. New York, NY: Cambridge University Press.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, *111*, 1061–1071. http://doi.org/10.1037/0033-295X.111.4.1061

Bourguignon, O. (2003). *Questions éthiques en psychologie*. Sprimont: Mardaga.

Bremner, D., McTaggart, B., Saklofske, D., & Janzen, T. (2011). WISC-IV GAI and CPI in psychoeducational assessment. *Canadian Journal of School Psychology*, *26*, 209–219. http://doi.org/10.1177/0829573511419090

Brody, N. (1997). Intelligence, schooling, and society. *American Psychologist*, *52*, 1046–1050. http://doi.org/10.1037/0003-066X.52.10.1046

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Publications.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 1904-1920*, *3*, 296–322. http://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist*, *26*, 543–570. http://doi.org/10.1080/13854046.2012.680913

Canivez, G. L., & Watkins, M. W. (1998). Long-term stability of the Wechsler Intelligence Scale for Children–Third Edition. *Psychological Assessment*, *10*, 285–291. http://doi.org/10.1037/1040-3590.10.3.285

Canivez, G. L., & Watkins, M. W. (1999). Long-term stability of the Wechsler Intelligence Scale for Children-among demographic subgroups: Gender, race/ethnicity, and age. *Journal of Psychoeducational Assessment*, *17*, 300–313. http://doi.org/10.1177/073428299901700401

Canivez, G. L., & Watkins, M. W. (2001). Long-term stability of the Wechsler Intelligence

Scale for Children–Third Edition among students with disabilities. *School Psychology Review*, *30*, 438–453.

Capel, R., Monod, D., & Müller, J.-P. (1997). De l'usage perverti des tests inférentiels en sciences humaines. *Genèses*, *26*, 123–142. http://doi.org/10.3406/genes.1997.1436

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge, UK: Cambridge University Press.

Cattell, J. M. (1890). Mental tests and measurements. *Mind*, *15*, 373–381. Retrieved from http://www.jstor.org/stable/2247264

Cattell, R. B. (1941). Some theoretical issues in adult intelligence testing. *Psychological Bulletin*, *38*, 10.

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin*, *40*, 153–193. http://doi.org/10.1037/h0059973

Cattell, R. B., & Tsujioka, B. (1964). The importance of factor-trueness and validity, versus homogeneity and orthogonality, in test scales. *Educational and Psychological Measurement*, *24*, 3–30. http://doi.org/10.1177/001316446402400101

Charter, R. A., & Feldt, L. S. (2001). Meaning of reliability in terms of correct and incorrect clinical decisions: The art of decision making is still alive. *Journal of Clinical and Experimental Neuropsychology*, *23*, 530–537. http://doi.org/10.1076/jcen.23.4.530.1227

Chen, H.-Y., Keith, T. Z., Chen, Y.-H., & Chang, B.-S. (2009). What does the WISC-IV measure? Validation of the scoring and CHC-based interpretative approaches. *Journal of Research in Education Sciences*, *54*, 85–108.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284–290. http://doi.org/http://dx.doi.org/10.1037/1040-3590.6.4.284

Cognet, G. (2006). *NEMI-2 Nouvelle échelle métrique de l'intelligence – 2e edition*. Paris: Editions du Centre de Psychologie Appliquée.

Cognet, G., & Bachelier, D. (2016). *Clinique de l'examen psychologique de l'enfant et de l'adolescent. Approches intégratives et neuropsychologique* (Dunod). Paris.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev.). Mahwah, NJ:

Lawrence Erlbaum Associates, Inc.

Colvin, S. S. (1921). Intelligence and its measurement: A symposium--IV. *Journal of Educational Psychology*, *12*, 136–139. http://doi.org/10.1037/h0065937

Cooper, D. B., Lacritz, L. H., Weiner, M. F., Rosenberg, R. N., & Cullum, C. M. (2004). Category fluency in mild cognitive impairment: reduced effect of practice in test-retest conditions. *Alzheimer Disease & Associated Disorders*, *18*, 120–122.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. http://doi.org/10.1007/bf02310555

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302. http://doi.org/10.1037/h0040957

D'Estampes, L., Garel, B., & Saint Pierre, G. (2003). Test séquentiel: Niveau de confiance après acceptation. *Revue de Statistique Appliquée*, *51*, 75–92. Retrieved from http://www.numdam.org/item?id=RSA_2003__51_3_75_0

Davis, F. (1959). Interpretation of differences among averages and individual test scores. *Journal of Educational Psychology*, *50*, 162–170. http://doi.org/10.1037/h0044024

Deary, I. J., Pattie, A., & Starr, J. M. (2013). The stability of intelligence from age 11 to age 90 years: The Lothian birth cohort of 1921. *Psychological Science*, *24*, 2361–2368. http://doi.org/10.1177/0956797613486487

Deary, I. J., Whalley, L. J., Lemmon, H., Crawford, J. R., & Starr, J. M. (2000). The stability of individual differences in mental ability from childhood to old age: Follow-up of the 1932 Scottish Mental Survey. *Intelligence*, *28*, 49–55. http://doi.org/10.1016/S0160-2896(99)00031-8

Deary, I. J., Whiteman, M. C., Starr, J. M., Whalley, L. J., & Fox, H. C. (2004). The impact of childhood intelligence on later life: Following up the Scottish Mental Surveys of 1932 and 1947. *Journal of Personality and Social Psychology*, *86*, 130–147. http://doi.org/10.1037/0022-3514.86.1.130

Dickes, P., Tournois, J., Flieller, A., & Kop, J.-L. (1994). *La psychométrie : théories et méthodes de la mesure en psychologie*. Paris: PUF.

Dikmen, S. S., Heaton, R. K., Grant, I., & Temkin, N. R. (1999). Test–retest reliability and practice effects of Expanded Halstead–Reitan Neuropsychological Test Battery. *Journal of the International Neuropsychological Society*, *5*, 346–356. http://doi.org/10.1017/S1355617799544056

Duff, K., Beglinger, L. J., Van Der Heiden, S., Moser, D. J., Arndt, S., Schultz, S. K., & Paulsen, J. S. (2008). Short-term practice effects in amnestic mild cognitive impairment: implications for diagnosis and treatment. *International Psychogeriatrics / IPA*, *20*, 986–999. http://doi.org/10.1017/S1041610208007254

Duff, K., Lyketsos, C. G., Beglinger, L. J., Chelune, G., Moser, D. J., Arndt, S., … McCaffrey, R. J. (2011). Practice effects predict cognitive outcome in amnestic Mild Cognitive Impairment. *The American Journal of Geriatric Psychiatry : Official Journal of the American Association for Geriatric Psychiatry*, *19*, 932–939. http://doi.org/10.1097/JGP.0b013e318209dd3a

Duff, K., Westervelt, H. J., McCaffrey, R. J., & Haase, R. F. (2001). Practice effects, test–retest stability, and dual baseline assessments with the California Verbal Learning Test in an HIV sample. *Archives of Clinical Neuropsychology*, *16*, 461–476. http://doi.org/10.1016/S0887-6177(00)00057-3

Elliott, S. N., Piersel, W. C., Witt, J. C., Argulewicz, E. N., Gutkin, T. B., & Galvin, G. A. (1985). Three-year stability of WISC-R IQs for handicapped children from three racial/ethnic groups. *Journal of Psychoeducational Assessment*, *3*, 233–244. http://doi.org/10.1177/073428298500300304

Estevis, E., Basso, M. R., & Combs, D. (2012). Effects of practice on the Wechsler Adult Intelligence Scale-IV across 3- and 6-month intervals. *The Clinical Neuropsychologist*, *26*, 239–254. http://doi.org/10.1080/13854046.2012.659219

Flanagan, D. P., & Dixon, S. G. (2013). The Cattell-Horn-Carroll Theory of Cognitive Abilities. In C. R. Reynolds, K. J. Vannest, & E. Fletcher-Janzen (Eds.), *Encyclopedia of Special Education* (pp. 368–381). Hoboken, NJ, USA: John Wiley & Sons, Inc. http://doi.org/10.1002/9780470373699.speced0381

Flanagan, D. P., & Kaufman, A. S. (2009). *Essentials of WISC®-IV assessment* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2013). *Essentials of cross-battery assessment* (3rd ed., Vol. 84). Hoboken, NJ: John Wiley & Sons, Inc.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin*, *95*, 29–51. http://doi.org/10.1037/0033-2909.95.1.29

Frenette, E., Bertrand, R., Valois, P., Dussault, M., & Hébert, M.-H. (2007). Comparaison empirique des paramètres d'items/sujets de la théorie des réponses aux items et de la théorie classique des tests. *Journal of Educational Measurement and Applied*

Cognitive Sciences, *1*, 1–13. Retrieved from http://www.jemacs.uni.lu/

Gergen, K. J. (1982). From self to science: What is there to know. In J. Suls (Ed.), *Psychological perspectives on the self* (pp. 129–149). London, UK: Lawrence Erlbaum.

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence*, *24*, 13–23.

Gottfredson, L., & Saklofske, D. H. (2009). Intelligence: Foundations and issues in assessment. *Canadian Psychology/Psychologie Canadienne*, *50*, 183–195. http://doi.org/10.1037/a0016641

Grégoire, J. (2007a). L'examen de l'intelligence. In M.-P. Noël (Ed.), *Bilan neurospychologique de l'enfant* (pp. 17–48). Belgique, Wavre: Mardaga.

Grégoire, J. (2007b). Les indices du Wisc-iv et leur interprétation. *Le Journal Des Psychologues*, *253*, 26–30. http://doi.org/10.3917/jdp.253.0026

Grégoire, J. (2009). *L'examen clinique de l'intelligence de l'enfant* (2e ed.). Belgique, Wavre: Mardaga.

Grégoire, J., & Wierzbicki, C. (2007). Analyse de la dispersion des indices du WISC-IV en utilisant l'écart significatif par rapport à la moyenne des quatre indices. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, *57*, 101–106. http://doi.org/10.1016/j.erap.2006.05.002

Gulliksen, H. (2013). *Theory of Mental Tests*. New York, NY: Wiley.

Haladyna, T. M., & Downing, S. M. (2005). Construct-irrelevant Variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*, 17–27. http://doi.org/10.1111/j.1745-3992.2004.tb00149.x

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement*, *10*, 287–302. http://doi.org/10.1177/014662168601000307

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. New York, NY: Springer. Retrieved from https://books.google.ch/books?hl=fr&lr=&id=dUbwCAAAQBAJ&oi=fnd&pg=PA33&dq=hambleton+swaminathan&ots=jsIPGvy0UG&sig=LZAbKXAtnou-A6ia7og3rrKBUk8

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item*

Response Theory. Newbury Park, CA: SAGE Publications.

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice*, *8*, 35–41. http://doi.org/10.1111/j.1745-3992.1989.tb00313.x

Hart, B., & Spearman, C. E. (1912). General Ability, its existence and nature. *British Journal of Psychology, 1904-1920*, *5*, 51–84. http://doi.org/10.1111/j.2044-8295.1912.tb00055.x

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*, 139–164. http://doi.org/10.1177/014662168500900204

Herrnstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life*. New York, NY: Free Press.

Hertzog, C., & Schaie, K. W. (1986). Stability and change in adult intelligence: I. Analysis of longitudinal covariance structures. *Psychology and Aging*, *1*, 159–171. http://doi.org/10.1037/0882-7974.1.2.159

Hoekstra, R. A., Bartels, M., & Boomsma, D. I. (2007). Longitudinal genetic study of verbal and nonverbal IQ from early childhood to young adulthood. *Learning and Individual Differences*, *17*, 97–114. http://doi.org/10.1016/j.lindif.2007.05.005

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*, 1157–1164. http://doi.org/10.3758/s13423-013-0572-3

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement*, *60*, 523–531. http://doi.org/10.1177/00131640021970691

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, N: Erlbaum.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70. http://doi.org/10.2307/4615733

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, *6*, 53–60.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55. http://doi.org/10.1080/10705519909540118

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement*, *6*, 249–260. http://doi.org/10.1177/014662168200600301

Huteau, M. (2006). Alfred Binet et la psychologie de l'intelligence. *Le Journal Des Psychologues*, 24–28. http://doi.org/10.3917/jdp.234.0024

Huteau, M., & Lautrey, J. (2003). *Evaluer l'intelligence*. Paris: PUF.

Huteau, M., & Lautrey, J. (2006). *Les tests d'intelligence*. Paris: La découverte.

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, *104*, 53–69. http://doi.org/http://dx.doi.org/10.1037/0033-2909.104.1.53

Johnson, W., & Bouchard Jr., T. J. (2005a). Constructive replication of the visual–perceptual-image rotation model in Thurstone's (1941) battery of 60 tests of mental ability. *Intelligence*, *33*, 417–430. http://doi.org/http://dx.doi.org/10.1016/j.intell.2004.12.001

Johnson, W., & Bouchard Jr., T. J. (2005b). Testing the grand old models of the structure of human intelligence: It's verbal, perceptual, and visualization (VPZ), not fluid and crystallized. *Intelligence*, *33*.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*, 319–342. http://doi.org/10.1111/j.1745-3984.2001.tb01130.x

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kaufman, A. S., & Kaufman, N. L. (1993). *KABC. Batterie pour l'examen psychologique de l'enfant*. Paris: Editions du Centre de Psychologie Appliquée.

Keith, T. Z., Fine, J. G., Taub, G. E., Reynolds, M. R., & Kranzler, J. H. (2006). Higher order, multisample, confirmatory factor analysis of the Wechsler Intelligence Scale for Children–Fourth Edition: What does it measure? *School Psychology Review*, *35*, 108–127.

Keith, T. Z., & Reynolds, M. R. (2010). Cattell–Horn–Carroll abilities and cognitive tests:

What we've learned from 20 years of research. *Psychology in the Schools*, *47*, 635–650. http://doi.org/10.1002/pits.20496

Kelly, E. L. (1961). Clinical psychology—1960: Report of survey findings. *Newsletter: Division of Clinical Psychology of the American Psychological Association*, *14*, 1–11.

Kelly, T. L. (1927). *Interpretation of educational measurements*. Yonkers-on-Hudson, NY: World Book Co.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*, 151–160. http://doi.org/10.1007/bf02288391

Lander, J. (2010). *Long-term stability of scores on the Wechsler Intelligence Scale for Children- fourth edition in children with learning disabilities*. ProQuest Information & Learning, US. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2010-99220-484&site=ehost-live

Lanz, P. (2000). The Concept of Intelligence in Psychology and Philosophy. In H. Cruse, J. Dean, & H. Ritter (Eds.), *Prerational Intelligence: Adaptive Behavior and Intelligent Systems Without Symbols and Logic, Volume 1, Volume 2 Prerational Intelligence: Interdisciplinary Perspectives on the Behavior of Natural and Artificial Systems, Volume 3* (pp. 19–30). Dordrecht: Springer Netherlands. http://doi.org/10.1007/978-94-010-0870-9_3

Lautrey, J. (2006). L'approche différentielle de l'intelligence. In J. Lautrey (Ed.), *Psychologie du développement et psychologie différentielle* (pp. 357–386). Paris: PUF.

Laveault, D., & Grégoire, J. (2014). *Introduction aux théories des tests en psychologie et en sciences de l'éducation* (3e ed.). Belgique: De boeck.

Lecerf, T., Golay, P., Reverte, I., Senn, D., Favez, N., & Rossier, J. (2012). Scores composites CHC pour le WISC-IV : Normes francophones. *Pratiques Psychologiques*, *18*, 37–50. http://doi.org/10.1016/j.prps.2011.04.001

Lecerf, T., Kieng, S., & Geistlich, S. (2016). WISC-IV : valeurs seuils pour des changements significatifs des scores de différence test–retest. *Pratiques Psychologiques*, XXX–XXX. http://doi.org/10.1016/j.prps.2016.07.003

Lecerf, T., Reverte, I., Coleaux, L., Favez, N., & Rossier, J. (2010). Indice d'aptitude général pour le WISC-IV : Normes francophones. *Pratiques Psychologiques*, *16*, 109–121. http://doi.org/10.1016/j.prps.2009.04.001

Lecerf, T., Reverte, I., Coleaux, L., Maillard, F., Favez, N., & Rossier, J. (2011). Indice d'aptitude général et indice de compétence cognitive pour le WISC-IV : Normes empiriques versus normes statistiques. *European Review of Applied Psychology/Revue Européenne de Psychologie Appliquée*, *61*, 115–122. http://doi.org/10.1016/j.erap.2011.01.001

Lecerf, T., Rossier, J., Favez, N., Reverte, I., & Coleaux, L. (2010). The four- vs. alternative six-factor structure of the French WISC-IV: Comparison using confirmatory factor analyses. *Swiss Journal of Psychology/Schweizerische Zeitschrift Für Psychologie/Revue Suisse de Psychologie*, *69*, 221–232. http://doi.org/10.1024/1421-0185/a000026

Lecoutre, B. (2005). Et si vous étiez un bayésien qui s'ignore? *Revue Modulad*, *32*, 92–105. Retrieved from http://www.univ-rouen.fr/LMRS/Persopage/Lecoutre/telechargements/Lecoutre_B-EtSiVousEtiezUnBayesien.pdf

Lécuyer, R. (2009). Intelligence, où es-tu? In M. Fournier & R. Lécuyer (Eds.), *L'intelligence de l'enfant* (pp. 10–16). France: Sciences humaines.

Leong, F. T. L., Park, Y. S., & Leach, M. M. (2013). Ethics in psychological testing and assessment. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology* (pp. 265–282). Washington, DC, USA: American Psychological Association. http://doi.org/10.1037/14047-015

Levine, A. H., & Marks, L. (1928). *Testing intelligence and achievement*. Oxford, UK: Macmillan.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Magnusson, D. (1967). *Test theory*. Reading, MA: Addison-Wesley .

Martin, O. (1997). *La mesure de l'esprit : origines et développements de la psychométrie, 1900-1950*. Paris: L'Harmattan.

Mayes, S. D., & Calhoun, S. L. (2007). Wechsler Intelligence Scale for Children-Third and -Fourth Edition predictors of academic achievement in children with attention-deficit/hyperactivity disorder. *School Psychology Quarterly*, *22*, 234–249. http://doi.org/10.1037/1045-3830.22.2.234

McCall, R. B. (1977). Childhood IQ's as predictors of adult educational and occupational status. *Science*, *197*, 482–483. http://doi.org/10.1126/science.197.4302.482

McDermott, P. A., Watkins, M. W., & Rhoad, A. M. (2014). Whose IQ is it?—Assessor bias variance in high-stakes psychological assessment. *Psychological Assessment*, *26*, 207–214. http://doi.org/10.1037/a0034832

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, *34*, 100–117. http://doi.org/10.1111/j.2044-8317.1981.tb00621.x

McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf-Gc framework. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 119–151). New York, NY: Guilford Press.

McGrew, K. S. (2005). The Cattell-Horn-Carroll Theory of Cognitive Abilities: Past, Present, and Future. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (pp. 136–181). New York. NY: Guilford Press.

McGrew, K. S., & Wendling, B. J. (2010). Cattell–Horn–Carroll cognitive-achievement relations: What we have learned from the past 20 years of research. *Psychology in the Schools*, *47*, 651–675. http://doi.org/10.1002/pits.20497

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educationnal measurement* (3rd ed., pp. 13–104). New York, NY: Americain Council on Education and Macmillan.

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept* (Vol. 53). Cambridge, UK: Cambridge University Press.

Michell, J. (2000). Normal Science, Pathological Science and Psychometrics. *Theory & Psychology*, *10*, 639–667. http://doi.org/10.1177/0959354300105004

Michell, J. (2004). Item Response Models, pathological science and the shape of error. *Theory & Psychology*, *14*, 121–129. http://doi.org/10.1177/0959354304040201

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*, 297–334. http://doi.org/10.1177/014662169301700401

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2015). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin &*

Review, *23*, 103–123. http://doi.org/10.3758/s13423-015-0947-8

Murphy, K. R., & Davidshofer, C. O. (2001). *Psychological testing: Principles and applications* (5th ed.). Upper Saddle River, NJ: Prentice-Hall.

Muthén, B., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*.

Neisser, U., Boodoo, G., Bouchard Jr, T. J., Boykin, A. W., Brody, N., Ceci, S. J., … Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*, 77–101. http://doi.org/10.1037/0003-066X.51.2.77

Nesselroade, J. R., Pruchno, R., & Jacobs, A. (1986). Reliability vs. stability in the measurement of psychological states: An illustration with anxiety measures. *Psychologische Beitrage*, *28*, 255–264.

Newton, J. H., & McGrew, K. S. (2010). Introduction to the special issue: Current research in Cattell–Horn–Carroll–based assessment. *Psychology in the Schools*, *47*, 621–634. http://doi.org/10.1002/pits.20495

Newton, P. E. (2012). Clarifying the consensus definition of validity. *Journal Measurement: Interdisciplinary Research and Perspectives*, *10*, 1–29. http://doi.org/http://dx.doi.org/10.1080/15366367.2012.669666

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. In *Philosophical Transactions of the Royal Society of London. Series A* (Vol. 236, pp. 333–380). Royal Society. Retrieved from http://www.jstor.org/stable/91337

Nunnally, J. C., & Bernstein, I. H. (2010). *Psychometric theory* (3rd ed.). New York, NY: Tata McGraw-Hill.

Oakman, S., & Wilson, B. (1988). Stability of WISC—R intelligence scores: Implications for 3-year reevaluations of learning disabled students. *Psychology in the Schools*, *25*, 118–120. http://doi.org/10.1002/1520-6807(198804)25:2<118::AID-PITS2310250204>3.0.CO;2-T

Pichot, P. (1999). *Les tests mentaux* (16e ed.). Paris: PUF.

Pintner, R. (1921). Intelligence and its measurement: A symposium--V. *Journal of Educational Psychology*, *12*, 139–143. http://doi.org/10.1037/h0069616

Prifitera, A., Saklofske, D. H., & Weiss, L. G. (1998). The WISC-III in context. In A. Prifitera

& D. H. Saklofske (Eds.), *WISC-III Clinical Use and Interpretation: Scientific-practitioner perspectives*. San Diego, CA: Elsevier Academic Press.

Raiford, S., Weiss, L. G., Rolfhus, E., & Coalson, D. (2005). *General Ability Index*. Hartcourt Assessment, Technical Report.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhague: Danish Institute for Educational Research.

Raven, J. (1998). *Progressives matrices de Raven: CPM / SPM /APM*. Paris: Editions du Centre de Psychologie Appliquée.

Reeve, C. L., & Bonaccio, S. (2011). The nature and the structure of "intelligence." In T. Chamorro-Premuzic, S. von Stumm, & A. Furnham (Eds.), *The Wiley-Blackwell handbook of individual differences* (pp. 187–216). England: John Wiley & Sons.

Rennes, P., Pichot, P., Anstey, & Kourovsky, F. (2000). *D2000 - Tests des dominos*. Paris: Editions du Centre de Psychologie Appliquée.

Reuchlin, M. (1992). Psychométrie. In *Grand dictionnaire de la psychologie*. Paris: Larousse.

Reverte, I. (2015). *L'analyse de la structure factorielle du WISC-IV selon la classification des aptitudes cognitives de Cattell-Horn-Carroll (CHC)*. Université de Genève, Suisse. Retrieved from http://archive-ouverte.unige.ch/unige:46570

Roid, G. H. (2003). *Stanford-Binet Intelligence Scales: Fifth Edition*. Itasca, IL: Riverside Publishing.

Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review*, *9*, 99–103.

Ryan, J. J., Glass, L. A., & Bartels, J. M. (2010). Stability of the WISC-IV in a sample of elementary and middle school children. *Applied Neuropsychology*, *17*, 68–72. http://doi.org/10.1080/09084280903297933

Saklofske, D. H., Prifitera, A., Weiss, L. G., Rolfhus, E., & Zhu, J. (2005). Clinical interpretation of the WISC-IV FSIQ and GAI. In A. Prifitera, D. H. Saklofske, & L. G. Weiss (Eds.), *WISC-IV clinical use and interpretation: Scientist-practitioner perspectives.* (pp. 33–65). San Diego, CA: Elsevier Academic Press. http://doi.org/10.1016/B978-012564931-5/50003-7

Saklofske, D. H., Weiss, L. G., Raiford, S., & Prifitera, A. (2006). Advanced Interpretive Issues with the WISC-IV Full-Scale IQ and General Ability Index Scores. In L. G.

Weiss, D. H. Saklofske, A. Prifitera, & J. Holdnack (Eds.), *WISC-IV advanced clinical interpretation* (pp. 99–138). San Diego, CA: Elsevier Academic Press.

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2012). *Assessment in special and inclusive education* (12th ed.). Belmont, CA: Wadsworth Publishing.

Samejima, F. (1970). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *35*, 139. http://doi.org/10.1007/BF02290599

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 85–100). New York, NY: Springer-Verlag. http://doi.org/10.1007/978-1-4757-2691-6_5

Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. In D. P. Flanagan & P. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 99–144). New York, NY: Guildford Press.

Schneider, W., Niklas, F., & Schmiedeler, S. (2014). Intellectual development from early childhood to early adulthood: The impact of early IQ differences on stability and change over time. *Learning and Individual Differences*, *32*, 156–162. http://doi.org/10.1016/j.lindif.2014.02.001

Schwartzman, A. E., Gold, D., Andres, D., Arbuckle, T. Y., & Chaikelson, J. (1987). Stability of intelligence: A 40-year follow-up. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *41*, 244–256. http://doi.org/10.1037/h0084155

Silverstein, A. (1982). Pattern analysis as simultaneous statistical inference. *Journal of Consulting and Clinical Psychology*, *50*, 234–240. http://doi.org/10.1037/0022-006X.50.2.234

Sireci, S. G., & Sukin, T. (2013). Test validity. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology* (pp. 61–84). Washington, DC, US: American Psychological Association. http://doi.org/10.1037/14047-004

Snyderman, M., & Rothman, S. (1987). Survey of expert opinion on intelligence and aptitude testing. *American Psychologist*, *42*, 137–144. http://doi.org/10.1037/0003-066X.42.2.137

Sparrow, S. S., Pfeiffer, S. I., & Newman, T. M. (2005). Assessment of children who are gifted with the WISC-IV. *WISC-IV: Clinical Use and Interpretation*, 282–299.

Spearman, C. E. (1904a). "General Intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*, 201–292. http://doi.org/10.2307/1412107

Spearman, C. E. (1904b). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*, 72–101. http://doi.org/10.2307/1412159

Spearman, C. E. (1907). Demonstration of formulae for true measurement of correlation. *The American Journal of Psychology*, *18*, 161–169. http://doi.org/10.2307/1412408

Spearman, C. E. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 1904-1920*, *3*, 271–295. http://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Spearman, C. E. (1914). The theory of two factors. *Psychological Review*, *21*, 101–115. http://doi.org/10.1037/h0070799

Spearman, C. E. (1927). *The abilities of man*. Oxford, UK: Macmillan.

Stavrou, E. (1990). The long-term stability of WISC-R scores in mildly retarded and learning-disabled children. *Psychology in the Schools*, *27*, 101–110. http://doi.org/10.1002/1520-6807(199004)27:2<101

Stern, W. (1911). *Die differentielle Psychologie: in ihren methodischen Grundlagen*. Lepizig: Barth.

Sternberg, R. J., Conway, B., Ketron, J., & Bernstein, M. (1981). People's conceptions of intelligence. *Journal of Personality and Social Psychology*, *41*, 37–55. http://doi.org/10.1037/0022-3514.41.1.37

Sternberg, R. J., Grigorenko, E. L., & Bundy, D. A. (2001). The predictive value of IQ. *Merrill-Palmer Quarterly*, *47*, 1–41. http://doi.org/10.1353/mpq.2001.0005

Teresi, J. A., & Jones, R. N. (2013). Bias in psychological assessment and other measures. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology* (pp. 139–164). Washington, DC, US: American Psychological Association. http://doi.org/10.1037/14047-008

Terman, L. M. (1921). Intelligence and its measurement: A symposium--II. *Journal of Educational Psychology*, *12*, 127–133. http://doi.org/10.1037/h0064940

Thorndike, E. L. (1921). Intelligence and its measurement: A symposium--I. *Journal of*

Educational Psychology, *12*, 124–127. http://doi.org/10.1037/h0064596

Thorndike, R. M., & Hagen, E. P. (1977). *Measurement and evaluation in psychology and education* (4th ed.). New York, NY: John Wiley & Sons, Inc.

Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review*, *38*, 406–427. http://doi.org/10.1037/h0069792

Truscott, S. D., Narrett, C. M., & Smith, S. E. (1994). WISC—R subtest reliability over time: Implications for practice and research. *Psychological Reports*, *74*, 147–156. http://doi.org/10.2466/pr0.1994.74.1.147

Turon-Lagot, E. (2012). *WISC-IV: une mesure des manifestations de l'intelligence chez l'enfant*. France: Eric Turon-Lagot.

van de Vijver, F. J. ., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. USA: SAGE Publications.

van der Linden, W. J. (1986). The changing conception of measurement in education and psychology. *Applied Psychological Measurement*, *10*, 325–332. http://doi.org/10.1177/014662168601000401

van der Maas, H. L. J., Dolan, C. V, Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*. Van Der Maas, Han L. J.: Department of Psychology, University of Amsterdam, Amsterdam, Netherlands, h.l.j.vandermaas@uva.nl: American Psychological Association. http://doi.org/10.1037/0033-295X.113.4.842

van der Maas, H. L. J., Kan, K.-J., & Borsboom, D. (2014). Intelligence is what the intelligence test measures. Seriously. *Journal of Intelligence*, *2*, 12–15. http://doi.org/10.3390/jintelligence2010012

Vance, H. B., Blixt, S., Ellis, R., & Debell, S. (1981). Stability of the WISC-R for a sample of exceptional children. *Journal of Clinical Psychology*, *37*, 397–399. http://doi.org/10.1002/1097-4679(198104)37:2<397

Vernon, P. E. (1950). *The Structure of Human Abilities*. London, UK: Methuen.

Voelkle, M. C., Brose, A., Schmiedek, F., & Lindenberger, U. (2014). Toward a unified framework for the study of between-person and within-Person structures: Building a bridge between two research paradigms. *Multivariate Behavioral Research*, *49*, 193–213. http://doi.org/10.1080/00273171.2014.889593

Voyazopoulos, R., Vannetzel, L., & Eynard, L. A. (2011). *L'examen psychologique de l'enfant et l'utilisation des mesures* (Dunod). Paris.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, *117*, 250–270. http://doi.org/10.1037/0033-2909.117.2.250

Vrignaud, P. (2002). Les biais de mesure, savoir les identifier pour y remédier. *Bulletin de Psychologie*, *55*, 625–634.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*, 426–482. http://doi.org/10.2307/1990256

Wasserman, J. D., & Bracken, B. A. (2013). Fundamental psychometric considerations in assessment. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of Psychology* (pp. 50–81). Hoboken, NJ, USA: John Wiley & Sons, Inc.

Watkins, M. W., Greenawalt, C. G., & Marcell, C. M. (2002). Factor Structure of the Wechsler Intelligence Scale for Children–Third Edition among Gifted Students. *Educational and Psychological Measurement*, *62*, 164–172. http://doi.org/10.1177/0013164402062001011

Watkins, M. W., Lei, P., & Canivez, G. L. (2007). Psychometric intelligence and achievement: A cross-lagged panel analysis. *Intelligence*, *35*, 59–68. http://doi.org/10.1016/j.intell.2006.04.005

Watkins, M. W., & Smith, L. G. (2013). Long-term stability of the Wechsler Intelligence Scale for Children–Fourth Edition. *Psychological Assessment*, *25*, 477–483. http://doi.org/10.1037/a0031653

Wechsler, D. (1949). *Manual for the Wechsler Intelligence Scale for Children*. New York, NY: Psychological Corporation.

Wechsler, D. (1955). *Manual for the Wechsler adult intelligence scale* . Oxford, England: Psychological Corp.

Wechsler, D. (1974). *Manual for the Wechsler Intelligence Scale for Children–Revised*. New York, NY: Psychological Corporation.

Wechsler, D. (1975). Intelligence defined and undefined: A relativistic appraisal. *American Psychologist*, *30*, 135–139. http://doi.org/10.1037/h0076868

Wechsler, D. (1991). *Manual for the Wechsler Intelligence Scale for Children–Third*

edition. San Antonio, TX: Psychological Corporation.

Wechsler, D. (2003). *Manual for the Wechsler Intelligence scale for children–Fourth edition*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (2005a). *WISC-IV – Manuel d'administration et de cotation*. Paris: Editions du Centre de Psychologie Appliquée.

Wechsler, D. (2005b). *WISC-IV – Manuel d'interprétation*. Paris: Editions du Centre de Psychologie Appliquée.

Williams, P. E., Weiss, L. G., & Rolfhus, E. L. (2003). *WISC-IV technical report# 2: Psychometric properties*. *The Psychological Corporation's WISC-IV Technical Manual* (Vol. 2). Retrieved from http://www.pearsonassess.ca/content/dam/ani/clinicalassessments/ca/programs/pdfs/wisc-iv-technical-report-number-2-psychometric-properties.pdf

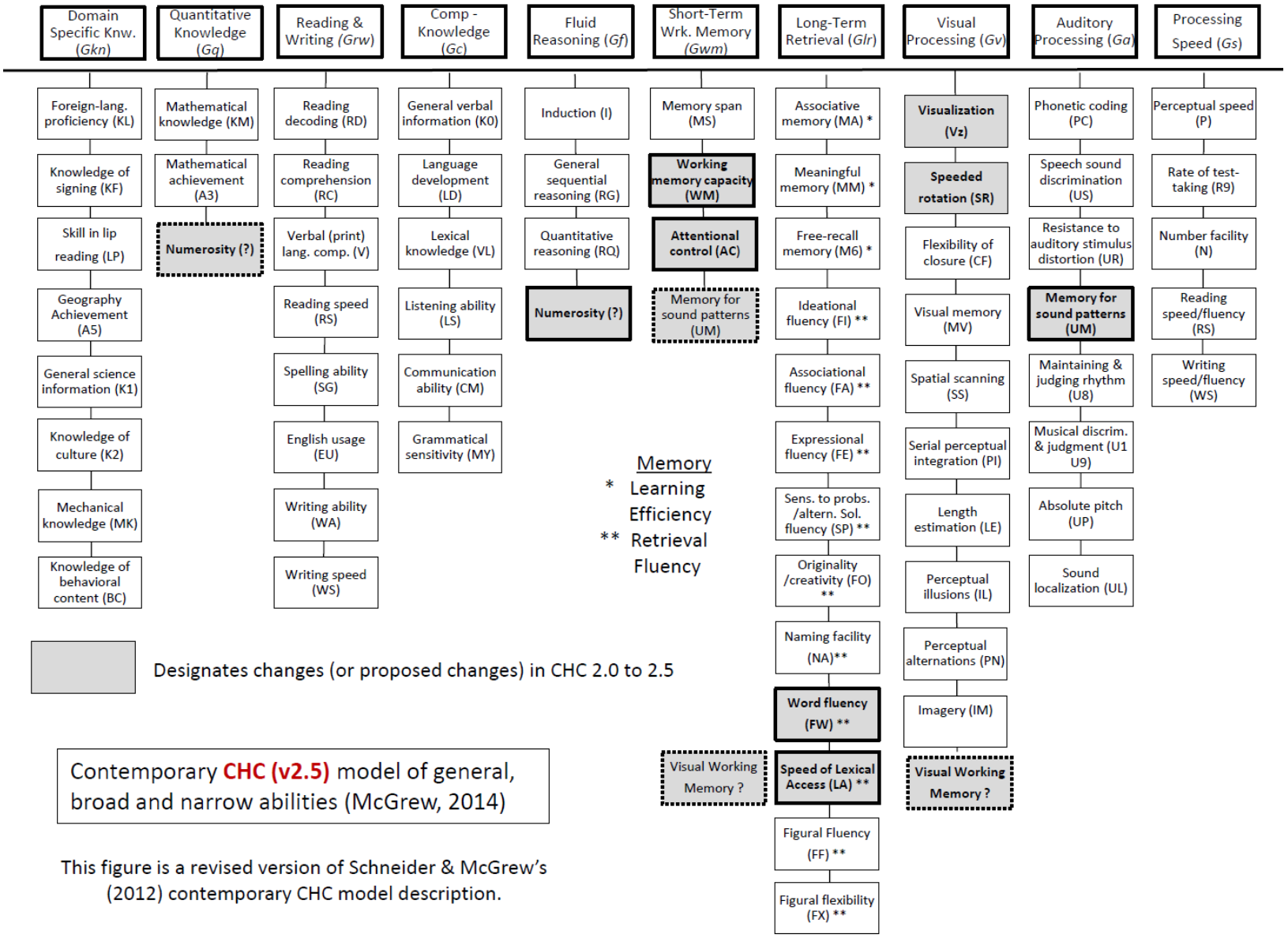
Woods, C. M. (2009). Evaluation of MIMIC-Model Methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, *44*, 1–27. http://doi.org/10.1080/00273170802620121

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. University of California, Los Angeles, USA.

ANNEXES

Annexe A

Représentation du modèle CHC selon Mcgrew (2014, version 2.5)


Domain Specific Knw. (Gkn)
Quantitative Knowledge (Gq)
Reading & Writing (Grw)
Comp - Knowledge (Gc)
Fluid Reasoning (Gf)
Short-Term Wrk. Memory (Gwm)
Long-Term Retrieval (Glr)
Visual Processing (Gv)
Auditory Processing (Ga)
Processing Speed (Gs)
Foreign-lang. proficiency (KL)
Knowledge of signing (KF)
Skill in lip reading (LP)
Geography Achievement (A5)
General science information (K1)
Knowledge of culture (K2)
Mechanical knowledge (MK)
Knowledge of behavioral content (BC)
Mathematical knowledge (KM)
Mathematical achievement (A3)
Numerosity (?)
Reading decoding (RD)
Reading comprehension (RC)
Verbal (print) lang. comp. (V)
Reading speed (RS)
Spelling ability (SG)
English usage (EU)
Writing ability (WA)
Writing speed (WS)
General verbal information (K0)
Language development (LD)
Lexical knowledge (VL)
Listening ability (LS)
Communication ability (CM)
Grammatical sensitivity (MY)
Induction (I)
General sequential reasoning (RG)
Quantitative reasoning (RQ)
Numerosity (?)
Memory span (MS)
Working memory capacity (WM)
Attentional control (AC)
Memory for sound patterns (UM)
Associative memory (MA) *
Meaningful memory (MM) *
Free-recall memory (M6) *
Ideational fluency (FI) **
Associational fluency (FA) **
Expressional fluency (FE) **
Sens. to probs. /altern. Sol. fluency (SP) **
Originality /creativity (FO) **
Naming facility (NA)**
Word fluency (FW) **
Speed of Lexical Access (LA) **
Figural Fluency (FF) **
Figural flexibility (FX) **
Visualization (Vz)
Speeded rotation (SR)
Flexibility of closure (CF)
Visual memory (MV)
Spatial scanning (SS)
Serial perceptual integration (PI)
Length estimation (LE)
Perceptual illusions (IL)
Perceptual alternations (PN)
Imagery (IM)
Visual Working Memory ?
Phonetic coding (PC)
Speech sound discrimination (US)
Resistance to auditory stimulus distortion (UR)
Memory for sound patterns (UM)
Maintaining & judging rhythm (U8)
Musical discrim. & judgment (U1 U9)
Absolute pitch (UP)
Sound localization (UL)
Perceptual speed (P)
Rate of test-taking (R9)
Number facility (N)
Reading speed/fluency (RS)
Writing speed/fluency (WS)
Memory
* Learning Efficiency
** Retrieval Fluency
Designates changes (or proposed changes) in CHC 2.0 to 2.5
Contemporary CHC (v2.5) model of general, broad and narrow abilities (McGrew, 2014)
This figure is a revised version of Schneider & McGrew's (2012) contemporary CHC model description.

Annexe B

Questionnaire sociodémographique à remplir par les parents d'enfants participant à la recherche

QUESTIONNAIRE À REMPLIR PAR LE(S) PARENT(S)

Nous vous rappelons que les informations fournies par ce questionnaire seront traitées de manière strictement confidentielle.

Nom, Prénom de votre

enfant : .....

Nom de l'école fréquentée par votre

enfant : .....

A. Quel est le plus haut niveau de formation que vous avez accompli ?

Pour la mère :

☐ Primaire

☐ Secondaire inférieur (cycle d'orientation)

☐ Formation professionnelle (apprentissage)

☐ Secondaire supérieur (collège, école de commerce, gymnase, etc.)

☐ École technique ou professionnelle supérieure (ex. : infirmières, assistants sociaux, maîtres d'école, ingénieur, etc.)

☐ Université. École polytechnique fédérale

☐ Autre (*veuillez préciser*) : .....

Pour le père :

☐ Primaire

☐ Secondaire inférieur (cycle d'orientation)

☐ Formation professionnelle (apprentissage)

☐ Secondaire supérieur (collège, école de commerce, gymnase, etc.)

☐ École technique ou professionnelle supérieure (ex. : infirmières, assistants sociaux, maîtres d'école, ingénieur, etc.)

☐ Université. École polytechnique fédérale

☐ Autre (*veuillez préciser*) : .....

B. Quelle est votre profession actuelle ?

Pour la mère : .....

Pour le père : .....

C. Quelle(s) langue(s) parlez-vous habituellement avec votre enfant ?

Si plusieurs langues, veuillez <u>numéroter</u> les langues de la plus parlée à la moins parlée.

☐ Anglais

☐ Albanais

☐ Allemand / Suisse-allemand

☐ Espagnol

☐ Français

☐ Italien

☐ Portugais

☐ Serbe, Croate ou Bosniaque

☐ Turc

☐ Autre (*veuillez préciser*) : .....

D. À quelle fréquence consultez-vous les médias suivants ?

Pour l'enfant :

Jamais	Rarement, moins d'une fois	Une ou quelques fois par semaine	Tous les jours, moins de 3h	Tous les jours, plus de 3h

Jamais	Rarement, moins d'une fois par semaine	Une ou quelques fois par semaine	Tous les jours, moins de 3h	Tous les jours, plus de 3h	
☐	☐	☐	☐	☐	Radio
☐	☐	☐	☐	☐	Télévision
☐	☐	☐	☐	☐	Journal
☐	☐	☐	☐	☐	Livre, revue
☐	☐	☐	☐	☐	Internet

Pour la mère :

Jamais	Rarement, moins d'une fois par semaine	Une ou quelques fois par semaine	Tous les jours, moins de 3h	Tous les jours, plus de 3h	
☐	☐	☐	☐	☐	Radio
☐	☐	☐	☐	☐	Télévision
☐	☐	☐	☐	☐	Journal
☐	☐	☐	☐	☐	Livre, revue
☐	☐	☐	☐	☐	Internet

Pour le père :

Jamais	Rarement, moins d'une fois par semaine	Une ou quelques fois par semaine	Tous les jours, moins de 3h	Tous les jours, plus de 3h	
☐	☐	☐	☐	☐	Radio
☐	☐	☐	☐	☐	Télévision
☐	☐	☐	☐	☐	Journal
☐	☐	☐	☐	☐	Livre, revue
☐	☐	☐	☐	☐	Internet

E. Dans quelle(s) langue(s) consultez-vous ces médias ?

A :	Anglais	**F** :	Italien
B :	Albanais	**G** :	Portugais
C :	Allemand / Suisse-allemand	**H** :	Serbe, Croate ou Bosniaque
D :	Espagnol	**I** :	Turc
E :	Français	**J** :	Autre (*veuillez préciser*)

Veuillez reporter ci-dessous la (les) lettre(s) correspondant à la (aux) langue(s).

Pour l'enfant :

Radio : .....

Livre, revue :.....

Télévision : .....

Internet : .....

Journal : .....

Pour la mère :

Radio : .....

Livre, revue :.....

Télévision : .....

Internet : .....

Journal : .....

Pour le père :

Radio : .....

Livre, revue :.....

Télévision : .....

Internet : .....

Journal : .....

Merci d'avoir pris le temps de remplir ce questionnaire ainsi que de bien vouloir nous le retourner ou de le remettre à l'enseignant(e) au moyen de l'enveloppe ci-jointe.

Annexe C

Tableau des dix catégories selon les types de profession

Catégorie	Type de profession
1	Directeurs, cadres de direction et gérants
2	Professions intellectuelles et scientifiques
3	Professions intermédiaires
4	Employés de type administratif
5	Personnel des services directs aux particuliers, commerçants et vendeurs
6	Agriculteurs et ouvriers qualifiés de l'agriculture, de la sylviculture et de la pêche
7	Métiers qualifiés de l'industrie et de l'artisanat
8	Conducteurs d'installations et de machines, et ouvriers de l'assemblage
9	Professions élémentaires
10	Sans emploi, au foyer

Annexe D

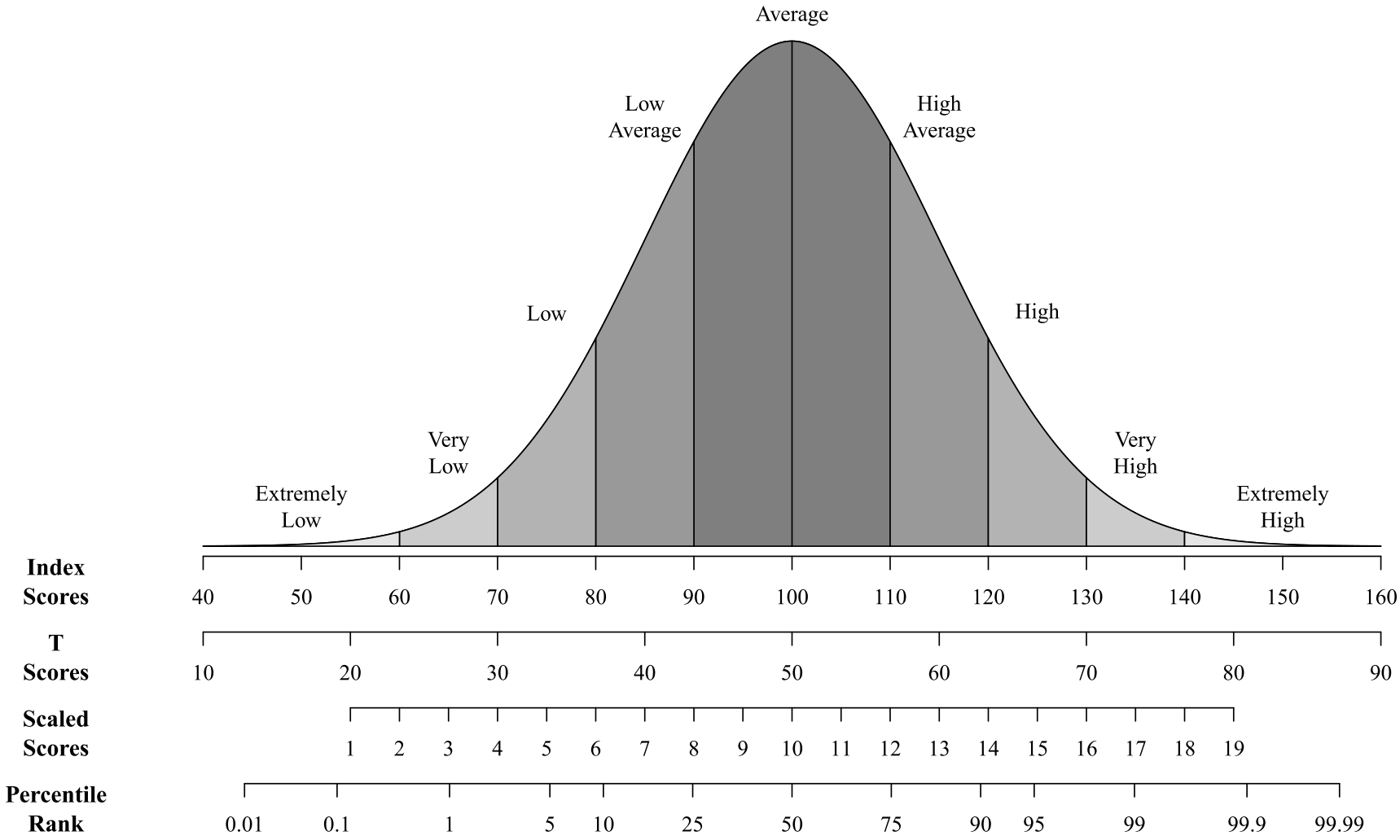
Tableau des fréquences, pourcentages et pourcentages cumulés pour les intervalles Test-Retest (N=277)

Intervalle Test-Retest [mois]	Fréquence	%	% cumulé
39	1	0.4	0.4
35	5	1.8	2.2
34	2	0.7	2.9
33	3	1.1	4.0
32	1	0.4	4.3
31	10	3.6	7.9
30	16	5.8	13.7
29	12	4.3	18.1
28	16	5.8	23.8
27	13	4.7	28.5
26	7	2.5	31.0
25	9	3.2	34.3
24	6	2.2	36.5
22	1	0.4	36.8
21	15	5.4	42.2
20	24	8.7	50.9
19	17	6.1	57.0
18	16	5.8	62.8
17	5	1.8	64.6
16	8	2.9	67.5
15	11	4.0	71.5
14	31	11.2	82.7
13	36	13.0	95.7
12	12	4.3	100

Annexe E

Courbe normale et distributions de scores standardisés pour les indices et les subtests du WISC-IV

Standard Scores



Annexe F

Tableau des coefficients de fidélité et des erreurs types de mesure utilisés pour les analyses de la stabilité intra-individuelle

	$r_{xx'}$	1 ETM	2 ETM	3 ETM
Subtest				
Cubes	.81	1.32	2.64	3.96
Similitudes	.80	1.38	2.76	4.14
Mémoire des chiffres	.86	1.16	2.32	3.48
Identification de concepts	.73	1.56	3.12	4.68
Code	.71	1.64	3.28	4.92
Vocabulaire	.78	1.40	2.80	4.20
Séquence Lettres-Chiffres	.82	1.33	2.66	3.99
Matrices	.86	1.16	2.32	3.48
Compréhension	.72	1.61	3.22	4.83
Symboles	.83	1.28	2.56	3.84
Complètement d'images [a]	.79	1.39	2.78	4.17
Composite				
ICV	.89	4.98	9.96	14.94
IRP	.88	5.24	10.48	15.72
IMT	.88	5.17	10.34	15.51
IVT	.84	6.01	12.02	18.03
QIT	.94	3.63	7.26	10.89
IAG	.92	4.24	8.48	12.72
ICC	.89	4.97	9.94	14.91
Gc	.84	5.26	10.52	15.78
Gf	.85	5.80	11.60	17.40
Gv [a]	.86	5.71	11.42	17.13
Gwm	.89	4.95	9.90	14.85
Gs	.85	5.85	11.70	17.55

Note. Les coefficients de fidélité et les erreurs types de mesure des 11 subtests, des 4 indices standard et du QIT proviennent du Manuel d'interprétation du WISC-IV (Wechsler, 2005b).

Les coefficients de fidélité et les erreurs types de mesures des indices IAG, ICC et des 5 facteurs CHC proviennent des travaux de Lecerf et collaborateurs (Lecerf et al., 2011, 2012)

ETM = erreur type de mesure ; ICV = Indice de Compréhension Verbale ; IRP = Indice de Raisonnement Perceptif ; IMT = Indice de Mémoire de Travail ; IVT = Indice de Vitesse de Traitement ; QIT = QI Total ; IAG = Indice d'Aptitude Générale ; ICC = Indice de Compétence Cognitive ; Gc = intelligence cristallisée ; Gf = intelligence fluide ; Gv = Traitement visuel ; Gwm = Mémoire à court terme; Gs = Vitesse de traitement.

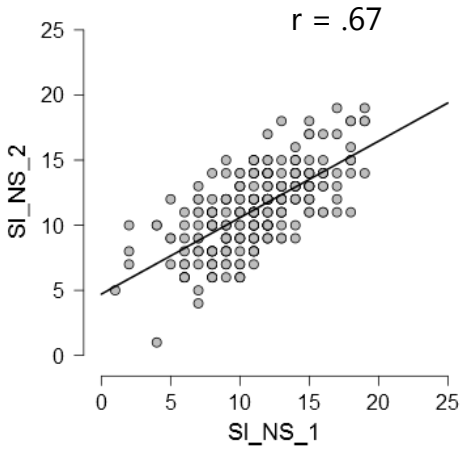
[a] N = 248.

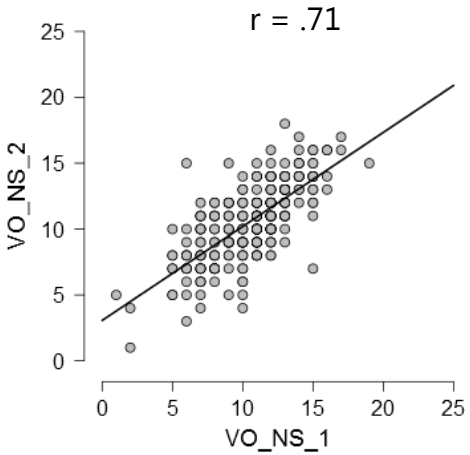
Annexe G

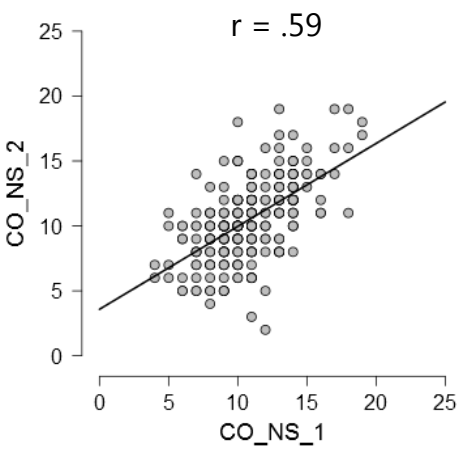
Graphiques des nuages de points pour les subtests du WISC-IV

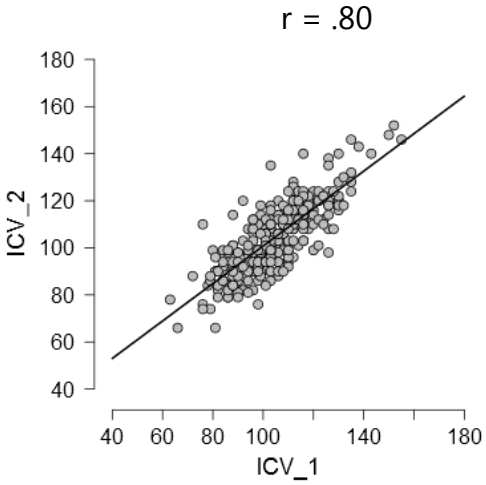
Graphiques des nuages de points pour les indices standards du WISC-IV

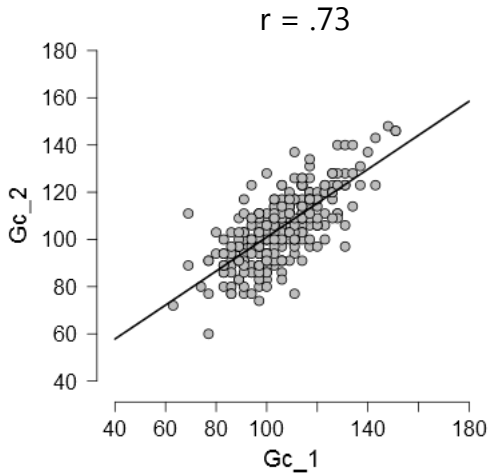
Graphiques des nuages de points pour les indices CHC du WISC-IV


r = .67
SI_NS_2
SI_NS_1
0
5
10
15
20
25

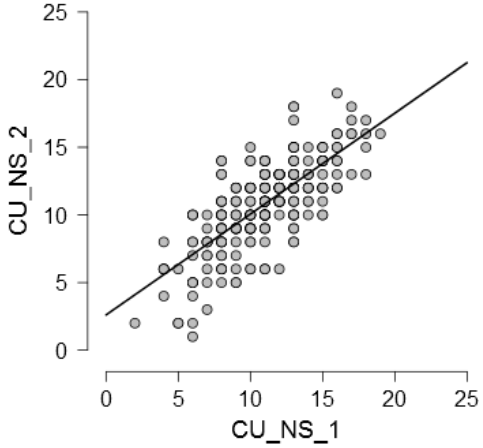

r = .71
VO_NS_2
VO_NS_1
0
5
10
15
20
25


r = .59
CO_NS_2
CO_NS_1
0
5
10
15
20
25

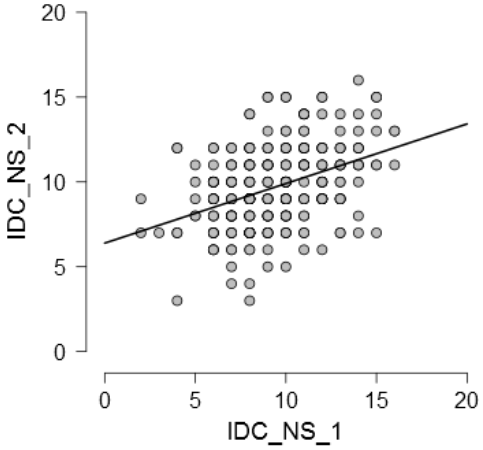

r = .80
ICV_2
ICV_1
40
60
80
100
120
140
160
180


r = .73
Gc_2
Gc_1
40
60
80
100
120
140
160
180

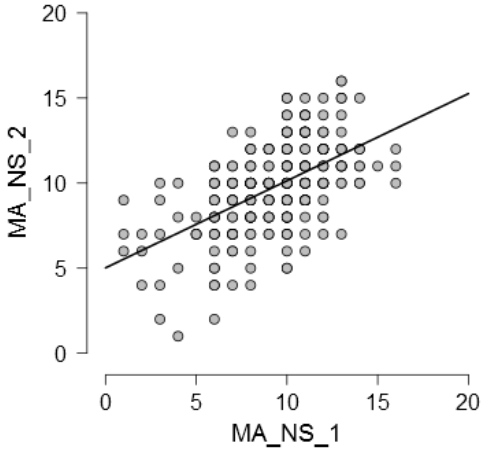
r = .74



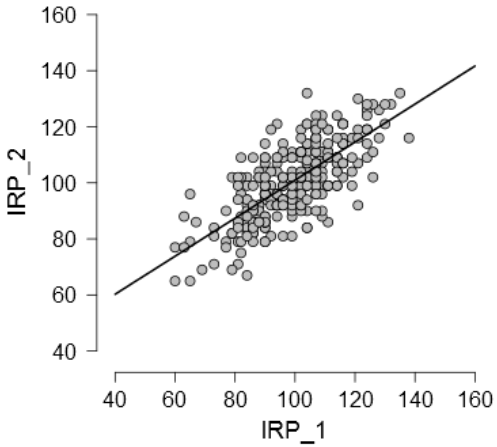
r = .40



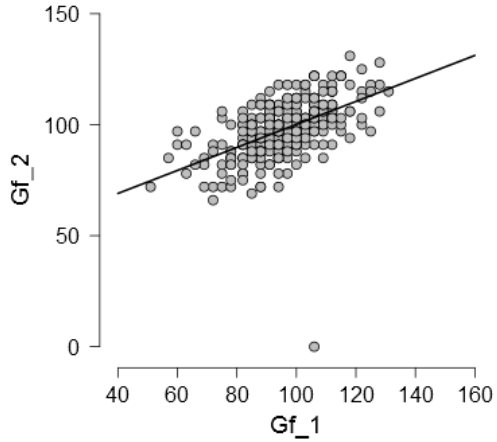
r = .56



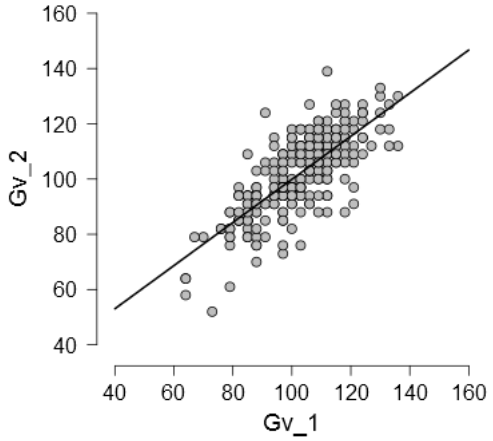
r = .73



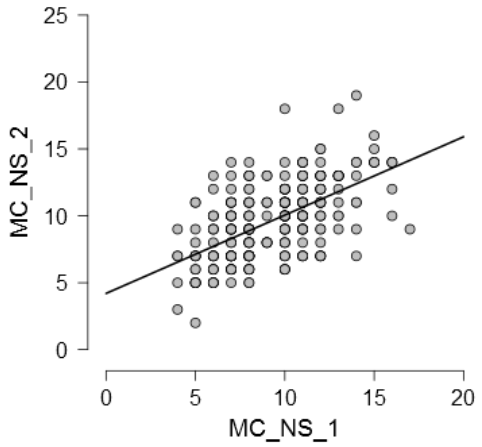
r = .53



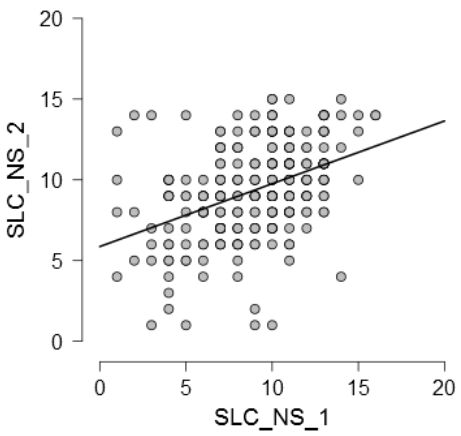
r = .75



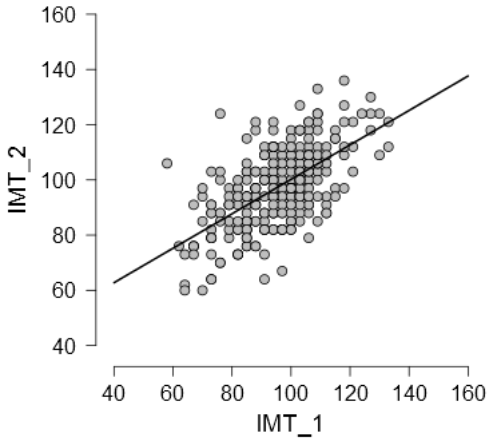
r = .58



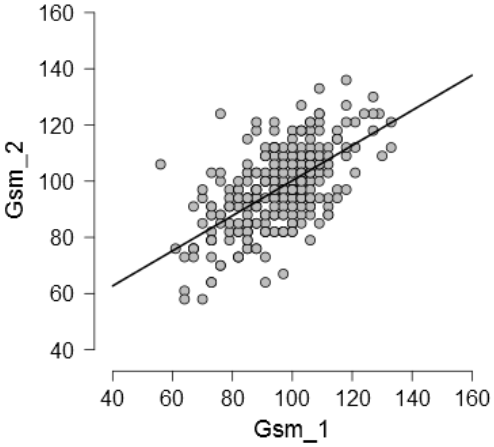
r = .42



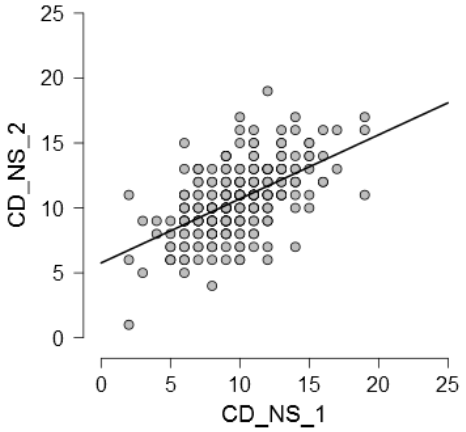
r = .61



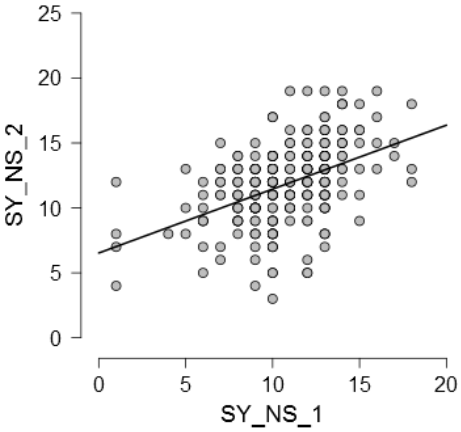
r = .61



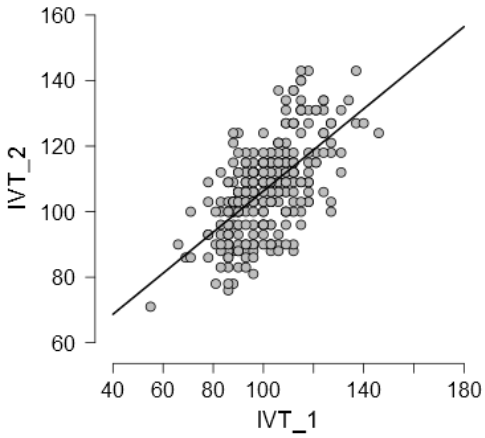
r = .56



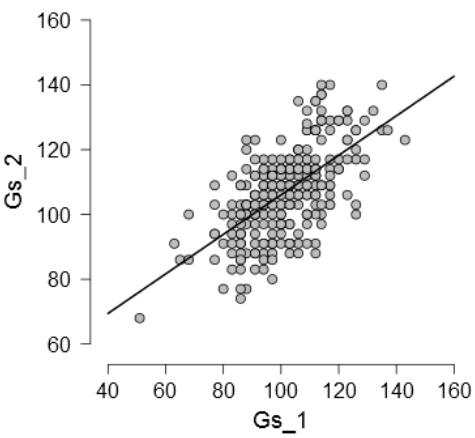
r = .47



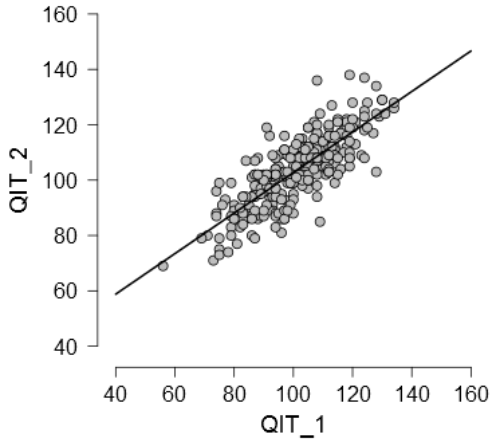
r = .62



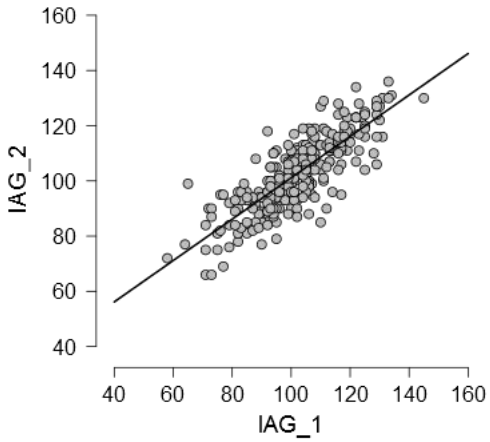
r = .62



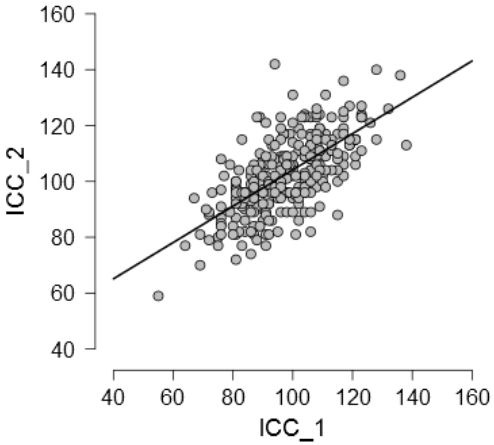
r = .80



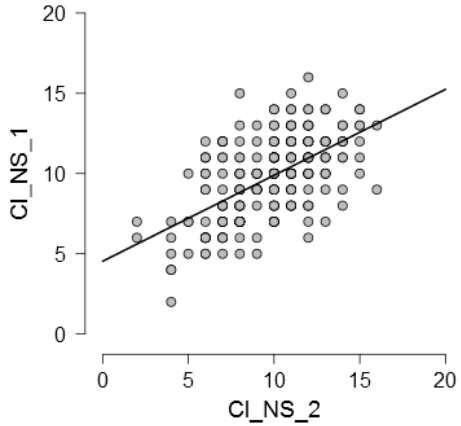
r = .82



r = .66



r = .57



Annexe H

Tableau des distributions de fréquences cumulées (en pourcentage) des scores de différence entre les deux passations du WISC-IV pour les subtests

Tableau des distributions de fréquences cumulées (en pourcentage) des scores de différence entre les deux passations du WISC-IV pour les indices standards

Tableau des distributions de fréquences cumulées (en pourcentage) des scores de différence entre les deux passations du WISC-IV pour les indices CHC

$\lvert T2 - T1\rvert$	CUB	SIM	MDC	IDC	COD	VOC	SLC	MAT	COM	SYM	CIM
0	20.9	19.5	19.9	14.1	16.6	22.7	16.6	16.6	15.9	11.6	19.4
1	55.6	44.0	53.1	41.9	47.3	55.6	38.3	41.2	50.5	38.6	54.8
2	75.1	67.1	70.8	61.0	69.7	76.9	62.1	69.0	71.5	58.8	72.6
3	87.0	83.8	83.4	80.1	81.2	90.3	77.6	83.4	83.8	75.1	85.1
4	94.2	92.1	91.7	90.3	89.9	97.1	86.3	92.4	91.0	85.6	91.9
5	98.9	96.4	96.0	94.2	95.3	98.6	93.1	97.1	96.4	91.7	97.6
6	100	98.9	98.2	98.2	97.5	99.3	96.4	99.3	97.8	96.0	99.2
7		99.6	99.3	98.9	98.9	-	97.1	99.6	98.9	98.6	100
8		100	100	100	99.3	99.6	97.5	100	99.6	99.6	
9					100	100	98.6		-	-	
10							98.9		100	-	
11							99.3			100	
12							100				

Note. Dans les colonnes sont présentés les pourcentages cumulés des scores de différence entre les deux passations. $\lvert T2 - T1\rvert$ = différence absolue entre Retest – Test; CUB = Cubes; SIM = Similitudes; MDC = Mémoire des chiffres; IDC = Identification de concepts; COD = Code; VOC = Vocabulaire; SLC = Séquence Lettres-chiffres ; MAT = Matrices ; COM = Compréhension ; SYM = Symboles ; CIM = Complètement d'images.

$\|T2 - T1\|$	QIT	ICV	IRP	IMT	IVT	IAG	ICC
0	6.5	6.9	8.3	10.5	10.1	7.2	3.6
1	17.3	8.3	-	-	-	18.4	6.1
2	24.5	25.3	23.1	10.8	11.2	27.1	17.3
3	32.5	27.8	30.7	29.2	20.2	36.8	20.6
4	37.9	39.0	37.9	29.6	23.1	44.4	26.7
5	45.8	42.6	45.1	-	23.8	50.9	31.8
6	53.8	49.1	51.6	44.4	32.5	60.6	40.8
7	58.8	56.7	54.9	-	37.2	66.8	46.6
8	67.5	64.6	57.4	-	38.6	73.6	53.1
9	72.9	69.0	62.8	63.5	46.2	76.2	57.0
10	77.3	72.9	69.7	63.9	54.9	79.4	61.4
11	82.3	75.8	72.2	-	-	82.3	66.8
12	84.6	81.2	76.9	71.5	59.6	85.6	70.4
13	86.6	84.8	79.1	-	65.3	88.1	74.4
14	91.7	88.8	83.0	71.8	66.1	89.9	76.5
15	94.2	91.0	97.0	81.9	75.1	92.1	79.8
16	95.3	93.5	87.4	-	81.6	93.9	82.7
17	-	94.2	90.3	-	82.7	94.6	84.8
18	-	95.7	91.7	89.9	84.5	96.8	87.0
19	96.0	96.4	93.1	-	87.4	97.1	89.9
20	97.1	96.8	94.6	-	87.7	97.5	92.4
21	-	97.5	95.3	92.4	88.1	97.8	93.5
22	-	97.8	95.7	-	91.0	98.6	93.9
23	97.8	-	96.4	-	91.7	98.9	94.9
24	98.9	98.2	97.1	95.7	93.1	-	95.3
25	99.3	-	97.8	-	96.0	99.3	96.0
26	-	98.6	-	-	-	99.6	-
27	-	-	98.9	97.5	96.8	-	97.5
28	100	99.3	99.3	-	97.5	-	-
29		-	96.6	-	98.2	-	-
30		-	-	98.9	-	-	97.8
31		-	100	-	98.9	-	98.2
32		-		-	-	-	98.9
33		-		99.3	99.3	-	-
34		100		-	99.6	100	99.3
35				-	-		99.6
36				-	100		-
48				100			100

Note. Dans les colonnes sont présentés les pourcentages cumulés des scores de différence entre les deux passations. $|T2 - T1|$ = différence absolue entre Retest – Test; QIT = QI Total; ICV = Indice de Compréhension Verbale; IRP = Indice de Raisonnement Perceptif; IMT = Indice de Mémoire de Travail; IVT = Indice de Vitesse de Traitement; IAG = Indice d'Aptitude Générale; ICC = Indice de Compétence Cognitive.

$\|T2 - T1\|$	Gc	Gf	Gv	Gwm	Gs
0	10.8	8.3	14.1	10.5	10.1
2	17.0	-	-	-	11.6
3	30.7	27.8	38.3	29.2	22.0
4	-	29.2	-	-	-
5	37.2	-	-	29.6	26.4
6	49.8	47.3	56.5	44.4	35.4
7	-	50.9	-	-	-
8	57.4	-	-	-	40.1
9	65.0	62.1	74.2	63.5	52.7
10	-	64.6	-	-	-
11	72.6	-	-	-	57.8
12	76.2	74.0	84.3	71.5	65.0
13	-	74.7	-	-	-
14	83.8	-	-	-	71.8
15	85.9	80.9	90.3	81.9	79.8
16	-	85.2	-	-	-
17	91.3	-	-	-	84.5
18	91.7	89.9	93.5	89.9	86.6
19	92.1	91.0	-	-	87.0
20	94.9	-	-	-	89.2
21	-	93.1	96.0	92.4	91.0
22	-	94.6	-	-	-
23	97.1	-	-	-	94.9
24	-	97.1	98.0	95.7	95.3
25	97.5	-	-	-	-
26	98.2	-	-	-	97.1
27	-	-	98.8	97.5	-
28	98.6	98.6	-	-	97.5
29	98.9	-	-	-	98.2
30	-	-	99.6	98.9	-
31	-	99.6	-	-	-
32	-	-	-	-	99.6
33	-	-	100	99.3	-
34	99.6	-		-	-
35	-	-		-	100
36	-	-		-	
37	-	100		-	
42	100			-	
48				99.6	
50				100	

Note. Dans les colonnes sont présentés les pourcentages cumulés des scores de différence entre les deux passations. $|T2 - T1|$ = différence absolue entre Retest – Test; Gc = intelligence cristallisée ; Gf = intelligence fluide ; Gv = Traitement visuel ; Gwm = Mémoire à court terme; Gs = Vitesse de traitement.

LISTE DES ABRÉVIATIONS ET DES SIGLES

Abréviation & Sigle	Signification
2-PLM	*Two Parameters Logistic Model* : Modèle logistique à deux paramètres (modèle de Birnbaum)
AIC	*Akaike Information Criterion* : Critère d'information d'Akaike
APA	*American Psychological Association*
BF	Facteur de Bayes
CCI	Courbe Caractéristique de l'Item
CCT	Courbe Caractéristique du Test
CFI	*Comparative Fit Index*
CHC	Cattell-Horn-Carroll
ETE	Erreur Type d'Estimation
ETM	Erreur Type de Mesure
FaN	Faiblesse Normative
FoN	Force Normative
FaP	Faiblesse Personnelle
FoP	Force Personnelle
IM	Indice Moyen
Indices CHC	Gc : Compréhension-connaissances
	Gf : Raisonnement fluide
	Gs : Vitesse de traitement
	Gwm : Mémoire de travail à court terme
	Gv : Traitement visuel
Indices standards du WISC-IV	QIT : Quotient Intellectuel Total
	ICV : Indice de Compréhension Verbale
	IRP : Indice de Raisonnement Perceptif
	IMT : Indice de Mémoire de Travail
	IVT : Indice de Vitesse de Traitement
	IAG : Indice d'Aptitude Générale
	ICC : Indice de Compétence Cognitive
FDI	Fonctionnement Différentiel de l'Item
LCS	*Latent Change Scores*
MoN	Moyenne Normative
MoP	Moyenne Personnelle
OPLM	*One Parameter Logistic Model* : Modèle logistique à un paramètre
QI	Quotient Intellectuel
RMSEA	*Root Mean Square Error of Approximation*
SES	*Socioeconomic status* : Statut socio-économique

Abréviation & Sigle	Signification
Subtests du WISC-IV	CUB : Cubes
	SIM : Similitudes
	MDC : Mémoire des chiffres
	IDC : Identification de concepts
	COD : Code
	VOC : Vocabulaire
	SLC : Séquence lettres-Chiffres
	MAT : Matrices
	COM : Compréhension
	SYM : Symboles
	CIM : Complètement d'images
TCT	Théorie Classique des Tests
TRI	Théorie de la Réponse à l'Item
WISC	*Wechsler Intelligence Scale for Children* : Échelle d'Intelligence de Wechsler pour Enfants et Adolescents