Université de Montréal

Modélisation, élaboration et évaluation de rapports à visée diagnostique des données du PIRLS 2011

par

Dan Thanh Duong Thi
Département d'administration et fondements de l'éducation
Faculté des sciences de l'éducation

Thèse présentée à la Faculté des études supérieures et postdoctorales
en vue de l'obtention du grade de Philosophiae Doctor (Ph.D)
en Sciences de l'éducation, option mesure et évaluation

Septembre, 2018

Université de Montréal

Faculté des études supérieures et postdoctorales

Cette thèse intitulée :

Modélisation, élaboration et évaluation de rapports à visée diagnostique des données du PIRLS 2011

par

Dan Thanh Duong Thi

a été évaluée par un jury composé des personnes suivantes :

Pascale Lefrançois, présidente du jury
Nathalie Loye, directrice de recherche
Michel Laurier, membre du jury
Martin Riopel, examinateur externe
Serge Larivée, représentant de la doyenne

Septembre, 2018

Résumé

Cette étude s'inscrit dans l'approche diagnostique cognitive (ADC), dont la finalité est de fournir aux enseignants des rétroactions fines et détaillées sur les forces et les faiblesses cognitives des élèves, à travers des rapports diagnostiques compréhensibles et interprétables. Elle a été menée dans le but de répondre au grand besoin qu'ont les enseignants de recevoir des informations diagnostiques à partir de tests à grande échelle, et de combler le manque de recherches empiriques sur l'élaboration et l'évaluation de rapports diagnostiques. Plus précisément, notre étude vise à : (1) modéliser des données de 4 762 élèves canadiens ayant fait le test PIRLS 2011 et (2) élaborer et évaluer des rapports diagnostiques destinés aux enseignants à partir des résultats des modélisations.

Afin de répondre au premier objectif, des habiletés sous-jacentes jugées nécessaires pour répondre aux 35 items du PIRLS 2011 ont été identifiées à partir du cadre de référence pour l'élaboration du PIRLS 2011 et avec un panel de trois experts. Deux matrices, Q1 et Q2, représentant les liens entre les items et ces habiletés, ont ensuite été élaborées. Avec ces deux matrices Q, nous avons modélisé des données à l'aide des modèles DINA et G-DINA, ce qui nous a permis d'évaluer l'ajustement des modèles aux données et d'estimer la qualité diagnostique des items ainsi que les profils de maîtrise des habiletés des élèves.

Quant au deuxième objectif de l'étude, un panel de cinq experts a, *dans un premier temps*, reformulé les descriptions des habiletés identifiées dans un langage accessible aux enseignants en salle de classe. *Dans un deuxième temps*, nous avons choisi le profil-type le plus représentatif des élèves parmi les profils de maîtrise des habiletés tirés des modélisations. *Dans un troisième temps*, trois formats de rapports ont été développés à partir de ce profil-type au moyen du protocole de conception de rapports créé grâce à une revue de la littérature. *Finalement*, l'évaluation des rapports a été effectuée auprès de 98 enseignants au primaire, conseillers pédagogiques et orthopédagogues à l'aide d'un questionnaire portant sur la préférence, l'évaluation de la qualité et la compréhension des rapports.

Les résultats obtenus indiquent le format de rapport de préférence des participants, lequel s'explique par la familiarité du type de graphique présenté. Cette préférence est constante, peu importe les informations contextuelles. Les participants évaluent très positivement la qualité des rapports, indépendamment des variables contextuelles. Par contre, il existe une différence significative entre l'évaluation de la qualité des rapports et la préférence des participants. Les participants qui évaluent leur format préféré démontrent des perceptions plus positives. D'autre part, les participants ont une très bonne compréhension des trois formats de rapports. Le groupe ayant le mieux compris les rapports est celui dont l'évaluation de la qualité est la plus positive.

Les résultats de l'étude mettent en évidence la possibilité de recevoir, à travers des rapports accessibles, fiables et compréhensibles, des rétroactions plus fines et détaillées sur les forces et les faiblesses des élèves dans des épreuves à grande échelle. Notre étude est la première à appliquer un cadre de référence pour l'élaboration et l'évaluation de rapports selon une approche diagnostique cognitive, et contribue ainsi à favoriser le soutien à l'apprentissage de la lecture au primaire.

***Mots-clés** : approche diagnostique cognitive, modèle de classification diagnostique, DINA, G-DINA, lecture, PIRLS 2011, élaboration des rapports diagnostiques, évaluation des rapports diagnostiques.*

Abstract

This study uses the cognitive diagnostic approach (CDA) to provide teachers with precise and detailed feedback on the cognitive strengths and weaknesses of students through comprehensible diagnostic reports. It was conducted with the aim of addressing teachers' pressing need for diagnostic information based on large-scale tests, as well as the lack of empirical research on diagnostic report development and evaluation. More specifically, our study aims to: (1) model data from 4,762 Canadian students who took the 2011 PIRLS test and (2) develop and evaluate diagnostic reports for teachers, using modeling results.

In order to attain the first objective, necessary underlying skills to solve the 35 items in PIRLS 2011 were identified based on the PIRLS 2011 development framework and with a panel of three experts. Two templates, Q1 and Q2, were then developed, representing the connexions between these skills and items. With these two "Q" templates, we modeled the data with the DINA and G-DINA models, which enabled us to assess how the models fit the data and to estimate the diagnostic quality of the items as well as the skill mastery profiles of the students.

As for the second objective, a panel of five experts *first* rephrased the identified skills descriptions in accessible terms for teachers in the classroom. *Next*, among the skill mastery profiles obtained from the modeling, we chose the profile that was most representative of the students. *Then*, three report formats were developed based on this standard profile, using the design protocol derived from our literature review. *Finally*, 98 elementary school teachers, educational consultants and remedial teachers evaluated these reports by means of a survey focusing on report preference, report quality evaluation and report comprehension.

The results point to a preferred report format among the participants. This preference is explained by the familiarity of the type of chart presented, and it is constant regardless of contextual information. The participants evaluated the quality of the reports very positively, with no variation according to contextual variables. However, there is a significant difference between report quality evaluation and report preference. Participants who evaluate their preferred format show more positive perceptions. Regarding report comprehension, participants have a very good understanding of all three formats. The group that best understands the reports is the one that made the most positive quality evaluation.

The results of the study point to the possibility of receiving, through accessible, reliable, and comprehensible diagnostic reports, more precise and detailed feedback on the strengths and weaknesses shown by students in large-scale tests. Our study is the first to apply a report development and evaluation framework based on a cognitive diagnostic approach, thus helping improve support for reading acquisition in elementary schools.

Keywords: *cognitive diagnostic approach, diagnostic classification model, DINA, G-DINA, reading, PIRLS 2011, diagnostic report development, diagnostic report evaluation*.

Table des matières

Résumé i

Abstract ii

Table des matières iii

Liste des tableaux vii

Liste des figures x

Liste des sigles xi

Remerciements xiii

Introduction 1

CHAPITRE 1 PROBLÉMATIQUE 3

1.1. Évaluation diagnostique 3

1.1.1. Origine du terme « diagnostic » 3

1.1.2. Définitions de l'évaluation diagnostique 4

1.1.3. Évaluation diagnostique cognitive 7

1.2. Évaluation diagnostique de la lecture 10

1.2.1. Nature de la lecture 10

1.2.2. Évaluation diagnostique cognitive en lecture 14

1. 3. Élaboration et évaluation des rapports diagnostiques 16

1.4. Objectifs et questions générales de la recherche 18

CHAPITRE 2 CADRE THÉORIQUE 21

2.1. Cadre de référence de l'élaboration du test PIRLS 2011 21

2.1.1. Vers une définition de la littératie et de la compréhension de l'écrit 21

2.1.2. Deux objectifs et quatre processus de la compréhension en lecture 23

2.1.3. Liens entre le test PIRLS et les modèles théoriques en lecture 29

2.1.4. Taxonomies des habiletés en lecture 41

2.1.5. Synthèse 46

2.2. Approche diagnostique cognitive 48

2.2.1. Définition de l'approche diagnostique cognitive 48
2.2.2. Études empiriques en lecture qui font appel au MCD 52
2.2.3. Limites de l'approche diagnostique cognitive 68
2.3. Élaboration et évaluation des rapports diagnostiques 69
2.3.1. Fonctions et destinataires des rapports 71
2.3.2. Cadres de référence d'élaboration et d'évaluation des rapports 72
2.3.3. Synthèse 96
2.4. Objectifs et questions spécifiques de la recherche 100
CHAPITRE 3. MÉTHODOLOGIE 104
3.1. Phase 1: Modéliser les données du test PIRLS 2011 des élèves en 4ème année au Canada dans une visée diagnostique 104
3.1.1. Épreuve PIRLS 2011 104
3.1.2. Données 107
3.1.3. Participants 108
3.2. Phase 2: Élaborer et évaluer des rapports diagnostiques destinés aux enseignants 119
3.2.1. Élaboration des rapports diagnostiques 119
3.2.2. Évaluation des rapports diagnostiques par des enseignants, des orthopédagogues et des conseillers pédagogiques 125
3.3. Synthèse de la méthodologie 130
CHAPITRE 4. RÉSULTATS 132
4.1. Élaboration des matrices Q 132
4.1.1. Matrice élaborée à partir du cadre de référence du PIRLS 2011 (Q1) 132
4.1.2. Matrices individuelles originales élaborées par les experts 135
4.1.3. Matrice Q2_finale proposée pour des modélisations 139
4.2. Estimation des paramètres d'items et de profils des élèves 142
4.2.1. Évaluation des ajustements relatifs et absolus des modèles aux données 142
4.2.2. Estimation des paramètres d'items 145
4.2.3. Profils de maitrise des habiletés des élèves 149
4.3. Élaboration des rapports diagnostiques avec le panel d'experts 152
4.3.1. Reformulation des habiletés en lecture 152

4.3.2. Choix d'un profil-type 155
4.4.3. Trois formats de rapports 159
4.4. Perceptions des enseignants au primaire, des orthopédagogues et des conseillers pédagogiques sur les formats de rapports diagnostiques 168
4.4.1. Description de l'échantillon 168
4.4.2. Préférence des enseignants, des orthopédagogues et des conseillers pédagogiques sur les trois formats de rapports 172
4.4.3. Évaluation de la qualité des rapports 191
4.4.4. Évaluation de la compréhension des rapports 216
4.5. Synthèse des résultats 222
CHAPITRE 5. INTERPRÉTATION, DISCUSSION ET CONCLUSION 226
5.1. Interprétation et discussion 226
5.1.1. Éléments qui influencent l'ajustement des modèles aux données et des paramètres d'items 228
5.1.2. Reformulation des habiletés en lecture; équivalence entre le programme de formation en anglais et le document du MELS et ses influences sur la matrice Q et les profils de la maitrise des habiletés des élèves 233
5.1.3. Retour aux cadres de référence de l'élaboration et de l'évaluation des rapports et aux critères d'un bon rapport 237
5.1.4. Discussion sur les perceptions des enseignants au primaire, des conseillers pédagogiques et des orthopédagogues sur les formats de rapports évalués 242
5.2. Limites et recommandations 250
5.3. Apports scientifiques et pratiques et conclusion 253
BIBLIOGRAPHIE i
ANNEXE 1-RÉSULTATS DESCRIPTIFS xvii
ANNEXE 2-VÉRIFICATION DES POSTULATS DE L'ANOVA ET DU TEST T xxx
ANNEXE 3-RÉSULTATS DE L'ANOVA À UN FACTEUR xxxii
ANNEXE 4-RÉSULTATS DU TEST T DE STUDENT xxxvi
ANNEXE 5-DOCUMENTS DE TRAVAIL AVEC LE PANEL D'EXPERTS 2 xl

ANNEXE 6-QUESTIONNAIRE DE L'ÉVALUATION DES RAPPORTS xliii
ANNEXE 7-APPROBATION ÉTHIQUE lviii

Liste des tableaux

Tableau 1. Description des processus de compréhension du test PIRLS 2011 (adapté de Labrecque et al., 2012; Mullis et al., 2009) 45
Tableau 2. Exemple d'une matrice Q du test METLAB (adapté de Li, 2011). 50
Tableau 3. Probabilités de réponse dans le modèle de DINA (adapté de Rupp, Templin et Henson, 2010). 63
Tableau 4. Synthèse des habiletés identifiées dans les recherches diagnostiques en lecture ... 67
Tableau 5. Synthèse des recherches réalisées en lecture avec les MCD 70
Tableau 6. Cadre de référence du développement et de l'évaluation des rapports (adapté de Hambleton et Zinsky, 2010). 74
Tableau 7. Synthèse des critères et principes pour l'élaboration des rapports 83
Tableau 8. Cadre de référence d'élaboration des rapports diagnostiques avec la méthode AHM 89
Tableau 9. Synthèse de la méthodologie des recherches en évaluation des rapports 97
Tableau 10. Pourcentages attribués aux deux objectifs et quatre processus en lecture du PIRLS 2011 (adapté de Labrecque et al., 2012; Mullis et al., 2009)..... 105
Tableau 11. Répartition des items du cahier 13 du PIRLS selon les deux objectifs de la lecture et les quatre processus de compréhension. 107
Tableau 12. Répartition des élèves au Canada selon les cahiers de tests du PIRLS 2011..... 108
Tableau 13. Exemple d'un extrait des classes latentes et leurs probabilités (adapté de Ravand, Barati et Widhiarso, 2012)..... 116
Tableau 14. Exemple de 5 attributs avec leurs probabilités de maitrise (adapté de Ravand, Barati et Widhiarso, 2012)..... 116
Tableau 15. Résumé des éléments méthodologiques pour la première question générale de recherche..... 118
Tableau 16. Résumé des éléments méthodologiques pour le deuxième objectif général de la recherche..... 129
Tableau 17. Description des attributs et la répartition des items par attribut. 133
Tableau 18. Matrice Q1 élaborée à partir du cadre de référence du PIRLS 2011. 134

Tableau 19. Définitions détaillées des attributs 136
Tableau 20. Matrice Q2 initiale élaborée par les experts..... 138
Tableau 21. Matrice Q2_finale proposée pour des modélisations du PIRLS 2011 141
Tableau 22. Ajustements relatifs des modèles aux données avec Q1 et Q2_finale..... 142
Tableau 23. Ajustements absolus des modèles aux données avec la Q1 et Q2_finale 143
Tableau 24. Estimation des paramètres d'items avec la matrice Q2_finale et avec DINA.... 147
Tableau 25. Profil des élèves et des pourcentages des participants..... 150
Tableau 26. Exemple des profils détaillés de 5 élèves de notre base de données..... 151
Tableau 27. Définitions des habiletés proposées par deux panels d'experts..... 154
Tableau 28. Synthèse de 10 profils les plus populaires avec des exemples 156
Tableau 29. Exemples des questions administrées aux enseignants..... 167
Tableau 30. Répartition des participants selon les formats de rapports..... 168
Tableau 31. Répartition des participants selon les informations sociodémographiques 171
Tableau 32. Préférence des participants sur les trois de rapports élaborés..... 172
Tableau 33. Lien entre la préférence des rapports pour la question 1 et les trois autres questions 174
Tableau 34. Lien entre la préférence des rapports et les postes occupées 177
Tableau 35. Lien entre la préférence des rapports et les tranches d'âge 180
Tableau 36. Lien entre la préférence des rapports et l'ancienneté 182
Tableau 37. Lien entre la préférence des rapports et le diplôme obtenu 184
Tableau 38. Lien entre la préférence des rapports et le domaine de formation..... 187
Tableau 39. Lien entre la préférence des rapports et le suivi des cours en méthodes quantitatives 189
Tableau 40. Indice de saturation de 32 items de l'évaluation de la qualité des rapports 193
Tableau 41. Indices de saturation pour les items de l'évaluation des directives..... 195
Tableau 42. Indice de saturation des items de l'évaluation du profil de l'élève. 196
Tableau 43. Indice de saturation des items de la description du profil et des pistes d'intervention 197
Tableau 44. Indices de saturation des items de l'évaluation du guide d'interprétation et de l'ensemble du rapport. 198
Tableau 45. Facteurs proposés par des analyses factorielles**.** 200

Tableau 46. Répartition des participants sur les questions de la préférence..... 211
Tableau 47. Résultats descriptifs de l'évaluation de la compréhension des rapports 218
Tableau 48. Lien entre l'évaluation de la qualité et la compréhension des rapports..... 221

Liste des figures

Figure 1. Exemple d'un processus interactif parallèle de traitement de l'information en lecture (adapté de Carrell et al., 1988). 30
Figure 2. Modèle contemporain en lecture (synthétisé et adapté selon Irvin, 1986, 1991; Giasson, 1996) 38
Figure 3. Modèle de construction-intégration en lecture (adapté de Wharton et Kintsch, 1990). 40
Figure 4. Liens entre les modèles théoriques en lecture et le test PIRLS 2011 47
Figure 5. Étapes de l'approche diagnostique cognitive 52
Figure 6. Cadre de référence du développement et de l'évaluation des rapports (adapté de Zapata-Rivera, VanWinkle et Zwick, 2012) 73
Figure 7. Synthèse des principes de la présentation des informations dans un rapport (synthétisé de Roberts et Gierl, 2010) 85
Figure 8. Exemple d'un profil de maitrise des habiletés d'un sujet issu du test METLAB (adapté de Li, 2012). 117
Figure 9. Schéma de la première page du rapport diagnostique 121
Figure 10. Schéma du guide d'interprétation du rapport diagnostique 122
Figure 11. Estimation des paramètres d'items et leur capacité diagnostique 148
Figure 12. Probabilités de maitrise des habiletés pour l'ensemble des élèves 151
Figure 13. Format A du rapport. 162
Figure 14. Format B du rapport 163
Figure 15. Format C du rapport 164
Figure 16. Guide d'interprétation du format C. 165
Figure 17. Évaluation de la qualité de différentes parties des rapports 201
Figure 18. Évaluation de la qualité des rapports selon le format évalué 202

Liste des sigles

ADC	Approche diagnostique cognitive
AHM	Méthode d'attribut hiérarchique
AIE	Association internationale pour l'évaluation du rendement scolaire
CBAL	Cognitively Based Assessment of, for, and as Learning
CMEC	Conseil des ministres de l'éducation du Canada
DINA	Deterministic Inputs, Noisy and Gate model
EM	Expectation-maximization
G-DINA	Generalized Deterministic Inputs, Noisy and Gate model
GDM	General Diagnostic Model
GET	General english test
IALS	Reading litteracy et l'International Adult Literacy Study
JMLE	Joint Maximum Likelihood Estimation
LCM	Latent Class Models
MCD	Modèle de classification diagnostique
MCMC	Markov Chain Monte Carlo
MELAB	Michigan English Language Assessment Battery
MLE	Maximum Likelihood Estimation
NAEP	National Assessment of Educational Progress
OCDE	Organisation de coopération et de développement économiques
pGDM	Crédit partiel de GDM
PIRLS	Programme international de recherche en lecture scolaire
PISA	Programme international pour le suivi des acquis des élèves
PPCE	Programme pancanadien d'évaluation
PSAT/NMSQT	The Preliminary SAT/National Merit Scholarship Qualifying Test
QCM	Question à choix multiples
RSM	Rule-Space Model
SAT	Standard Assessment Test
TCF	Test de connaissance du français
TOEFL	Test of English as a foreign language
TOEFL iBT	TOEFL Internet-based Test
TOEIC	Test of English for International Communication

À mon oncle et à ma tante qui m'ont appris les premiers mots du français.

Remerciements

Je souhaite d'abord exprimer ma reconnaissance profonde et mes remerciements sincères à ma directrice de recherche, Nathalie Loye pour ses conseils précieux et son soutien académique et moral tout au long de mon parcours doctoral. Son talent de chercheur engagé est une source d'inspiration et ses mots d'encouragements m'ont donné des forces, de la détermination et de la persévérance dans les moments difficiles. Nathalie, je te remercie d'avoir cru en moi et d'être toujours là lorsque j'avais besoin de toi. Je n'oublierai jamais la tasse de café qu'on a bue à Chicago lors du congrès AERA\NCME en 2015. Le fait d'être sous ta direction est l'un des meilleurs choix que j'ai pu faire dans ma vie.

J'exprime également mes immenses remerciements aux membres du jury: Nathalie Loye (Université de Montréal), Michel Laurier (Université d'Ottawa), Pascale Lefrançois (Université de Montréal) et Martin Riopel (Université du Québec à Montréal) pour leurs commentaires précieux et constructifs.

Un grand merci au personnel et aux professeurs du Département d'administration et fondements de l'éducation ainsi qu'au comité des études supérieures de la Faculté des sciences de l'éducation, notamment François Bowen et Claudine Jomphe pour leurs directives dans la recherche du soutien financier pour mon cheminement au doctorat. Un merci particulier vient à Jean-Guy Blais et du GRIÉMÉtic, pour sa confiance et son appui financier lors de mon stage de recherche aux États-Unis ainsi que mes études doctorales.

Je tiens à remercier le professeur Jimmy de la Torre du Département de Psychologie de Rutgers, New Jersey pour son accueil chaleureux lors de mon stage de recherche. Grâce à lui et à son équipe, j'ai énormément appris sur les modèles DINA et G-DINA. Merci de m'avoir donné accès aux codes de DINA et G-DINA pour réaliser des modélisations avec le logiciel OxEdit.

Mes remerciements sincères vont à Josée Lambert-Chan de Brisson-Legris pour sa collaboration précieuse dans le processus de l'élaboration et de l'évaluation des rapports diagnostiques. Josée, j'aimerais te remercier, non seulement parce que tu m'as aidée

énormément dans le recrutement des participants pour mon projet, mais surtout, merci de ta confiance à la beauté du fruit de notre travail.

Je tiens d'ailleurs à reconnaitre le support et les encouragements de mes collègues de doctorat pour les discussions constructives, les moments de partage de nos angoisses, nos doutes et nos stress ainsi que nos réalisations. Un grand merci à Carole B La Grenade et Ben Moustapha Diedhiou d'avoir accepté de faire la révision linguistique de ma thèse. Un gros merci va également à tous mes amis vietnamiens à Montréal d'avoir partagé avec moi les bons moments ainsi que les moins bons de mon cheminement au doctorat.

Enfin, mes plus vifs remerciements à ma famille, mes beaux parents, mes soeurs et surtout à ma mère, qui ne connait aucun mots français, mais qui croit très bien à la beauté de l'apprentissage. Merci de leur soutien et leur amour inconditionnel pendant toutes ces années de mon doctorat. Les remerciements spéciaux à mon mari, Chi Nhan Duong, pour sa patience, sa compréhension et sa confiance dans ce que je fais. Nous avons appris à se connaitre au début de nos parcours au doctorat et nous sommes arrivés à la destination en même temps. J'ai donc hâte de commencer une nouvelle étape de vie avec toi.

Introduction

Au cours des trois dernières décennies, le monde de l'éducation a connu une augmentation des demandes en lien avec l'évaluation diagnostique afin d'améliorer l'enseignement et l'apprentissage des élèves (de la Torre, 2009; Jang, 2009). À titre d'exemple, une enquête nationale menée aux États-Unis auprès de 400 enseignants sur leurs croyances et pratiques liées à l'utilisation des informations diagnostiques des évaluations à grande échelle révèle l'importance qu'ils accordent à recevoir des telles informations afin de réguler l'enseignement et l'apprentissage (Huff et Goodman, 2007; Templin et al., 2010).

Malgré la mise en évidence de besoins d'informations diagnostiques de la part des enseignants et des apprenants (Leighton et Gierl, 2007), il existe peu d'outils qui ont été conçus spécifiquement à visée diagnostique (Alderson, 2010; Jang, 2009; Lee et Sawaki, 2009). C'est pour cette raison que de nombreuses recherches ont été effectuées avec les données issues de diverses épreuves à grande échelle telles que le TOEFL, le TOEFL iBT, le TOEIC ou le MELAB, etc. afin de fournir des informations diagnostiques plus détaillées sur le niveau de compétences des élèves que les informations de classification ou de rangs obtenues habituellement.

Parmi les épreuves à grande échelle administrées actuellement, le Programme international de recherche en lecture scolaire (PIRLS) est une épreuve qui vise à dégager les tendances dans le rendement en lecture des élèves de 4ème année, dans les politiques et dans les pratiques en matière de littératie (Labrecque et al., 2012). En 2011, le test PIRLS a eu lieu dans 45 pays, dont le Canada, et a été conçu à partir d'un cadre de référence basé sur deux objectifs et quatre processus cognitifs en lecture. Ce test a donc un potentiel pour fournir des informations plus fines et détaillées sur les profils des élèves que les scores totaux ou rangs habituellement connus. En d'autres mots, les résultats actuels nous renseignent peu sur la maitrise ou non-maitrise de ces objectifs ou processus cognitifs, et sur les stratégies à mettre en place pour les améliorer.

Afin de combler le manque d'informations diagnostiques lors de la communication des résultats des tests à grande échelle, notre recherche vise à modéliser à visée diagnostique les données en lecture du test PIRLS 2011 auprès de 4762 élèves de 4ème année au Canada. Plus précisément, elle s'intéresse, dans un premier temps, à vérifier la capacité diagnostique de

cette épreuve. Dans un deuxième temps, elle envisage de fournir des profils de maitrise des élèves ayant pris le test. Dans un troisième temps, l'étude vise à élaborer des rapports diagnostiques destinés aux enseignants à partir des profils. Finalement, ces rapports sont évalués par des enseignants au primaire, des conseillers pédagogiques et des orthopédagogues.

Afin de répondre à ces objectifs de recherche, nous avons organisé cette thèse en cinq chapitres. Le premier chapitre est consacré à la problématique ainsi qu'à la présentation des objectifs et questions générales de la recherche. Le deuxième chapitre constitue le cadre conceptuel et théorique dans lequel le cadre de référence de l'élaboration du PIRLS 2011, les modèles théoriques en lecture, l'approche diagnostique cognitive, les cadres de référence de l'élaboration et de l'évaluation des rapports ainsi que les objectifs et questions spécifiques de recherche sont présentés. Le troisième chapitre traite de la méthodologie de notre étude en décrivant le processus de l'élaboration de la matrice Q, du développement et de l'évaluation des rapports ainsi que des techniques du traitement et de l'analyse des données. Le quatrième chapitre vise à présenter les résultats des modélisations des données et de l'élaboration et de l'évaluation des rapports diagnostiques auprès des enseignants au primaire, des conseillers pédagogiques et des orthopédagogues. Finalement, le cinquième chapitre est consacré à la discussion des résultats principaux de la recherche, des limites ainsi qu'aux contributions scientifiques et pratiques de notre étude.

CHAPITRE 1 PROBLÉMATIQUE

Dans ce premier chapitre, nous expliciterons la fonction diagnostique de l'évaluation à travers ses différentes définitions. Vient ensuite la description de la nature de la lecture, de l'évaluation diagnostique cognitive en lecture ainsi que de l'élaboration des rapports diagnostiques. Le chapitre se terminera par l'énoncé des questions générales de notre recherche.

1.1. Évaluation diagnostique

L'évaluation diagnostique constitue un des concepts centraux de notre étude. Il est donc indispensable de commencer cette partie de problématique par le définir. Et, avant d'analyser des définitions de l'évaluation diagnostique proposées par différents auteurs, commençons par expliciter l'origine du terme « diagnostic ».

1.1.1. Origine du terme « diagnostic »

Le terme « diagnostic » vient du mot grec « diagignoskein » qui veut dire « connaitre précisément », « décider », et « mettre d'accord sur... » (Templin et al., 2010). Plus précisément, le diagnostic se définit comme une analyse critique de la nature d'une chose ou d'un phénomène, puis des conclusions tirées à partir de ces analyses. Dans le domaine de la médecine, le diagnostic est considéré comme une action ou un processus qui consiste à identifier ou déterminer la nature ou les causes d'une maladie ou d'une blessure en se basant sur l'historique du patient ou les résultats des tests en laboratoire afin d'arriver à prendre des décisions sur le traitement (Loye, 2008; Templin et al., 2010; Yang et Embretson, 2007). En biologie, le diagnostic est une brève description des caractéristiques distinctes d'un organisme pour des fins de classification taxonomique (Templin et al., 2010). En d'autres termes, le diagnostic est l'acte d'analyser d'une manière précise un problème et d'en identifier les causes afin de prendre des décisions nécessaires.

Malgré les différences dans la définition de ce terme, retenons que le diagnostic doit représenter les trois aspects suivants: (a) une description des caractéristiques distinguées d'un objet ou d'un phénomène; (b) une identification de la nature de cet objet

ou des causes de ce phénomène; et (c) une explicitation des décisions prises ou conclusions tirées à partir de ces descriptions ou analyses. Ces conclusions peuvent être la classification d'un objet dans des catégories spécifiques ou la description du mécanisme d'observation d'un objet (Yang et Embretson, 2007).

En éducation, l'évaluation diagnostique correspond à l'identification des « particularités des élèves à travers leurs réponses aux questions d'un test » (Loye, 2008, p.1). Son rôle va au-delà du « dépistage des élèves en difficulté mais doit permettre de découvrir les forces et les faiblesses ainsi que le degré de participation des élèves avant que ceux-ci n'entreprennent d'une séquence importante d'apprentissage » (Scallon, 2000, p.15). Les définitions de ce concept, son rôle dans le domaine de l'éducation ainsi que ses caractéristiques vont être détaillés dans la partie qui suit.

1.1.2. Définitions de l'évaluation diagnostique

Il existe plusieurs définitions de l'évaluation diagnostique et celles-ci varient d'un auteur à l'autre. Si, dans les ouvrages de base traitant de l'évaluation des apprentissages, les auteurs proposent des définitions générales, dans des documents plus récents sur le sujet, nous pouvons constater une évolution de ces définitions où l'accent est mis davantage sur l'aspect cognitif et psychologique du processus.

Bloom et ses collaborateurs (1971), dans leur « Handbook on formative and summative evaluation of student learning », ont défini l'évaluation diagnostique comme étant celle qui permet de découvrir les forces et les faiblesses des élèves, soit avant l'entrée dans une unité d'apprentissage, soit pendant son déroulement. Ainsi, elle aboutit à des décisions de soutien, de remédiation pour certains ou des décisions d'adaptation de l'enseignement aux caractéristiques des élèves.

La définition de Scallon (1988, 2000) va dans le même sens que celle proposée par Bloom et ses collaborateurs (1971). D'après Scallon (1988, 2000), l'évaluation diagnostique peut être effectuée au début ou pendant le déroulement d'une période d'apprentissage. Si l'évaluation diagnostique est réalisée au début de la formation, elle joue un rôle de prévention dont l'objectif est de déterminer le niveau initial de l'élève. Les décisions prises visent à identifier les rythmes et les méthodes d'apprentissage ou les modalités convenant aux élèves ou à chaque élève. L'évaluation diagnostique peut être

effectuée pendant le processus d'apprentissage lorsque les symptômes persistent. Dans ce cas, au-delà d'aider à identifier des solutions de nature pédagogique, son but est de relever les éléments externes de la situation d'enseignement et d'apprentissage qui peuvent influencer le processus comme les problèmes de santé, la situation familiale, les intérêts, la motivation, etc. L'évaluation diagnostique joue alors une fonction corrective.

En parallèle avec ces fonctions de prévention et de correction telles que mentionnées dans les définitions précédentes, la définition de Nadeau (1978) met l'accent sur la fonction de placement de l'évaluation diagnostique. Il reconnait qu'elle peut jouer un rôle important dans le processus de l'enseignement et de l'apprentissage, car elle permet de: (1) déterminer la présence ou l'absence d'habiletés jugées nécessaires (prérequis) pour l'apprentissage d'une nouvelle unité d'enseignement ; (2) déterminer le niveau de maitrise des objectifs d'un cours en vue de situer l'élève au point de départ le plus approprié et (3) classer les élèves dans les groupes distincts selon certaines caractéristiques telles que l'intérêt, la personnalité, l'aptitude ou toute autre variable reconnue comme étant reliée à une stratégie particulière d'enseignement ou à un type d'apprentissage donné.

Selon Roegiers (2004), l'évaluation diagnostique est considérée comme une évaluation d'orientation, car elle vise à détecter les forces et les faiblesses d'un élève en vue d'y remédier ou de l'orienter vers un type d'apprentissage approprié. Ainsi, à la différence des auteurs cités précédemment qui croient que l'évaluation diagnostique est effectuée avant et pendant une période d'apprentissage, l'évaluation diagnostique selon Roegiers (2004) intervient seulement avant cette période.

Cette définition rejoint le point de vue de Figari et Achouche (2001) qui considère l'évaluation diagnostique comme une « évaluation avant l'action » qui exerce une fonction de prévision. Elle peut être considérée comme une évaluation initiale jouant aussi le rôle de diagnostic ou pronostic. Dans certains cas, elle est aussi « l'évaluation du contexte », car elle permet de définir l'environnement en cause, d'identifier des besoins à combler, de diagnostiquer les problèmes qui empêchent que ces besoins soient atteints. Dans une perspective macro-analytique, ce type d'évaluation s'intéresse à établir les limites du système, à décrire ses valeurs et ses buts à étudier et à déterminer dans quelle mesure la pratique est en accord avec les principes d'une théorie pertinente. Dans ce sens,

la définition proposée par Figari et Achouche (2001) se rapproche de celle du diagnostic de Yang et Embretson (2007) et de Templin et al., (2010) dans le sens qu'elle s'intéresse à identifier la nature d'un objet ou à déterminer les caractéristiques d'un phénomène ainsi que son mécanisme de fonctionnement.

Les définitions plus récentes de l'évaluation diagnostique s'intéressent davantage aux aspects cognitifs du processus de l'apprentissage. Selon Legendre (2005), l'évaluation diagnostique vise à détecter les forces et les faiblesses des apprenants dans leurs apprentissages permettant de proposer des pistes d'intervention pour remédier aux difficultés rencontrées. D'après Loye (2008), le but ultime des évaluations diagnostiques est de dresser le portrait individuel de chaque élève. Enfin, d'après Laplante (2011), l'objectif de l'évaluation diagnostique est de comprendre les processus sous-jacents impliqués par des élèves dans la résolution des tâches en compréhension écrite et les stratégies compensatoires permettant de proposer des pistes d'intervention pédagogique appropriées. Toujours selon Loye (2008), lors de la proposition et de la mise en place des pistes d'intervention, il est possible de regrouper des élèves selon les types de difficultés ou de forces complémentaires, notamment dans le contexte des évaluations à grande échelle.

Templin et al., (2010), en se basant sur la documentation en médecine clinique ainsi qu'en psychopédagogie, ont défini l'évaluation diagnostique comme étant un processus systématique qui vise à obtenir les informations spécifiques sur les caractéristiques psychologiques d'une personne en utilisant des méthodes variées. Son but est de justifier, de contrôler, de prendre des décisions et de proposer des solutions appropriées. Ainsi, la définition de Templin et ses collaborateurs (2010) est bien liée au diagnostic des problèmes de la vie quotidienne. Elle souligne qu'un processus de diagnostic efficace devrait être systématique et ancré dans une théorie. Par ailleurs, l'efficacité de ce processus réside non seulement dans l'identification des symptômes ou des causes du problème, mais aussi dans l'intervention et l'évaluation régulière de ces méthodes.

L'évaluation diagnostique peut prendre place aussi bien dans le domaine de la psychologie, lors de l'évaluation des entreprises ou des besoins des employés, dans le domaine de l'éducation que dans le domaine des sciences de la santé. Bien que ses objectifs diffèrent d'un domaine à un autre, elle possède, selon Templin et al., (2010),

sept caractéristiques communes qui sont: (1) l'objet de l'évaluation, (2) l'objet des décisions prises, (3) le moment ciblé par l'évaluation, (4) l'objectif de l'évaluation, (5) les méthodes d'évaluation, (6) les types d'intervention et (7) les différences de pouvoir entre les agents.

En résumé, à la différence de l'évaluation formative qui intervient seulement pendant le déroulement d'une période d'apprentissage et joue plutôt le rôle de régulation de l'enseignement et de l'apprentissage, l'évaluation diagnostique intervient avant et pendant cette période et a pour fonction l'orientation, le placement, la prévention ou la correction. Elle s'intéresse non seulement aux facteurs endogènes du processus tels que les interventions pédagogiques ou les styles d'apprentissage des élèves, mais aussi aux facteurs exogènes de ce processus comme les problèmes de santé des élèves, leur situation familiale, leurs intérêts et leurs attentes, ou encore les caractéristiques du système. Quant à l'évaluation diagnostique cognitive, elle relève d'un processus systématique et ancré dans des théories d'apprentissage avec des modèles cognitifs sous-jacents dont nous parlerons plus en détail dans les lignes suivantes.

1.1.3. Évaluation diagnostique cognitive

Il est à noter qu'avec les développements récents de la psychologie cognitive, les outils d'évaluation diagnostiques tiennent compte du processus cognitif de l'apprentissage. En d'autres termes, ils sont donc conçus pour mesurer les structures de connaissances spécifiques à chaque élève afin de leur fournir des rétroactions sur leurs points forts et leurs faiblesses cognitives en se basant sur un cadre de référence comprenant trois éléments essentiels: (1) le modèle d'apprentissage de l'élève dans un domaine spécifique; (2) un ensemble de croyances ou d'hypothèses sur les types d'observations des habiletés définies par ce modèle et (3) un cadre de référence permettant d'interpréter les résultats d'évaluation (Leighton et Gierl, 2007; Nichols, 1994). Ainsi, Nichols (1994) ajoute que l'évaluation diagnostique permet de comprendre d'une façon plus explicite le processus que l'élève a utilisé pour résoudre un problème, la manière dont il se développe et comment distinguer les élèves les plus compétents.

Une des forces de l'évaluation diagnostique cognitive réside dans la précision des informations fournies. Si les tests traditionnels sont construits dans une approche

psychométrique qui met l'accent sur les qualités psychométriques des items afin de classer les élèves sur une échelle de mesure fiable comme dans le cas de l'évaluation certificative, l'évaluation diagnostique cognitive permet de mesurer les structures sous-jacentes de connaissances et des compétences spécifiques de l'élève afin de lui fournir des informations très fines sur ses forces et faiblesses cognitives (Jang, 2009; Leighton et Gierl, 2007; Sawaki et al., 2009; Yang et Embretson, 2007). De plus, selon Scallon (1988) et Loye (2008), les tests diagnostiques permettent de faire une description plus détaillée de la performance des élèves en fonction d'un profil qualitatif et non pas au moyen d'un score total. Ainsi, elle s'intéresse non seulement aux éléments « intérieurs » du processus d'apprentissage tels que les styles d'apprentissage, l'intérêt de l'élève, le processus de raisonnement ou de résolution des problèmes, les types d'intervention pédagogique, mais aussi aux éléments extérieurs, à savoir, le milieu et l'ambiance de la classe. Par ailleurs, la conception des outils d'évaluation basés sur les modèles et les théories de l'apprentissage ainsi que les recherches empiriques sur le domaine nous permettent d'obtenir des informations plus détaillées sur les points forts et les faiblesses cognitives des élèves.

De plus, la description du processus d'évaluation diagnostique nous démontre qu'il s'agit d'un processus qui est très rigoureux dans sa démarche (Templin et al., 2010). En effet, on peut y constater la cohérence ou l'interdépendance entre les différentes étapes, car les résultats d'une étape influencent ceux de l'étape suivante. Par exemple, l'évaluation diagnostique d'un élève ayant des difficultés en mathématiques doit passer par cinq étapes systématiques suivantes. *Premièrement*, le travail consiste à établir l'anamnèse, un profil qualitatif de l'élève comprenant les informations sur son cheminement scolaire, les résultats acquis en maths ainsi que le développement du statut émotionnel ou social de cet enfant. *Deuxièmement*, un diagnostic de base sera effectué en administrant des tests standardisés ou des questionnaires visant à élaborer un profil de l'élève en comparaison avec un groupe de référence afin d'identifier des problèmes de déficit ou de santé. *Troisièmement*, ce diagnostic de base est transmis en diagnostic différentiel. En d'autres termes, l'élève est évalué plus en profondeur dans le domaine ciblé à l'aide d'outils spécifiquement conçus. *La quatrième étape* consiste à établir un rapport synthèse des informations obtenues sur la base duquel des personnes responsables

du développement sont consultées afin de suggérer des solutions de remédiation appropriée. *Enfin*, l'évaluation de ces outils de traitement utilisés pour cet élève afin de prendre de nouvelles décisions, soit de terminer le processus diagnostique si le traitement est jugé efficace, soit chercher d'autres moyens mis en place (Templin et al., 2010). Dans une évaluation diagnostique cognitive, la construction du modèle et de la théorie de l'apprentissage va décider du choix des observations et celui-ci, à son tour va exercer une certaine influence sur la conception des items. Par ailleurs, les résultats obtenus à partir de l'administration des tests vont permettre d'évaluer le modèle et la théorie proposés lors de la première étape (Templin et al., 2010).

Toutefois, afin d'assurer la qualité des informations fournies et la rigueur entre les étapes, l'évaluation diagnostique exige une longue démarche de collecte de données qui demande beaucoup de temps et d'énergie de la part des personnes responsables, ce qui constitue une des difficultés importantes de la mise en œuvre de ce type d'évaluation. Selon Yang et Embretson (2007), McKenna et Stahl (2012), Templin et al., (2010), afin d'assurer la justesse d'une évaluation diagnostique, une très grande attention doit être accordée à la conception de l'épreuve, à la pré-validation de l'épreuve ainsi qu'à l'analyse et l'évaluation des résultats obtenus. Pour y parvenir, il faut avoir une grande collaboration entre les élèves, les enseignants, les parents, des spécialistes en mesure et en évaluation ainsi que l'équipe du cycle ou de l'école (Henson et al., 2007; Templin et al., 2010).

Par ailleurs, la réalité du développement des évaluations diagnostiques ainsi que la validation de certains modèles d'évaluation diagnostique cognitive dans les contextes concrets montrent qu'il manque une théorie adéquate et des outils qui peuvent nous permettre de poursuivre les recherches dans ce domaine (Jang, 2010). C'est pour cette raison que jusqu'à présent, la plupart des recherches en évaluation diagnostique cognitive dans le domaine des langues sont réalisées avec des tests standardisés à grande échelle qui n'ont pas été conçus spécifiquement à cette visée.

Malgré ces limites, l'évaluation diagnostique cognitive joue un rôle de plus en plus important dans le processus d'enseignement/apprentissage, en s'intéressant non seulement aux éléments endogènes du processus de l'enseignement et de l'apprentissage, mais également à des facteurs exogènes tels que le contexte de l'enseignement et de

l'apprentissage, les besoins, les intérêts et les styles d'apprentissage des élèves, etc. Ainsi, elle permet de dresser un profil très fin, détaillé et complet sur les points forts et les faiblesses des élèves.

Ainsi, dans cette première partie, nous avons dressé un aperçu général de l'évaluation diagnostique et de l'évaluation diagnostique cognitive, les définitions ainsi que les caractéristiques lors de la mise en application de ces évaluations. Dans le contexte de la présente recherche, nous nous intéressons à l'évaluation diagnostique de la lecture. Ainsi, la partie suivante est réservée à l'explicitation de la nature de la lecture ainsi que l'évaluation de cette compétence dans une approche diagnostique cognitive.

1.2. Évaluation diagnostique de la lecture

Dans cette partie, nous commencerons par la description de la nature de la compétence en lecture en dressant un aperçu très général des modèles et théories qui sous-tendent les différentes conceptions de la lecture au fil du temps. Certains modèles et théories mentionnés dans cette partie seront présentés plus en détail dans notre cadre conceptuel parce qu'ils représentent des liens explicites avec le cadre de référence du test PIRLS 2011. Nous aborderons ensuite l'évaluation diagnostique cognitive de la lecture.

1.2.1. Nature de la lecture

La lecture est une compétence complexe et multidimensionnelle et il s'avère difficile de dresser un aperçu général de toutes les études traitant de sa nature (Alderson, 2005). Et, les défis des recherches sur la compétence en lecture dépendent également des perspectives théoriques et des paradigmes dominants selon les différents contextes historiques et politiques (Jang, 2005).

Au début du 20^{e} siècle, avec la dominance du ***paradigme béhavioriste*** qui cherche à rejeter la possibilité de l'implication des processus mentaux sous-jacents à la lecture (Hiebert et Raphael, 1996) et voulant que seuls les comportements observables puissent être décrits (St-Pierre et al., 2011), la lecture est conçue comme un ensemble de variables quantifiables: les associations entre des lettres et des sons, la capacité de reconnaissance des mots. Selon Clapham (1996), Langer et Allignton (1992) et Jang (2005), l'objectif principal est de découvrir les sous-habiletés isolées de la lecture pour comprendre

littéralement des textes décontextualisés et d'expliquer les relations entre des variables observables telles que le décodage des symboles ou la reconnaissance des mots.

À la différence du paradigme béhavioriste, les théories des ***sciences cognitives*** conceptualisent la lecture comme un ensemble de mécanismes mentaux qui s'enclenchent chez l'individu et qui peuvent être étudiés hors de leur contexte habituel d'utilisation (St-Pierre et al., 2011; Sfard, 1998). Ainsi, les théoriciens en lecture ont porté l'attention sur différents modèles mentaux de la lecture tels que les modèles ***cognitifs de traitement des informations*** et ***la théorie du schéma*** (Jang, 2005). Selon Gough (1977), les modèles cognitifs de traitements des informations suggèrent que les mots doivent être décodés visuellement avant la construction du sens et que le processus de la lecture implique de multiples étapes hiérarchiques de la fixation des yeux, des stimulations visuelles, de la reconnaissance des mots, de la production phonologique au décodage d'une série de mots. Cette conception de la lecture est liée au processus ***«Bottom-up »*** dans lequel le lecteur est le décodeur passif des systèmes graphiques, phonologiques, syntaxiques et sémantiques (Alderson, 2005).

Toutefois, plusieurs auteurs tels que Anderson et Pearson (1984), Carell (1983), Carell et al. (1988), et Hudson (1982), en rejetant cette approche « Bottom-up » de la lecture, ont mis l'accent sur l'importance des connaissances antérieures en lecture et initié la ***théorie du schéma*** qui souligne l'interaction entre les différents types de connaissances lors de la compréhension d'un texte (Bartlett, 1932 ; Adams et Collins, 1979 ; Rumelhart, 1978 ; An, 2013). Sous la lumière de cette théorie du schéma, l'approche ***« Top-down »*** qui reconnait l'importance des schémas et la contribution du lecteur dans la compréhension du texte a été proposée (Alderson, 2005). En effet, Goodman (1967) considère la lecture comme « un jeu de devinettes psycholinguistique » dans lequel le lecteur devine ou prédit le sens du texte en se basant sur un minimum d'informations textuelles et un maximum de connaissances acquises antérieures. De plus, Smith (1971) prétend que les informations non visuelles transcendent le texte et incluent les expériences du lecteur avec les connaissances antérieures, la familiarité avec les structures de la langue et des types de textes spécifiques ainsi que des connaissances générales sur le monde et des connaissances spécifiques sur le sujet. Autrement dit, si l'objectif de la lecture est de comprendre le texte, le lecteur construit directement du sens

en utilisant ses connaissances et ses expériences antérieures (Jang, 2005). Goodman (1976) suggère aussi que la construction de sens demande l'utilisation active des indices grapho-phoniques, syntaxiques et sémantiques. Le lecteur n'est pas un identificateur passif des lettres, mais un constructeur actif de ses propres connaissances. La théorie du schéma sera présentée plus en détail dans notre cadre conceptuel.

Toutefois, les approches « Bottom-up » et « Top-down » ne semblent pas suffisantes pour caractériser les processus de lecture. Ainsi, les théories récentes proposent des ***modèles interactifs*** de la lecture en tenant compte à la fois de l'interaction de ces deux approches où chaque composante du processus de lecture peut interagir avec une autre composante (Alderson, 2005; Grabe, 1991, 2000; Jang, 2005; Rumelhart, 1976; Stanovich, 1980). La lecture passe donc d'un processus de traitement de l'information en série à un processus de traitement de l'information parallèle (Grabe, 1991). Par exemple, le modèle de Rumelhart (1978) intègre des mécanismes de rétroactions permettant l'interaction entre les différentes sources de connaissances (linguistiques et sur le monde) et l'entrée visuelle (visual input). Ainsi, dans son modèle, l'hypothèse finale proposée pour le texte est synthétisée à partir de multiples sources de connaissances qui interagissent continuellement et simultanément. D'autre part, Stanovich (1980) a développé le ***modèle interactif compensatoire*** qui suggère que le degré d'interaction entre les composantes dépend du déficit des connaissances dans chaque composante séparée. Des lecteurs de faible niveau pour ce qui est de la reconnaissance des mots peuvent donc utiliser les connaissances « Top-down » pour compenser (Alderson, 2005). Grabe (1991) ajoute également que le terme « interactif » ne réfère pas seulement à l'interaction entre les approches « Bottom-up » et « Top-Down », mais aussi à l'interaction entre le lecteur et le texte en mettant l'accent sur l'importance du rythme de lecture comme un indice de lecture efficace, car un taux inférieur à 200 ou à 300 mots par minute peut engendrer des problèmes de compréhension en lecture (Grabe, 2000; Jang, 2005).

Les études modernes en lecture distinguent différents niveaux de compréhension en lecture. En effet, selon Kintsch et Yarbrough (1982), il est possible de comprendre les mots, mais pas le sens du texte et les phrases, ni l'organisation du texte. Ainsi, van Dijk et Kintsch (1983) ont proposé le ***modèle de traitement du discours*** (discourse processing

model) comme un modèle interactif dans lequel le sens des phrases est représenté par des propositions qui sont liées les unes aux autres (Jang, 2005). Ils séparent en deux niveaux le traitement du discours: les ***micro-processus*** qui traitent la compréhension locale, la reconnaissance phrases par phrases du texte et les ***macro-processus*** portant sur la compréhension globale du texte. Dans ce modèle de traitement du discours, les auteurs proposent deux niveaux de compréhension: la compréhension du sens littéral du texte (Van Dijk et Kintsch, 1983) et le dynamisme cognitif de la compréhension avec le modèle de la ***construction-intégration*** (Kintsch, 1988). Ce modèle de construction-intégration assume que l'architecture cognitive du lecteur (par exemple la limite de la mémoire), les procédures cognitives (la récupération) ainsi que les dispositifs du texte (par exemple, les arguments) soutiennent la compréhension (Perfertti et Stafura, 2014). Ce modèle de construction-intégration va être mieux décrit dans la partie suivante parce qu'il représente des liens explicites avec la conception du test PIRLS 2011 qui est au cœur de notre recherche.

D'autres auteurs considèrent cette activité comme ***une pratique socioculturelle***. En effet, la lecture n'est pas une activité isolée qui peut se réaliser dans un milieu vide. Par contre, elle est faite pour un but précis dans un contexte social qui peut lui-même contribuer à former ce que signifie la lecture (Alderson, 2005). Dans ce point de vue, la lecture n'est pas une activité mentale individuelle, mais plutôt un accomplissement social des membres de la communauté des lecteurs (Baker, 1991; Barton, 1994; Jang, 2005; Street et al., 2001). Dans cette communauté, des relations sociales, des pratiques culturelles et des habitudes différentes de la lecture exercent des influences fortes sur les processus de lecture (Health, 1980). Ainsi, la lecture peut porter des significations différentes pour les membres de différents groupes ethniques, avec une variété de modes, de fonctions et d'usage (Health, 1980; Jang, 2005). Ce point de vue socioculturel en lecture pose comme défi la possibilité d'avoir différentes interprétations du sens des textes dans l'élaboration et l'évaluation des épreuves standardisées à grande échelle (Jang, 2005).

Dans le cadre du PIRLS, avec la mise en place de deux enquêtes de l'AIE reading litteracy et l'International Adult Literacy Study (IALS, 1994), l'AIE a relié le terme « reading » au terme « literacy » et leur traduction en français devient « compréhension »

et « écrit » (Labrecque et al., 2012). On peut ainsi « associer la capacité de réfléchir à ce qu'on lit à la capacité d'utiliser la lecture comme outil permettant d'atteindre des objectifs individuels et sociétaux » (Labrecque et al., 2012, p.6; Mullis et al., 2009). En associant la compréhension et les usages sociaux, l'AIE a adopté, pour la première fois, une définition de « reading litteracy » ou la « compréhension de l'écrit » en français (Labrecque et al., 2012) comme étant « la capacité de comprendre et d'utiliser ces formes du langage écrit requises par la société ou importantes pour l'individu » (Elley, 1994, p.5). Cette définition de la compréhension s'ouvre à une littératie fonctionnelle avec une variété de formats de textes dits non continus qui sont liés à la vie quotidienne comme supports d'évaluation: cartes graphiques, tableaux, petites annonces, etc. Toutefois, cette définition s'arrête à la compréhension et ne comprend pas les aspects plus critiques ou réflexifs (Lafontaine, 2001). En développant le cadre de référence pour le test PIRLS, les concepteurs se basent principalement sur la définition de la compréhension de l'écrit proposée par l'AIE en y ajoutant l'aspect actif et réflexif des lecteurs lors de la construction du sens. Nous reviendrons sur la notion de la littératie et la définition de la compréhension de l'écrit dans notre cadre théorique.

1.2.2. Évaluation diagnostique cognitive en lecture

La compréhension de la lecture est une compétence cognitive qui a été beaucoup étudiée par les chercheurs dans les domaines des langues, de la psychologie cognitive, de la théorie des curriculums ou même de la psychométrie. Chaque discipline adopte un point de vue théorique différent sur la structure et le processus cognitif de la lecture (Svetina et al., 2011). À titre d'exemple, certains chercheurs tels que Carver (1992), et Rost (1993) ont souligné que la lecture se composait d'une seule structure globale simple ou d'une construction bi-divisible comprenant la lecture globale et le vocabulaire. L'évaluation diagnostique de la lecture basée sur ces deux composantes peut fournir des informations utiles pour le diagnostic des lecteurs débutants, faibles ou dyslexiques (Alderson, 2005). Cependant, elle ne permet pas aux chercheurs d'identifier les variations des problèmes en lecture et les sources de ces problèmes (Jang, 2005, 2009). Ainsi, selon d'autres auteurs comme Alderson, Bachman, Perkins et Cohen, (1991), et Weir et al., (2000), la compétence de la lecture doit être multi-divisible, car elle implique une vaste

gamme de processus, connaissances et habiletés comme les processus de décodage, le traitement de la structure linguistique et des processus cognitifs d'ordre supérieur. Par exemple, Hudson (1996) et Jang (2005) ont proposé que les habiletés suivantes devaient être identifiées dans les processus de lecture : (a) l'automaticité dans la reconnaissance des mots et des phrases ; (b) le contexte et le schéma (c'est-à-dire, la forme, le contenu) ; (c) les stratégies et les habiletés métacognitives, et (d) l'objectif et le contexte de lecture. Ainsi, la lecture devrait couvrir un ensemble d'habiletés: de la compréhension textuelle locale aux interprétations globales des textes et l'inférence (Hudson, 1996; Jang, 2005). L'évaluation diagnostique de la lecture devrait aussi permettre l'identification des forces et des faiblesses pour chacune de ces habiletés. Ces habiletés et sous-habiletés en lecture vont être décrites plus en détail dans notre cadre conceptuel.

Dans le contexte du test PIRLS 2011, par exemple, les concepteurs ont relié deux termes « la compréhension » et « écrit » pour exprimer une notion générale qui peut définir la capacité de lire qui « associe la capacité de réfléchir à ce qu'on lit à la capacité d'utiliser la lecture comme outil permettant d'atteindre des objectifs individuels et sociétaux » (Labrecque et al., 2012, p. 6). Depuis 1991, le terme « compréhension de l'écrit » a été utilisé par l'Association internationale pour l'évaluation du rendement scolaire (AIE) pour désigner la capacité de lire dans une étude sur « reading litteracy » qui fournit des éléments fondamentaux dans la conception du cadre d'évaluation du test PIRLS 2011 (Labrecque et al., 2012). Cette définition de la compréhension de l'écrit adoptée pour le PIRLS 2011 qui considère la lecture comme un processus *constructif et interactif* (Alexander et Jetton, 2000; Anderson et Pearson, 1988; Chapelle et al., 1997; Mullis et al., 2009) ainsi que ses liens explicites avec les modèles théoriques en lecture vont être décrits plus en détail dans notre cadre conceptuel.

Avec les avancements des théories sur l'acquisition du langage et de la cognition, de nouvelles initiatives en évaluation diagnostique des compétences langagières en général et de la lecture en particulier ont émergé (Alderson, 2005; Jang, 2005, 2009; Kunnan et Jang, 2009). En effet, l'évaluation diagnostique cognitive a été récemment développée afin de répondre à une grande demande d'informations formatives qui peuvent être obtenues à partir des tests. Son objectif principal est de fournir des informations diagnostiques formatives basées sur une compréhension profonde des

compétences des élèves à partir d'un ensemble d'habiletés et de sous-habiletés plus fines (Jang, 2005, 2009; Leighton et Gierl, 2007; Nichols, 1994). Selon Jang (2005, 2009), la particularité de l'évaluation diagnostique cognitive est l'établissement des liens explicites entre les compétences en termes de structures latentes d'intérêt et des caractéristiques des items des tests utilisés pour diagnostiquer les habiletés.

Sous la lumière de cette approche, un grand nombre de recherches empiriques ont été effectuées en lecture à partir des tests de langues à grande échelle, à savoir celles de Buck, Tatsuoka et Kostin (1997); Jang (2005, 2009); Lee et Sawaki (2009); Li (2011); Li et Suen (2013); Sawaki, Him et Gentile (2009); von Davier (2008), etc. Avec l'application des modèles diagnostiques cognitifs tels que le Rule-Space (Buck et Tatsuoka, 1998), le Fusion (Hartz, 2002) et le modèle diagnostique général (GDM) (Davier, 2008), le DINA (de la Torre, 2009; Haertel, 1989; Junker et Sijtsma, 2001) et le G-DINA (de la Torre, 2011), ces recherches ont permis d'identifier les points forts et les faiblesses cognitives de chaque individu et de leur fournir les rétroactions plus détaillées sur leur degré de maitrise des connaissances et des habiletés permettant ainsi aux enseignants de prendre des mesures de remédiation appropriées (Lee et Sawaki, 2009). Ces recherches ont établi un pont entre les épreuves standardisées à grande échelle et l'approche diagnostique cognitive permettant de mieux exploiter et d'utiliser ces tests à des visées diagnostiques.

1. 3. Élaboration et évaluation des rapports diagnostiques

L'approche diagnostique cognitive est conçue spécifiquement pour mesurer les structures de connaissances et des processus cognitifs des habiletés des élèves. Ainsi, sa finalité est de fournir au public visé des rétroactions formatives sur leurs forces et leurs faiblesses à travers des rapports diagnostiques détaillés (Roberts et Gierl, 2010). La valeur d'un rapport réside dans la cohérence entre le rapport, le public visé et les intentions précédemment envisagées. Il peut être conçu pour les apprenants, les enseignants, les parents, les directions d'école ou les conseils d'administration. C'est pour cette raison qu'un rapport approprié et utile pour un groupe ne l'est pas nécessairement pour l'autre (Ryan, 2006).

Aux États-Unis, dans le cadre du programme « No Child Left Behind », les résultats individuels doivent maintenant être communiqués à tous les élèves qui y participent. Plus spécifiquement, une des demandes du programme consiste à produire des rapports individuels descriptifs, interprétables et qui peuvent jouer le rôle de diagnostic permettant aux enseignants, aux parents, aux directeurs d'école de mieux comprendre les niveaux de maitrise des compétences des élèves ainsi que leurs besoins spécifiques. Ces informations doivent être présentées dans un format uniforme et interprétable et dans un langage que les parents peuvent comprendre (Goodman et Hambleton, 2004; Ryan, 2006).

Malgré le rôle important des rapports dans le processus d'évaluation, la revue de littérature sur le sujet montre qu'il existe peu de recherches empiriques qui ont été réalisées sur l'élaboration et l'évaluation de ces rapports (Roberts et Gierl, 2010; Ryan, 2006). Toutefois, des recherches modestes, nous ont montré une image très consistante de l'efficacité ou de l'inefficacité dans la communication des résultats de certains rapports comme dans le cas des recherches de Aschbacher et Herman (1991), de Zenisky, Hambleton et Sireci (2009). La conclusion générale tirée de ces recherches est que les utilisateurs rencontrent de grandes difficultés dans l'interprétation des résultats des rapports issus des évaluations à grande échelle (Ryan, 2006; Roberts et Gierl, 2010). De plus, ces recherches ont identifié des problèmes majeurs lors du développement des rapports qui peuvent être regroupés en cinq grandes catégories: (1) présenter une quantité excessive d' informations comme les différents types de scores et omettre des informations essentielles comme les objectifs du test, et la façon dont les résultats du test peuvent être obtenus et utilisés; (2) ne pas fournir d'informations sur le précision des scores du test; (3) utiliser un jargon statistique; (d) ne pas définir les mots clés ou ne pas fournir de guides pour l'interprétation; et (5) fournir une grande quantité d'information dans un espace limité du rapport, ce qui le rend plus dense et difficile à lire (Roberts et Gierl, 2010).

Suite à ces critiques, les auteurs ont proposé un ensemble de principes afin d'assurer une bonne qualité des rapports qui peuvent être regroupés en 4 grands critères, à savoir : l'accessibilité, l'utilité, la lisibilité et la validité (Aschbacher et Herman, 1991; Goodman et Hambleton, 2004; Klesch, 2010; Roberts et Gierl, 2010; Ryan, 2006;

Sinharay et al., 2010; Zenisky et al., 2009). Ces principes et ces quatre critères seront mieux explicités dans notre cadre conceptuel.

Dans les rapports élaborés à visée diagnostique, les informations présentées, y compris la description des habiletés et des concepts de l'apprentissage, sont fondamentalement différentes de celles qui figurent dans les rapports issus des tests à grande échelle tels que les scores totaux ou des rangs percentiles. Les concepteurs doivent donc déterminer et présenter de nouveaux types d'informations provenant de ces tests diagnostiques. En bref, le défi des rapports diagnostiques réside dans l'intégration des informations substantives et techniques correspondant aux besoins du public visé aux informations sophistiquées provenant des modélisations de l'approche diagnostique cognitive (Roberts et Gierl, 2010).

1.4. Objectifs et questions générales de la recherche

Avec la demande croissante des évaluations à grande échelle dans l'enseignement et l'apprentissage au primaire, il y a une forte pression pour rendre ces évaluations plus informatives avec le diagnostic sur les points forts et les faiblesses cognitives des élèves (Chiu et al., 2009; Leighton et Gierl, 2007). Par ailleurs, les recherches empiriques en évaluation diagnostique cognitive nous suggèrent qu'il est possible de décomposer la compétence en lecture en un ensemble de connaissances et d'habiletés possibles à diagnostiquer grâce à des modélisations psychométriques (Jang, 2005; Leighton et Gierl, 2007). Ainsi, des épreuves standardisées à grande échelle peuvent permettre des rétroactions diagnostiques utiles sur les points forts et les faiblesses des apprenants, ce qui peut exercer des influences positives sur l'enseignement et l'apprentissage (Gierl et al., 2008; Hartz, 2002; Jang, 2005; Templin et Henson, 2006). Toutefois, la plupart des recherches en langues réalisées avec l'approche diagnostique cognitive n'utilisent que des tests administrés aux adultes tels que le TOEFL, le TOEFL iBT, le TOEIC ou le MELAB. Rares sont les études qui ont choisi un test en lecture destiné aux élèves au primaire comme le cas du test PIRLS 2011.

Le test PIRLS, administré tous les cinq ans, permet aux pays participants d'accéder à des informations importantes sur le rendement en lecture des élèves en 4ème année au primaire et de faire des comparaisons à l'échelle internationale. C'est la tranche d'âge qui

marque une transition importante dans l'apprentissage de la lecture où les élèves ont déjà appris à lire et utilisent maintenant la lecture pour apprendre (Labrecque et al., 2012). Il est à noter que le PIRLS est le seul test international qui évalue le rendement en lecture au primaire au Canada et qui fournit « un moyen sans équivalent d'obtenir les données sur le rendement en lecture de 4e », ce qui permet de les comparer au rendement des élèves des autres provinces et de 45 pays (Labrecque et al., 2012, p.5).

L'instrument est administré pendant une durée de 40 minutes comprenant à la fois des questions à choix multiples et des questions de court développement. Les élèves sont aussi invités à remplir un questionnaire de 30 minutes sur leurs habitudes personnelles en lecture (Labrecque et al., 2012; Mullis et al., 2009). Au Canada, les résultats du PIRLS ne servent qu'à des fins de recherche et à l'élaboration des politiques. De plus, ils ne figurent pas dans le relevé de notes des élèves, puisque le Conseil des ministres de l'Éducation du Canada (CMEC) n'attribue aucun résultat à titre individuel à un élève, à une école ou à un conseil ou une commission scolaire. Chaque province peut choisir une approche différente pour la divulgation des résultats et des informations (Labrecque et al., 2012). Ces contraintes limitent ainsi l'exploitation et l'utilisation des résultats du test pour améliorer l'enseignement et l'apprentissage en lecture dans les classes au primaire.

Dans le but de répondre à ce besoin de recevoir des informations diagnostiques lors de la communication des résultats des tests à grande échelle, notre recherche vise, dans un premier temps, à modéliser les données du test PIRLS 2011 des élèves en 4ème année au Canada pour des visées diagnostiques afin d'établir la capacité diagnostique de ces épreuves et de fournir des profils détaillés sur les niveaux de compétences des candidats. Ainsi, afin de répondre à cet objectif, notre première question générale de recherche est la suivante :

À quel diagnostic peut-on aboutir en modélisant les données du test PIRLS des élèves de la 4ème année au Canada?

Par ailleurs, avec les contributions cruciales des évaluations à grande échelle dans l'enseignement et l'apprentissage au primaire, une grande attention a été accordée à concevoir des épreuves qui peuvent répondre aux besoins du public et du monde professionnel. Toutefois, peu d'attention est accordée à l'organisation et à la communication des résultats (Goodman et Hambleton, 2004). Ce qui a conduit à des

confusions de la part des responsables, des éducateurs et du public en général dans la compréhension et dans l'interprétation des certains résultats des évaluations à grand échelle faute des connaissances statistiques suffisantes (Hambleton et al., 2002; Jaeger, 1998; Vezzu et al., 2012). De plus, même si ces rapports sont très importants pour les enseignants, les parents et les élèves, les recherches sur l'élaboration et de l'évaluation de ces rapports ne soulignent pas leur rôle actif lors de ce processus (Vezzu et al., 2012).

Ainsi, afin de pouvoir fournir aux enseignants des rétroactions plus fines et détaillées sur les forces et les faiblesses des élèves et de combler le manque des rapports diagnostiques destinés aux enseignants, notre recherche vise, dans un deuxième temps, à élaborer et à évaluer des rapports diagnostiques à partir des résultats obtenus des modélisations avec un guide d'interprétation des résultats. Notre recherche tient compte également de la participation active des enseignants lors de l'élaboration et de l'évaluation des rapports. Nous avons donc posé la deuxième question générale de notre recherche :

Comment peut-on élaborer et évaluer des rapports diagnostiques destinés aux enseignants en tenant compte de leur participation active lors de ce processus?

Ces questions générales de recherche seront éclaircies grâce à notre cadre conceptuel portant sur les théories et les modèles cognitifs expliquant les processus de la lecture et l'approche diagnostique cognitive ainsi que les processus de l'élaboration des rapports diagnostiques. Elles vont être également raffinées par les objectifs et questions de recherche spécifiques, qui seront au cœur du chapitre suivant.

CHAPITRE 2 CADRE THÉORIQUE

Cette recherche s'intéresse à la modélisation des données en lecture du test PIRLS 2011 et à l'élaboration et évaluation des rapports issus de cette modélisation à visée diagnostique destinés aux enseignants. Ainsi, dans ce chapitre, nous allons, dans un premier temps, définir le cadre de référence du test PIRLS 2011 avec ses deux objectifs et quatre processus de la compréhension en lecture et expliciter les liens de ce test avec les modèles théoriques en lecture, à savoir les modèles interactifs, le modèle contemporain en lecture, la théorie du schéma et le modèle de construction-intégration en lecture. Dans un deuxième temps, nous décrirons l'approche diagnostique cognitive et présenterons la recension des écrits relative aux recherches empiriques en lecture effectuées à la lumière de cette approche. Nous présenterons également les cadres de référence d'élaboration et d'évaluation des rapports diagnostiques. Finalement, nous conclurons le chapitre par l'énoncé des objectifs et des questions spécifiques de notre recherche.

2.1. Cadre de référence de l'élaboration du test PIRLS 2011

Dans les cadres de référence du test PIRLS, nous observons très souvent les termes « literacy » et « reading literacy » pour désigner la littératie au sens large et la compréhension de l'écrit au sens spécifique (la lecture). L'utilisation de ces deux termes de manière interchangeable dans des écrits et des cadres de référence, amène la nécessité de les distinguer.

2.1.1. Vers une définition de la littératie et de la compréhension de l'écrit

En 1971, l'AIE a mené, pour la première fois, une recherche sur la compréhension des textes comprenant deux études séparées soit la compréhension en lecture et l'appréciation des oeuvres littéraires. L'étude portant sur la compréhension en lecture s'ouvre sur les aspects cognitifs et repose sur un modèle explicite du lecteur avec le principe « get meaning », autrement dit, le lecteur récupère le sens unique déposé dans le texte (Lafontaine, 2001). L'étude traitant de la littérature se concentre davantage sur les aspects esthétiques et affectifs, sur la dimension réflexive ou critique et sur l'interprétation; elle fait référence au courant « response to litterature » dans lequel la

réponse est définie comme « *l'interaction* entre l'individu et l'œuvre qui peut se prolonger longtemps après la lecture » (Lafontaine, 2001; Puvres, 1973, p.36). Toutefois, selon Lafontaine (2001), le fait de distinguer la lecture en deux composantes - les aspects cognitifs et culturels et la compréhension et l'interprétation- a créé une rupture dans l'évaluation alors qu'elles devraient être réunies, d'où le concept de « littératie ». Ainsi, ce concept permet d'articuler « les aspects sociaux et les aspects culturels aux aspects développementaux » (Grossman, 1999, p.146) et de se questionner sur « des frontières communément admises, parce qu'elles correspondent à des partages institutionnels entre cognitif et culturel, mais aussi interdisciplinaires, entre psychologie, anthropologie et linguistique » (Grossman, 1999, p.144). Ce concept de « littératie » veut donc désigner « l'ensemble des habiletés, des comportements et des attitudes reliées à l'écrit » (Pierre, 1992, p.3).

Dans le contexte du PIRLS 2011, la compréhension de l'écrit se définit comme « l'aptitude à comprendre et à utiliser les formes écrites de la langue qui sont exigées par la société ou valorisées par l'individu. Les jeunes peuvent construire du sens à partir de toutes sortes de textes. Ils lisent pour apprendre, pour participer aux communautés de lecture à l'école et dans la vie de tous les jours et par plaisir » (Labrecque et al., 2012, p.6; Mullis et al., 2009). Dans le cadre de notre travail, les deux termes « compréhension de l'écrit » et « la lecture » sont utilisés alternativement pour désigner la même compétence.

Cette définition du test PIRLS se base sur les théories qui articulent la lecture comme un processus ***constructif et interactif*** (Alexander et Jetton, 2000; Anderson et Pearson, 1984, 1988; Labrecque et al., 2012; Mullis et al., 2009; Rudell et Unrau, 2004; Walter, 1999). Les lecteurs sont des *constructeurs actifs* du sens parce qu'ils maitrisent des stratégies efficaces en lecture qu'ils appliquent dans ce processus (Afflerbach et Cho, 2009; Clay, 1991; Mullis et al., 2009; Langer, 1990). Ils mobilisent un ensemble de connaissances linguistiques, de stratégies cognitives et métacognitives ainsi que leurs acquis antérieurs (Labrecque et al., 2012). D'ailleurs, la construction du sens est réalisée dans l'*interaction* entre *le texte* et le *lecteur* dans *un contexte* particulier en lecture (Mullis et al., 2009; Snow, 2002; Grabe, 1991). Ce contexte en lecture devrait favoriser

l'engagement et la motivation de lire pour répondre à des besoins spécifiques du lecteur (Mullis et al., 2009).

Il existe des similitudes entre les définitions de la lecture du test PIRLS, celle du Programme pancanadien d'évaluation (PPCE) et celle du programme international pour le suivi des acquis des élèves (PISA) bien que les populations ciblées soient différentes dans les trois programmes. Ces définitions soulignent la nature interactive et constructive de la lecture (Labrecque et al., 2012). Ainsi, dans le PISA, la compréhension de l'écrit est définie par « comprendre, utiliser des textes écrits, mais aussi réfléchir à leur propos, pour réaliser des objectifs personnels, développer ces connaissances, ses potentialités, et participer à la vie en société » (OCDE, 2011, p.23). L'accent est mis sur les aspects réflexifs et l'esprit critique des lecteurs qui jouent un rôle actif et génératif dans la construction du sens (Lafontaine, 2001). De façon semblable au PIRLS et au PISA, le PPCE définit la lecture comme « un processus dynamique et interactif par lequel le lecteur construit le sens d'un texte » (CMEC, 2011, p.39). Ainsi, la lecture est décrite comme étant le résultat d'une interaction entre le lecteur, le texte, l'intention et le contexte (Labrecque et al., 2012). Dans le contexte du test PIRLS 2011, cette vision interactive et constructive en lecture se traduit en deux objectifs et quatre processus de compréhension en lecture qu'il est important d'aborder en détail dans les lignes suivantes.

2.1.2. Deux objectifs et quatre processus de la compréhension en lecture

Le PIRLS 2011 examine **deux objectifs de la lecture** : (1) lire pour l'expérience littéraire et (2) lire pour acquérir et utiliser des informations. Ces objectifs font intervenir **quatre processus en lecture** : (1) se concentrer sur les informations énoncées de façon explicite et les extraire du texte; (2) faire des inférences simples; (3) interpréter et combiner des idées et des informations et (4) examiner et évaluer le contenu, le langage et les éléments textuels. Ces objectifs et processus ne fonctionnent pas d'une manière isolée l'un de l'autre ou du contexte dans lequel les élèves vivent et apprennent (Labrecque et al., 2009).

Étant donné que ces deux objectifs sont importants pour la tranche d'âge de 9 à 10 ans, ce qui marque une transition importante dans le développement en lecture des élèves,

le PIRLS 2011 accorde un nombre égal de questions dans l'évaluation de chaque objectif. Bien que chaque question puisse distinguer les deux objectifs de lecture, les processus et les stratégies utilisés par les élèves sont plus ou moins similaires dans les deux objectifs (Mullis et al., 2009; Labrecque et al., 2012). Les élèves peuvent, par exemple, utiliser le même processus « se concentrer sur les informations énoncées de façon explicite et les extraire du texte » pour atteindre les deux objectifs. Chaque objectif de lecture est associé à certains types de textes, par exemple, la lecture pour l'expérience littéraire est souvent réalisée avec les textes fictifs tandis que lire pour acquérir et utiliser des informations est généralement associé à des textes informatifs. Ainsi, avec ces deux objectifs, une variété de formats de textes peut être utilisée. Ces textes diffèrent dans la manière dont les idées sont organisées et présentées et suscitent des façons différentes de construire du sens (Goldman et Rakestraw, 2000; Kobayashi, 2002; Mullis et al. 2009). Autrement dit, le contenu, l'organisation et le style typique d'un genre de texte peuvent influencer l'approche des lecteurs pour comprendre les textes (Alexander et Jetton, 2000; Mullis et al., 2009; Weaver et Kintsch, 1996). C'est dans l'interaction entre le lecteur et le texte que la construction du sens est réalisée et que ces deux objectifs en lecture sont atteints (Mullis et al., 2009). Cette idée rejoint celle des modèles interactifs (Rumelhart, 1977) et du modèle contemporain en lecture (Irwin, 1986; Giasson, 1996).

La lecture précoce de la plupart des jeunes enfants est centrée sur la littérature et les textes narratifs. En outre, de nombreux jeunes lecteurs acquièrent des informations à partir des livres ou d'autres types de textes. Ce genre de lecture devient plus important lorsque les élèves développent leurs habiletés en lecture pour répondre aux exigences du curriculum (Duke, 2004; Langer, 1990; Palincsar et Duke, 2004; Labrecque et al., 2012). Ainsi, un large éventail de types de textes est utilisé pour chaque processus en lecture, ce qui vise à créer une expérience de lecture aussi semblable que possible à des expériences authentiques de lecture que les élèves peuvent avoir dans leur vie quotidienne (Mullis et al, 2009).

2.1.2.1. Deux objectifs en lecture

Il existe une multitude de raisons pour lesquelles nous lisons : pour les intérêts personnels, pour le plaisir, pour participer à la société et pour l'apprentissage. Pour les

jeunes lecteurs dans le cadre du test PIRLS, l'accent est mis sur la lecture pour le plaisir et la lecture pour l'apprentissage qui se présentent sous forme de deux objectifs : lire pour l'expérience littéraire et lire pour acquérir et utiliser des informations. Chacun de ces deux objectifs correspond à un type de texte différent proposé dans le PIRLS 2011.

Lire pour l'expérience littéraire

La fiction est le type de texte le plus souvent utilisé pour évaluer l'expérience littéraire (Labrecque et al., 2012). En effet, lors de la lecture de ce type de texte, les élèves « se plongent dans des évènements, des actions, des personnages et des idées inventées tout en appréciant la langue elle-même » (Labrecque et al., 2012, p.8; Mullis et al., 2009). Afin de comprendre et d'apprécier la littérature, le lecteur doit apporter au texte ses propres expériences et sentiments, l'appréciation du langage et ses connaissances sur des formes littéraires (Mullis, 2009). Cette idée reflète une des caractéristiques fondamentales du modèle de construction-intégration en lecture (Kinstch et al., 1990) et de la théorie du schéma (Alderson et Pearson, 1984, 1988) lorsque le lecteur, en lisant, doit mobiliser des propres expériences en lecture et les différents types de connaissances afin d'interpréter et de créer le sens du texte.

L'utilisation des textes de fictions narratifs (nouvelles, romans, etc.) permet aux élèves d'explorer des situations qu'ils ne pourraient pas encore rencontrer dans la vie quotidienne et d'approfondir leur réflexion sur ces situations (Labrecque et al., 2012). Le texte peut présenter le point de vue du narrateur ou d'un personnage principal, ou parfois plusieurs points de vue dans un texte plus complexe. Les informations et les idées peuvent être décrites directement ou par un dialogue ou une description des évènements. Des histoires courtes ou des romans racontent parfois les évènements chronologiquement, d'autres utilisent un temps plus complexe avec des décalages temporels (Mullis et al, 2009).

Lire pour acquérir et utiliser des informations

En lisant pour acquérir et utiliser des informations, les élèves s'impliquent dans les faits de l'univers réel plutôt que dans des mondes imaginaires. À travers des textes informatifs, les élèves peuvent comprendre le fonctionnement du monde réel et pourquoi les évènements se produisent de telle façon (Mullis et al., 2009). Les différences dans l'organisation des textes informatifs peuvent demander aux jeunes lecteurs de mobiliser

différentes stratégies cognitives et métacognitives pour répondre aux questions. Par exemple, certains textes ne comportent pas de titres ou d'organisateurs textuels. Parfois, ils sont organisés selon un ordre logique plutôt que chronologique et ne doivent pas être nécessairement lus du début jusqu'à la fin puisque les élèves peuvent choisir les parties dont ils ont besoin pour répondre à des questions. Ce type de texte informatif inclut les récits factuels, les tests procéduraux (recettes, instructions), les textes persuasifs (annonces publicitaires, etc.) (Mullis et al., 2009). L'influence des éléments textuels sur la compréhension en lecture a été mentionnée dans les modèles interactifs en lecture (Rumelhart, 1977) ainsi que le modèle contemporain en lecture (Irvin, 1986; Giasson, 1996) qui met l'accent sur l'interaction entre le lecteur et le texte dans un contexte de lecture particulier.

2.1.2.2. Quatre processus en lecture

Dans la compréhension en lecture, le lecteur peut construire le sens de différentes manières : se concentrer et récupérer des idées spécifiques; faire des inférences; interpréter et intégrer des informations et des idées; examiner et évaluer les caractéristiques du texte. Afin de mettre en œuvre ces processus, le lecteur doit utiliser des stratégies métacognitives lui permettant d'examiner et d'ajuster sa compréhension **(**Kintsch et Kintsch, 2005; Van Dijk et Kintsch, 1983; Jacobs, 1997; Pressley, 2006; Mullis et al., 2009). En outre, les connaissances et les expériences acquises permettent une compréhension de la langue, des textes et du monde et de filtrer les informations à partir de différentes sources (Alexander et Jetton, 2000; Beach et Hynds, 1996; Clay, 1991; Mullis et al., 2009), ce qui reflète les principes fondamentaux de la théorie du schéma (Alderson et Pearson, 1984).

Dans le contexte du test PIRLS 2011, les quatre processus de compréhension ont été identifiés à partir des textes présentés aux élèves. Il est à noter que ces quatre processus sont tous sollicités à des degrés divers dans les différents types de textes. Le fait que chacune des questions posées correspond à un processus différent en lecture permet aux élèves de démontrer un ensemble d'habiletés afin de construire du sens à partir des textes (Mullis et al, 2009). Ces questions sont conçues dans l'interaction entre la longueur, la complexité du texte et les caractéristiques sophistiquées des processus en

lecture (Labrecque et al., 2012; Mullis et al., 2009). En effet, il se peut que le processus de repérage et d'extraction des informations explicites du texte soit plus facile que, par exemple, faire des interprétations à travers un texte entier et les intégrer aux idées et expériences externes du lecteur. Cependant, les textes peuvent varier énormément dans la longueur, la complexité syntaxique, le niveau d'abstraction des idées et la structure organisationnelle. Ainsi, la nature du texte peut exercer des impacts substantiels sur la difficulté des questions posées dans les quatre processus en lecture (Mullis et al., 2009).

Dans le processus « ***se concentrer et récupérer des informations explicites*** », les lecteurs utilisent différentes façons de repérer et de comprendre le contenu pertinent pour la question posée (Mullis et al., 2009). Extraire des informations appropriées du texte exige que le lecteur comprenne non seulement ce qui est énoncé explicitement dans le texte, mais aussi comment ces informations sont reliées à celles recherchées (Mullis et al., 2009). Ainsi, le repérage des informations explicites demande une compréhension automatique du texte sans que les lecteurs recourent à l'interprétation ou à l'inférence. Ce processus de compréhension en lecture reste une compréhension au niveau des phrases ou des propositions (Labrecque et al., 2012), ce qui correspond à une compréhension locale ou microstructure telle que présentée dans le modèle de la construction-intégration en lecture (Kintsch, 1988) dont nous parlerons plus en détail dans la partie suivante.

Lors de la construction de sens, les lecteurs vont ***faire des inférences*** sur les idées et les informations qui ne sont pas énoncées explicitement leur permettant d'aller au-delà de la surface du texte et de combler les lacunes du texte (Labrecque et al., 2012). Certaines inférences sont simples et se basent principalement sur les informations fournies dans le texte. Bien que les informations soient explicitement déclarées dans le texte, la connexion entre elles doit être inférée. Bien souvent, les lecteurs habiles font ces inférences automatiquement en reliant les informations les unes aux autres et en reconnaissant leurs relations (Mullis et al., 2009). Ce processus correspond à la fois à une compréhension globale (la macrostructure) et à une compréhension locale proposée dans le modèle de la construction-intégration en lecture (Kintsch, 1988).

Afin ***d'interpréter et d'intégrer les informations*** dans le texte, les lecteurs se basent sur leurs connaissances antérieures sur le monde leur permettant de faire des connexions implicites selon leur propre perspective, ce qui reflète bien la théorie du

schéma (Alderson et Pearson, 1984, 1988). Dans ce processus, les lecteurs doivent mobiliser plus de connaissances et d'expériences antérieures que lorsqu'ils font des inférences simples. C'est pour cette raison que le sens construit à travers l'intégration et l'interprétation des idées et des informations varie selon les lecteurs, tout dépendant des expériences et connaissances mobilisées pour les tâches (Labrecque et al., 2012; Mullis et al., 2009).

Lors du processus « ***Examiner et évaluer le contenu, la langue et les éléments textuels »***, l'intérêt se déplace de la construction de sens vers la considération critique du texte lui-même (Mullis et al., 2009). Ceci confirme l'aspect réflexif et le regard critique en lecture proposé par l'AIE dans sa définition sur la compréhension de l'écrit. Dans ce processus, les lecteurs doivent « prendre du recul par rapport au texte, afin de porter un regard critique sur son contenu, le langage utilisé ou les éléments textuels » (Labrecque et al., 2012, p.9; Mullis et al., 2009). Ils doivent, par exemple, comparer la compréhension d'un mot dans le texte à des informations en provenance d'autres sources (Labrecque et al., 2012). L'évaluation des éléments textuels porte sur la clarification du sens de certaines expressions en se basant sur le genre du texte et sa structure ou ses conventions linguistiques (Labrecque et al., 2012). Ce processus fait partie de la compréhension globale du texte. Les processus sont mis en œuvre par l'ensemble des tâches demandées pour chacun des processus qui vont être analysés plus en détail dans la partie sur les taxonomies des habiletés en lecture.

En somme, les lignes précédentes servent à décrire le cadre de référence du test PIRLS 2011. Plus précisément, nous avons explicité la définition de la littératie et de la compréhension de l'écrit proposée par l'AIE, les similitudes entre celle-ci et les définitions du PISA et du PPCE. Dans le contexte du PIRLS 2011, la lecture est décrite comme un processus constructif et interactif dans lequel, le lecteur est l'agent actif dans la construction du sens. Avant, pendant et après la lecture, le lecteur doit mobiliser un ensemble de connaissances linguistiques, de stratégies cognitives et métacognitives ainsi que les connaissances antérieures sur le monde afin d'atteindre ses objectifs. D'ailleurs, la construction du sens est réalisée dans l'interaction entre le texte, le lecteur et le contexte de la lecture. Cette vision interactive et constructive en lecture se traduit par deux objectifs et quatre processus de compréhension en lecture. La partie suivante sert à

décrire les liens explicites entre cette vision interactive et constructive et les modèles théoriques sous-jacents en lecture.

2.1.3. Liens entre le test PIRLS et les modèles théoriques en lecture

La vision interactive et constructive en lecture peut faire référence à quatre modèles : les modèles interactifs (Rumelhart, 1977), le modèle contemporain en lecture (Irwin, 1986; Giasson, 1996), la théorie du schéma (Alderson et Pearson, 1984, 1988) et le modèle de construction-intégration en lecture (Kintsch, 1988).

2.1.3.1. Modèles interactifs en lecture

Si, au début du 20e siècle, la lecture est considérée comme une activité comprenant des étapes hiérarchiques, suivant un processus en série comprenant des phases non interactives, les développements plus récents en lecture suggèrent que cette activité devrait être conçue comme une activité interactive, dont le processus de traitement de l'information est parallèle (Grabe, 1991; Rumelhart, 1977). En général, le terme « interactif » peut référer à deux conceptions différentes (Grabe, 1991): (a) l'interaction entre *le lecteur et le texte* (Carrell, 1983; Grabe, 1991) et (b) l'interaction entre *des habiletés et des types de connaissances de différents niveaux* (Carrell, 1988; Dubin et al., 1986; Grabe, 1991; Rayner et Pollatsek, 1998; Samuels et Kamil, 1984). Ainsi, différentes perspectives de recherches ont été développées en se basant sur ces deux manières de concevoir l'interaction en lecture. Les psychologues cognitifs semblent s'intéresser à l'interaction entre les habiletés en lecture tandis que les chercheurs en langue seconde mettent plus d'accent sur l'interaction entre le texte et le lecteur (Grabe, 1991).

Le premier type d'interaction est celui entre le texte lu et le lecteur. Dans le processus de lecture, les caractéristiques du lecteur telles que les connaissances, les stratégies cognitives et métacognitives, les caractéristiques physiques, les motivations pour la lecture et la manière dont ces caractéristiques interagissent avec les raisons pour lesquelles l'élève fait la lecture influencent le résultat de ce processus (Alderson et Bachman, 2005). En ce qui concerne le texte, les recherches de Grabe (1984, 1986) et Biber (1986) suggèrent que les éléments textuels se combinent interactivement pour créer ce qu'on appelle la « textualité » qui devrait être traitée par le lecteur (Carrell et al.,

1988). En effet, selon Alderson (2005), ces éléments textuels comme le contenu du texte, les types ou les genres de texte, l'organisation du texte et la structure des phrases, la typographie du texte, la relation entre le texte verbal et non verbal (les images ou les figures, par exemple) et enfin les moyens par lesquels le texte est présenté facilitent ou rendent plus difficile la compréhension du texte.

Le deuxième type d'interaction est celui entre les différents types de connaissances et d'habiletés mobilisées dans un processus en lecture que Rumelhart (1977) est le premier chercheur à introduire. Ces connaissances ou habiletés peuvent se distinguer en différents niveaux (inférieur et supérieur) qui varient de la reconnaissance des caractéristiques graphiques des mots à l'interprétation du texte comme l'inférence. Une des hypothèses principales de ce type d'interaction est que dans un processus en lecture, le traitement de l'information se compose de différentes étapes parallèles qui interagissent simultanément et continuellement au lieu d'étapes hiérarchiques qui se passent les unes après les autres (Carrell et al., 1988). La figure 1 illustre un exemple simple du processus interactif parallèle de traitement de l'information dans laquelle la partie à gauche représente les différents niveaux de traitement de l'information et celle à droite représente le processus en lecture.

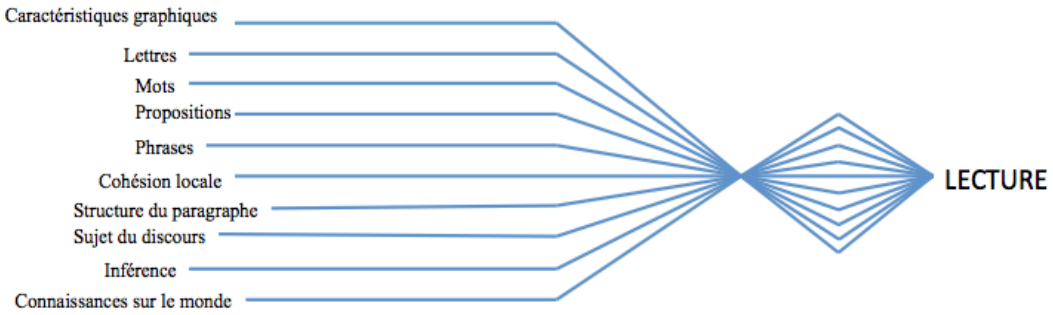


Figure 1. Exemple d'un processus interactif parallèle de traitement de l'information en lecture (adapté de Carrell et al., 1988).

Toutefois, il n'y a pas de modèle interactif unique en lecture mais plutôt une famille de modèles interactifs (Carrell et al., 1988). À la suite de la vision interactive en lecture proposée par Rumelhart (1977), plusieurs modèles interactifs ont été développés, à savoir

le *modèle interactif-activation* de McClelland et Rumelhart (1981), le *modèle interactif compensatoire* de Stanovich (1980) et le *modèle de l'efficience verbale* de Perfetti (1988). Ces modèles sont très utiles pour comprendre le processus en lecture tant en première langue qu'en langue seconde (Carrell et al., 1988).

Le modèle interactif-activation de McClelland et Rumelhart (1981) a été développé en se basant sur les recherches de la reconnaissance des mots et sur le modèle d'activation mentale et de récupération des informations de Morton (1979, cité par Carrell et al., 1988). Le processus d'activation est essentiellement l'étape où toutes les caractéristiques personnelles telles que les connaissances sur les lettres, le contexte, les connaissances syntaxiques, sémantiques, le sujet du discours et les connaissances antérieures activent des groupes de sens ou la sélection de la compréhension (Carrell et al., 1988; McClelland et Rumelhart, 1981; Van Dijk et Kintsch, 1983). Ainsi, l'automaticité de ce processus nous permet de nous concentrer sur la compréhension plutôt que la sélection active ou la prédiction des mots (Carrell et al., 1998, Leong, 1982; McClelland et Rumelhart, 1981; Van Dijk et Kintsch, 1983).

Stanovich (1980) a aussi introduit la notion « activation » dans son modèle interactif compensatoire en lecture qui assume que l'interprétation d'un texte est synthétisée à partir des informations fournies simultanément provenant de toutes sources de connaissances et qu'il y a une compensation entre les différentes sources de connaissances ou de stratégies au cours de ce processus. Ce modèle constitue une des balises très importantes pour différencier les lecteurs forts des plus faibles. Un lecteur qui est faible dans une stratégie, par exemple, va compter sur une autre stratégie pour compenser ses déficits, ce qui explique le fait que les lecteurs faibles en reconnaissance des mots vont se baser principalement sur les éléments contextuels pour interpréter le texte (Alderson, 2005; Carrell et al., 1988). Il est à noter que les modèles interactifs en lecture sont nés pour compléter les limites des approches « Bottom -up » et « Top-down » en représentant également l'interaction entre les stratégies de ces approches (Carrell et al., 1988). C'est pourquoi ces modèles fournissent une conceptualisation plus exacte des performances en lecture des élèves que ces deux approches (Stanovich, 1980). En outre, lorsqu'ils sont combinés à une hypothèse de traitement compensatoire de l'information, ces modèles interactifs offrent une meilleure prise en compte des données qui existent sur

l'utilisation de la structure orthographique et du contexte des phrases chez de bons lecteurs et chez ceux qui sont plus faibles (Carrell et al., 1988; Stanovich, 1980).

Lorsque nous parlons de l'interaction entre les différents types de connaissances, une des balises théoriques importantes est ***la théorie du schéma*** (Bartlett, 1932; Adams et Collins, 1979; Alderson et Pearson, 1984, 1988; An, 2013; Rumelhart, 1980). Ces différents types de connaissances, y compris des connaissances linguistiques, des connaissances sur le monde, organisées sous forme des schémas (Alderson, 1977; Minsky, 1975; Rumelhart, 1976) font l'objet de nombreuses recherches sur la contribution de ces connaissances dans la compréhension en lecture (Carrell, 1981; Carrell, 1983; Hudson, 1982; Mcvee, Dunsmore et Gavelek, 2005). Dans une telle théorie, le schéma se définit comme *une structure des données pour représenter des concepts génériques stockés dans la mémoire* (Rumelhart, 1978) ou *une structure de connaissances abstraite* (Alderson et Pearson, 1984) ou tout simplement *une structure générale de connaissances utilisées pour la compréhension*. Le principe fondamental de la théorie des schémas suppose que le texte ne porte pas le sens en lui-même. Par contre, il ne fournit que des indications au lecteur lui permettant de récupérer ou de construire du sens à partir des connaissances acquises précédemment (connaissances antérieures), donc de ses propres schémas (Bartlett, 1932; Adams et Collins, 1979 ; Rumelhart, 1980; An, 2013). Les schémas d'un lecteur sont organisés d'une manière hiérarchique, du plus général, en haut, au plus spécifique, en bas. La compréhension d'un texte est donc une interaction entre les connaissances antérieures du lecteur (background knowledge) et le texte. Ainsi, une compréhension efficace repose sur la capacité de relier les éléments fournis par le texte à ces connaissances antérieures (An, 2013). Alderson (1977) a souligné également que tout acte de compréhension implique aussi des connaissances de l'individu sur le monde.

Dans le contexte du test PIRLS 2011, les concepteurs mettent plus d'accent sur le premier type d'interaction, soit l'interaction entre le lecteur et le texte. En effet, les lecteurs, lors de la lecture, doivent mobiliser un ensemble des connaissances, des stratégies cognitives et métacognitives pour répondre aux tâches demandées. Le texte contient certains éléments linguistiques et structurels et se focalise sur un sujet particulier. Les lecteurs peuvent apprendre à partir de multiples types et formes de textes

incluant des formes traditionnelles comme les livres, les revues et les journaux, mais aussi des sources technologiques: l'internet, le courriel, les messages téléphoniques, les médias ou la télévision (Leu, et al., 2004; Mullis et al., 2009). Ceci affirme bien l'importance de différents types et formes des textes qui peuvent influencer l'intérêt et la compréhension du texte, d'où vient l'idée de la contribution des variables textuelles à la compréhension en lecture telle que présentée dans les modèles interactifs en lecture. Malgré le rôle important de ces facteurs dans le processus de compréhension du texte, les études suggèrent que l'analyse textuelle n'est souvent réalisée que dans une perspective linguistique sans tenir compte du lecteur (Alderson, 2005; Carrell et al., 1988).

Toutefois, l'interaction en lecture telle qu'abordée dans le PIRLS ne s'arrête pas au niveau d'interaction entre le texte et le lecteur sans tenir compte du contexte. Autrement dit, la construction du sens doit se faire dans l'interaction entre trois composantes: le texte, le lecteur et le contexte. La référence théorique de l'interaction de ces trois composantes est affirmée dans le modèle contemporain en lecture (Irwin, 1986; Giasson, 1996) que nous allons décrire en détail dans les lignes suivantes.

2.1.3.2. Modèle contemporain en lecture

Le modèle contemporain en lecture d'Irvin (1986, 1991) articule la compréhension en lecture comme l'interaction entre trois composantes : **le lecteur, le texte et le contexte (**Deschênes, 1986; Irvin, 1986, 1991 et Giasson, 1996). Ces trois composantes sont indissociables et le niveau de compréhension en lecture varie selon le degré de relation entre ces trois variables. Plus les variables sont imbriquées les unes dans les autres, meilleure sera la compréhension (Giasson, 1996).

Texte

La composante **« texte »** englobe toutes les caractéristiques qui sont liées au texte, à savoir l'intention de l'auteur et le genre littéraire ainsi que la structure et le contenu du texte. La structure fait référence à la façon dont les idées sont organisées dans un texte alors que le contenu renvoie aux thèmes, aux concepts présentés dans le texte (Giasson, 1996). Les caractéristiques du texte ont un effet important sur la compréhension, parce qu'elle ne se produit pas simplement par l'extraction du sens à partir du texte. Pendant la lecture, le lecteur construit différentes représentations du texte qui sont importantes pour

sa compréhension comme le code de surface (libellé exact du texte), la base du texte (des unités d'idées représentant le sens) et une représentation des modèles mentaux noyés dans le texte (Snow, 2002).

Le degré de difficulté dans l'interprétation des textes dépend des facteurs inhérents au texte et de la relation entre celui-ci et les connaissances et les habiletés du lecteur et des activités dans lesquelles il est engagé. Lors de la lecture, les connaissances du domaine du lecteur interagissent avec le contenu du texte. En plus du contenu, la charge du vocabulaire du texte, la structure linguistique, le style du discours et le genre du texte interagissent aussi avec les connaissances du lecteur. Lorsqu'il y a trop de facteurs qui ne correspondent pas à ses connaissances et ses expériences, le texte lui parait trop difficile pour une compréhension optimale (Snow, 2002).

Contexte

Le **contexte** comprend les conditions dans lesquelles se trouve le lecteur lorsqu'il est en contact avec le texte. Ces conditions incluent le **contexte psychologique** du lecteur, le **contexte social** et le **contexte physique**. *Le contexte psychologique* concerne les facteurs propres au lecteur lui-même comme la motivation, l'intérêt et l'intention de lecture. *Le contexte social*, quant à lui, englobe toutes les formes d'interaction qui peuvent se produire lors de la réalisation de la tâche de lecture, par exemple, l'intervention de l'enseignant ou des pairs, les situations de lecture individuelles ou en groupe, la lecture automne ou la lecture guidée, etc. Finalement, *le contexte physique* renvoie à toutes les conditions matérielles dans lesquelles se déroule la lecture, à savoir le bruit, la température ambiante, la qualité de la reproduction des textes, etc. (Giasson, 1996). Snow (2002) ajoute que ce contexte social ne se limite pas à la salle de classe ou de l'école, mais il peut aussi se transférer à un contexte plus large comme le milieu socioéconomique, le voisinage, les disparités ethniques, etc. Selon Tharp et Gallimore (1988), au point de vue socioculturel, l'acquisition des connaissances des élèves et la littératie sont influencées par quatre caractéristiques socioculturelles : (1) l'identité des lecteurs; (2) la façon dont l'activité de lecture est définie et effectuée; (3) le moment de l'activité et (4) les raisons ou les motivations qui ont suscités l'activité. Évidemment, ces caractéristiques varient en fonction des facteurs économiques et sociaux.

Dans le contexte du PIRLS 2011, la construction du sens est faite dans une communauté de lecteurs comprenant des individus de différents groupes sociaux et ethniques, ce qui peut engendrer une variété de manières d'interpréter des textes. En effet, les discussions sur ce qu'ils lisent avec ces différents groupes d'individus permettent aux élèves de construire du sens dans une variété de contextes (Mullis et al., 2009). Ainsi, les interactions sociales qui se passent dans cette communauté de lecture aident les élèves à mieux comprendre et apprécier les textes (Galda et Beach, 2001; Mullis et al., 2009), ce qui reflète bien la conception socioculturelle de la lecture (Baker, 1991; Jang, 2005; Street et al., 2001).

Lecteur

La composante « lecteur » constitue la composante centrale et la plus complexe dans la compréhension en lecture. Le lecteur résout les tâches en lecture avec des ***structures cognitives*** et des ***structures affectives*** qui lui sont propres. De plus, il mobilise ***des processus*** différents pour mieux comprendre le texte. Les ***structures cognitives*** incluent les connaissances sur la langue, dont celles phonologiques, syntaxiques, sémantiques et pragmatiques et les connaissances générales sur le monde. Ces connaissances générales sont celles antérieures que le lecteur utilise pour interpréter le texte afin de créer de nouvelles connaissances (Adams et Bruce, 1982).

En complément des structures cognitives**, les structures affectives** sont liées aux attitudes générales face à la lecture et aux intérêts développés par le lecteur. Cette attitude générale interviendra lorsque l'individu sera confronté à une tâche dont l'enjeu est la compréhension en lecture (Giasson, 1996). Quant aux intérêts spécifiques de l'individu, ils peuvent se développer non seulement dans le contexte de lecture, mais aussi hors de ce contexte, et doivent être considérés comme un facteur indispensable lors de la lecture d'un texte spécifique. Outre les attitudes générales et les intérêts, certains éléments susceptibles d'intervenir dans les structures affectives sont le concept de soi en général et le concept de soi comme lecteur, et la peur de l'échec (Giasson, 1996).

Lors de la compréhension en lecture, en parallèle avec les structures cognitives et affectives, le lecteur doit mobiliser **des processus** différents pour accomplir la tâche. Irwin (1986) a proposé une classification qui distingue cinq catégories de processus: (a) les **microprocessus** qui servent à comprendre l'information contenue dans une phrase;

(b) **les processus d'intégration** qui ont pour fonction de réaliser les liens entre les propositions ou les phrases; (c) **les macro processus** qui visent la compréhension globale du texte et les liens de cohérence du texte; (d) **les processus d'élaboration** permettant aux lecteurs d'effectuer les inférences imprévues par l'auteur et (e) **les processus métacognitifs** qui gèrent la compréhension et permettent au lecteur de s'ajuster au texte et à la situation (Irvin, 1986; Giasson, 1996). Plus précisément, les microprocessus servent à la reconnaissance des mots, la lecture par le groupe de mots ou la micro-sélection. Les processus d'intégration visent à l'identification des référents, l'utilisation des connecteurs et les inférences fondées sur les schémas. Les macro-processus, quant à eux, sont orientés vers l'identification vers des idées principales, le résumé et l'utilisation de la structure du texte. Les processus d'élaboration permettent les prédictions, l'imagerie mentale, les réponses affectives, l'établissement des liens avec les connaissances et le raisonnement. Finalement, les processus métacognitifs gèrent l'identification et la réparation de la perte de la compréhension (Irvin, 1986; Giasson, 1996). La figure 2 présente une synthèse du modèle contemporain en lecture avec les éléments qui sont jugés influencer la compréhension en lecture.

Ainsi, si nous pouvons faire référence à ce modèle contemporain en lecture, parmi les quatre processus de la compréhension proposés dans le PIRLS 2011, « se concentrer sur les informations énoncées de façon explicite » fait partie des microprocessus; « faire des inférences simples » est un des processus d'intégration; « interpréter et combiner des idées et des informations » est un processus d'élaboration tandis que le processus « examiner et évaluer le contenu, le langage et les éléments textuels » fait partie des macro-processus. Nous pouvons également regrouper ces quatre processus en deux catégories : la *microstructure* ou *compréhension locale* et la *macrostructure* ou *compréhension globale*. La compréhension locale concerne la lecture de chaque phrase ou chaque paragraphe tandis que celle globale porte sur le texte entier. Ainsi, le processus « se concentrer sur les informations énoncées de façon explicite » fait partie de la compréhension locale du texte. Ce processus demande que les lecteurs se concentrent et repèrent quelques morceaux d'informations, mais dans chaque cas, l'information se trouve dans une phrase ou une proposition.

Le processus « faire des inférences simples » est à la fois un processus de la compréhension locale et globale étant donné que les lecteurs ne se concentrent pas sur le niveau de compréhension des phrases ou des propositions, mais plutôt sur une compréhension locale dans certaines parties du texte ou encore sur une compréhension globale du texte. En outre, les inférences simples peuvent demander aux lecteurs de connecter le sens local et le sens global (Mullis et al., 2009). Par contre, « interpréter et combiner des idées et des informations » et « examiner et évaluer le contenu, le langage et les éléments textuels » sont des processus de la compréhension globale, car les élèves doivent porter un regard critique sur le contenu global et sur les éléments textuels du texte (Labrecque et al., 2012). Ils peuvent réfléchir également sur les dispositifs pour transmettre le sens de l'auteur et juger sa pertinence ainsi que se questionner sur l'objectif, la perspective ou l'habileté de l'auteur. Ainsi, le contenu ou le sens peuvent être examinés à partir d'une perspective très personnelle avec une vision critique et objective en se basant sur les connaissances du monde et les expériences du passé (Labrecque et al., 2012; Mullis et al., 2009). Toutefois, l'idée de distinguer la compréhension en lecture globale et la compréhension locale est introduite pour la première fois dans les travaux de Van Dijk et Kintsch (1983) et ensuite dans le modèle de construction et intégration en lecture de Kintsch (1988) dont nous parlerons plus en détail dans les lignes suivantes.

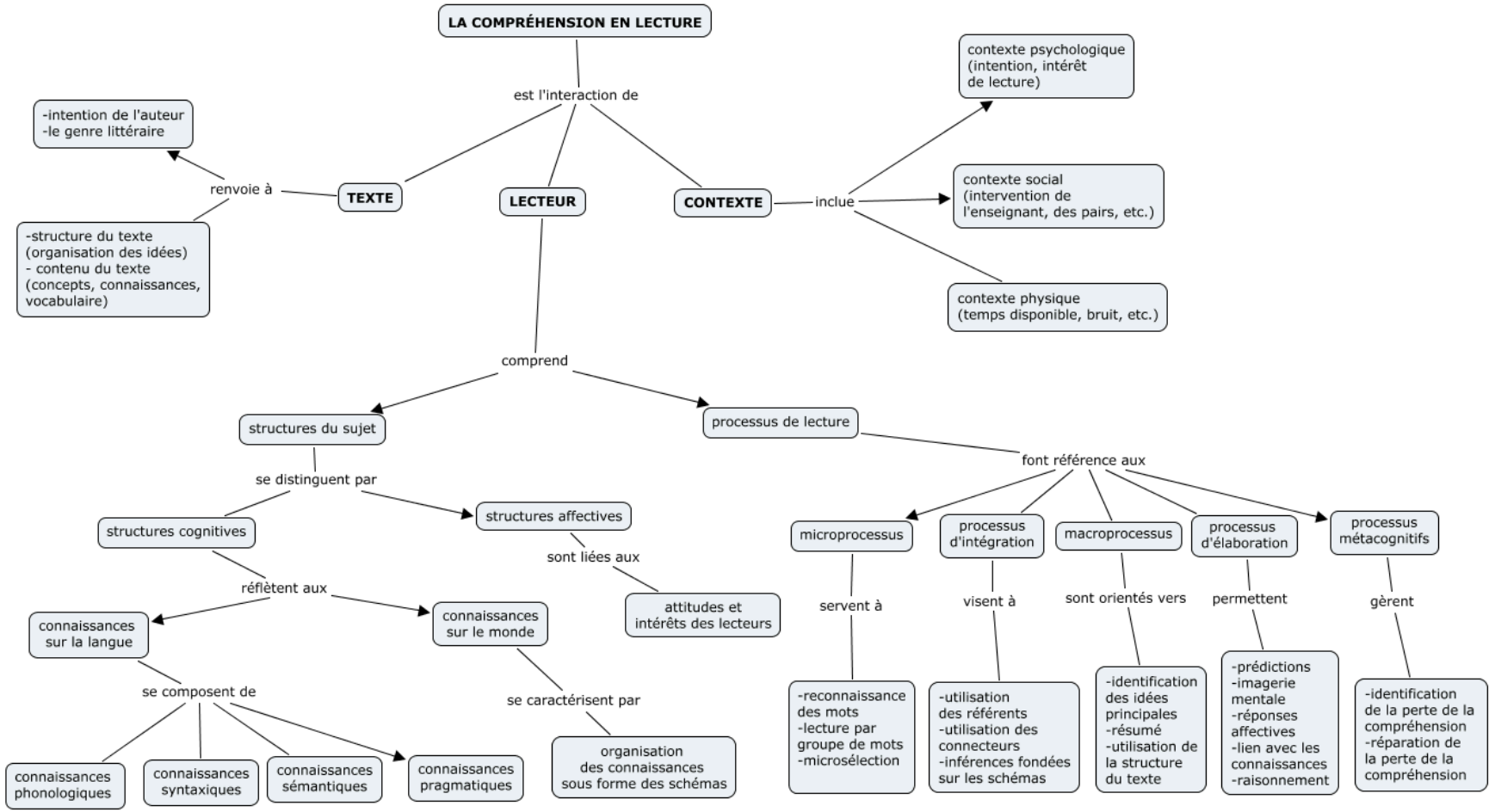


Figure 2. Modèle contemporain en lecture (synthétisé et adapté selon Irwin, 1986, 1991; Giasson, 1996)

2.1.3.3. Modèle construction-intégration en lecture

Le modèle de construction-intégration en lecture est un modèle de la compréhension du discours (Kintsch, 1988; Wharton et Kintsch, 1991; Kintsch et Welsch, 1990; Otero et Kintsch, 1992) qui traite de la représentation du discours dans la mémoire. Selon Van Dijk et Kintsch (1983), et Kintsch et al., (1990), cette représentation du discours dans la mémoire se distingue en trois niveaux. Dans un premier niveau, le texte est caractérisé par les mots ou les phrases utilisés d'une manière exacte, autrement dit, un niveau de surface de la présentation. Dans un deuxième niveau, ce ne sont plus les mots ou les phrases, mais plutôt le contenu sémantique du texte qui doit être présenté où les stratégies de la *compréhension locale (microstructure)* et celles de la *compréhension globale (la macrostructure)* vont intervenir. Le dernier niveau, ce qui constitue le modèle de la situation, assume que la compréhension du texte ne peut pas être produite uniquement à partir du texte, mais aussi à partir de la situation décrite par le texte grâce à la structure du texte et à différents domaines de connaissances préétablies (Van Dijk et Kinstch, 1983). Par conséquent, ce niveau ne concerne plus la macrostructure du texte, mais plutôt les schémas de connaissances (knowledge schemata) (Kintsch et al., 1990).

Le modèle de construction-intégration en lecture décrit comment les textes sont représentés dans la mémoire lors du processus de la compréhension et comment ils sont intégrés dans la base des connaissances des lecteurs (Kintsch, 1988; Wharton et Kintsch, 1991; Kintsch et al., 1990; Otero et Kintsch, 1992). Ce modèle propose deux étapes : **la construction des connaissances** dans laquelle la base du texte est construite à partir de l'entrée linguistique ainsi que des connaissances des lecteurs et **l'intégration des connaissances** où la base du texte est intégrée pour former un ensemble cohérent (Kintsch, 1988).

L'étape de la construction des connaissances est modélisée comme un système de production, dont les règles peuvent fonctionner à des niveaux différents : certains peuvent construire des propositions à partir des informations linguistiques fournies par le texte ou générer par la compréhension globale. D'autres peuvent se remémorer des connaissances à partir de la mémoire à long terme (Kintsch, 1988; Wharton et Kintsch, 1991; Kintsch et al., 1990). Le résultat de cette étape de construction est la production d'un réseau des propositions (propositional network). Ce réseau est donc influencé non seulement par

leurs connaissances mais aussi par leurs expériences antérieures, ce qui exerce ensuite des influences sur le réseau de propositions élaboré (Elaborated propositional network) (Wharton et Kintsch, 1990). Ce réseau de propositions élaboré est ensuite transféré dans un ensemble cohérent et interprétable à travers l'étape de l'intégration dont le résultat est une représentation finale du texte qui peut être interprétée et évaluée (Wharton et Kintsch, 1990). Il est à noter qu'il existe une complémentarité entre l'étape de la construction et celle de l'intégration, étant donné que l'étape de la construction n'a pas à être parfaite et sera ajustée dans la phase d'intégration. Ainsi, la phase d'intégration permet d'éliminer des réseaux de propositions incohérents (Kintsch, 1988). La figure 3 présente ces deux étapes du modèle de construction-intégration en lecture.

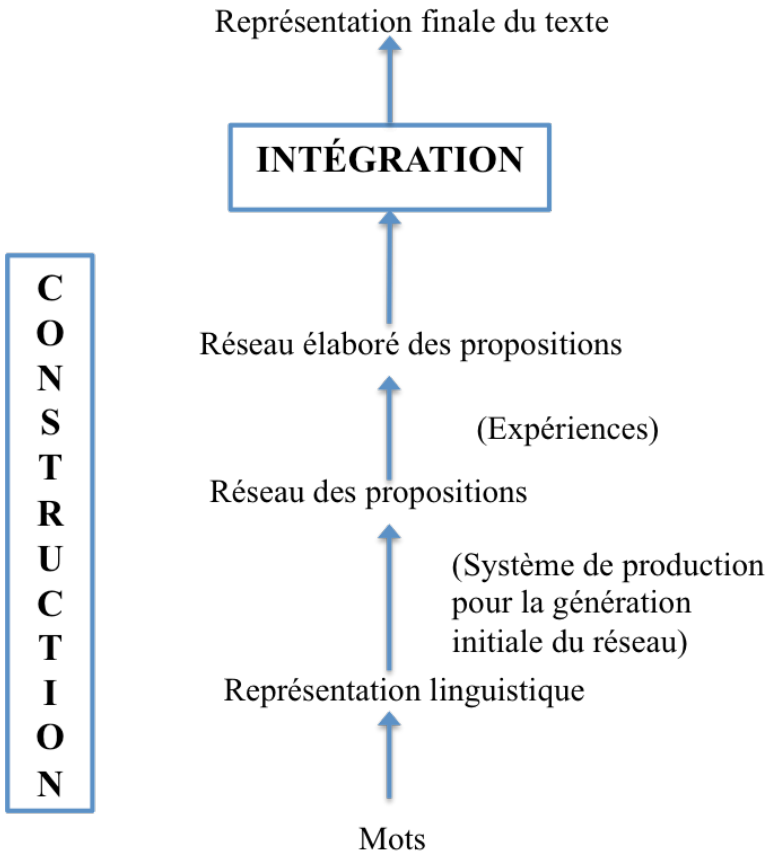


Figure 3. Modèle de construction-intégration en lecture (adapté de Wharton et Kintsch, 1990).

Le modèle de construction-intégration en lecture fournit une architecture générale pour le processus de la compréhension qui peut répondre aux différences entre individus,

car en lisant un même texte, ce ne sont pas les mêmes propositions qui seront construites par tous les individus (Otero et Kinstch, 1992). En effet, il existe des différences dans l'interprétation des textes parce que chacun va activer différents éléments de ses connaissances (Kintsch et al., 1990).

Une des hypothèses fondamentales du modèle de la construction-intégration en lecture (Kintsch, 1983) est que la compréhension d'un texte peut se passer à deux niveaux différents : une compréhension globale (la macrostructure) et une locale (la microstructure). Cette distinction semble cohérente avec les quatre processus de la compréhension en lecture proposés dans le cadre de référence du test PIRLS 2011 que nous avons analysés dans le modèle contemporain en lecture (Irwin, 1986; Giasson, 1996). D'ailleurs, les discussions sur les différents niveaux de compréhension des textes ont fréquemment mené à la fusion avec celles sur les habiletés ou sous-habiletés des lecteurs. Autrement dit, ces habiletés permettent de distinguer la compréhension de microstructure de celle de macrostructure (Alderson, 2005).

Ainsi, afin de mieux connaitre les différents niveaux de compréhension des textes et les habiletés qui leur sont liées, de nombreux chercheurs en lecture ont essayé d'identifier des habiletés et des sous-habiletés de la compréhension et plusieurs taxonomies de ces habiletés ont été développées afin de répondre à cette question (Alderson et Lukmani, 1989). La partie suivante est consacrée à décrire quelques taxonomies d'habiletés et de sous-habiletés de lecture selon différentes études et à faire émerger des convergences et des divergences entre elles ainsi que les liens explicites avec les quatre processus en lecture proposés dans le PIRLS 2011.

2.1.4. Taxonomies des habiletés en lecture

Une des recherches empiriques majeures sur l'identification des habiletés en lecture a été menée par Davis (1968, 1972) avec une grande banque d'items élaborée pour mesurer huit habiletés hypothétiques en lecture. À la différence d'autres recherches, chaque item est basé sur un champ de lecture séparé afin de créer l'indépendance locale entre les items (Spiro, 1980). Quatre recherches ont été ensuite menées par Davis (1968, 1972); Spearritt (1972) et Thorndike (1973) à partir de cette base de données et ont fait ressortir huit habiletés de lecture, à savoir : (a) rappeler le sens des mots; (b) faire des

inférences sur le sens des mots dans un contexte; (c) trouver **des réponses explicites** aux questions ou en paraphrase; (d) construire des idées ensemble dans le contenu; (e) faire des **inférences** à partir du contenu; (f) reconnaitre l'objectif, l'attitude, le ton et l'humeur de l'auteur; (g) identifier le style de l'auteur; (h) suivre la structure d'un passage.

En outre, Carver (1973) a décrit les quatre habiletés suivantes de la compréhension : (1) décoder des mots et déterminer leur sens dans une phrase particulière; (2) combiner les sens des mots individuels pour la compréhension complète des phrases; (3) comprendre des paragraphes et leurs idées implicites principales ainsi que la cause effet, les preuves et les hypothèses, les conclusions non déclarées et les idées associées à celles principales d'un paragraphe et (4) évaluer des idées, y compris des questions logiques, des preuves, l'authenticité et des arguments valables. Carver (1973) et Spiro (1980) suggèrent que les niveaux 1 et 2 correspondent à la lecture tandis que les niveaux 3 et 4 représentent le raisonnement. Ils ajoutent que la plupart des tests en lecture mettent l'accent sur ces deux niveaux et mesurent ainsi le raisonnement. En même temps, cette manière de distinguer la lecture du raisonnement de Carver (1973) semble similaire au repérage *des réponses explicites* et *implicites* (Spiro, 1980).

D'ailleurs, une vue très commune dans la littérature sur les recherches en lecture est *la conception simple en lecture (the simple view of reading)* qui suppose que la lecture se divise en deux composantes : le décodage et la compréhension (Gough et al., 1996). La compréhension est souvent décrite comme: (a) l'analyse des phrases; (b) la compréhension des phrases dans un discours; (c) la construction de la structure du discours et (d) l'intégration de cette compréhension dans nos connaissances *(*Alderson, 2005). Il est à noter que ces processus de compréhension ne sont pas utilisés uniquement pour la compréhension écrite, mais aussi pour la compréhension orale.

Une autre vision de la conception simple en lecture est celle proposée par Carver (1973, 1992) dans laquelle la lecture doit être analysée selon trois composantes séparables : la reconnaissance des mots; la vitesse ou la fluidité de la lecture et des habiletés de la résolution des problèmes en compréhension avec ces processus suivants: (1) la lecture typique faite dans les conditions où le lecteur n'a aucune difficulté à comprendre les phrases; (2) la mémorisation; (3) le survol (skimming) et (4) le repérage (scanning). Il a montré également que la fluidité en lecture change selon le

développement des lecteurs, autrement dit, la vitesse de lecture augmente avec le développement de cette compétence (Alderson, 2005).

En partageant le même point de vue de Carver (1973, 1992), Grabe (1991) a proposé six composantes suivantes dans le processus de lecture : (1) l'habileté de reconnaissance automatique; la reconnaissance globale d'un mot (2) les connaissances sur le vocabulaire et la structure; (3) les connaissances de la structure du discours formel; (4) les connaissances culturelles; (5) les habiletés de synthèse et d'évaluation et (6) les connaissances métacognitives et les habiletés d'autorégulation. Les stratégies métacognitives proposées incluent: (1) la reconnaissance des informations importantes dans le texte; (2) l'ajustement de la vitesse de lecture; (3) l'utilisation du contexte pour résoudre un problème de compréhension; (4) la formulation des questions sur les informations et (5) la reconnaissance des problèmes présentés dans le texte (Alderson, 2005; Grabe, 1991).

Avec les avancements des sciences cognitives, des recherches se sont intéressées au développement des habiletés de littératie critique. Par exemple, Abdullah (1994) a identifié les sept habiletés suivantes dans le processus de littératie critique : (1) évaluer des inférences déductives; (2) évaluer des inférences inductives; (3) évaluer la solidité de la généralisation; (4) évaluer des hypothèses cachées; (5) identifier les biais dans les déclarations; (6) reconnaitre le motif des auteurs et (7) évaluer la force des arguments. Il est important de tenir compte du fait que ces habiletés correspondent très bien à l'aspect de lecture réflexive et critique que le test PIRLS envisage.

Ces listes ou ces taxonomies d'habiletés en lecture permettent de justifier théoriquement les moyens de diviser les tâches ou les items dans un test et d'isoler les habiletés à évaluer. Ils suggèrent également la possibilité de diagnostiquer des problèmes en lecture chez les lecteurs afin d'émettre des stratégies de remédiation et constituent donc un cadre de référence fort utile pour la construction des tests (Alderson, 2005). Toutefois, ces listes d'habiletés devraient être utilisées avec soin, étant donné qu'il existe encore une certaine confusion dans le fait d'identifier et de séparer des habiletés et que des justifications empiriques claires soient nécessaires pour appuyer ce débat (Spiro, 1980). De plus, ces habiletés sont souvent mal définies ou indéfinies, ce qui rend difficile le consensus par les experts sur le choix des habiletés utilisées selon les items (Alderson,

1990). Finalement, l'équipe de Alderson (Alderson, 1990; Alderson, 2005; Alderson et Lukmani, 1989) suggère que l'analyse des tests de performance ne révèle pas la séparabilité des habiletés ni la hiérarchie de la difficulté ou la discrimination. Malgré ces limites, cette approche d'identification des habiletés en lecture ne peut pas être exclue lors de l'analyse du niveau de compétence exercé (Alderson, 2005).

En effet, en analysant quelques taxonomies d'habiletés proposées par les auteurs, nous avons constaté qu'il existe des correspondances entre ces habiletés et les processus de compréhension en lecture dans le PIRLS 2011. Par exemple, l'habileté « Faire des inférences sur le sens des mots dans un contexte » et « trouver des réponses explicites aux questions ou en paraphrase » proposées dans la taxonomie de Davis (1968) correspondent très bien au processus « Faire des inférences simples » et « Se concentrer sur les informations énoncées de façon explicite et les extraire du texte » dans le test PIRLS 2011. Cette même habileté « Faire des inférences » a été identifiée dans la taxonomie de Carver (1973). Ces deux habiletés font partie de la compréhension locale du texte. Quant à la compréhension globale, l'habileté « Comprendre des paragraphes et leurs idées implicites principales » de Carver (1973) correspond au processus « Interpréter et combiner des idées et des informations », tandis que « Évaluer des idées, y compris des questions logiques, des preuves, l'authenticité et des arguments valables » (Carver, 1973); « Habiletés de synthèse et d'évaluation » (Grabe, 1991); « Reconnaître le motif des auteurs » et « Évaluer la force des arguments » (Abdullah, 1994) ont un lien avec le processus de « Examiner et évaluer le contenu, le langage et les éléments textuels » du PIRLS 2011. En parallèle avec les quatre processus de la compréhension en lecture, les concepteurs ont identifié l'ensemble des tâches demandées pour chaque processus dont nous avons fait une synthèse dans le tableau 1. Ces tâches identifiées permettent de décrire plus en détail chaque processus de la compréhension en lecture et de fournir une image plus fine sur le potentiel diagnostic de chaque processus.

Tableau 1. Description des processus de compréhension du test PIRLS 2011 (adapté de Labrecque et al., 2012; Mullis et al., 2009)

Processus de compréhension	**Description**	**Tâches demandées**
Se concentrer sur les informations énoncées de façon explicite et les extraire du texte	-Comprendre et repérer les informations énoncées de façon explicite et faire le lien avec la question posée -Reconnaitre la pertinence de l'information et de l'idée	-Mettre en évidence des éléments d'information pertinents par rapport au but spécifique -Rechercher des idées précises -Rechercher des définitions pour certains mots ou certaines expressions -Définir le contexte d'une histoire (époque, lieu, etc.) -Déterminer le thème ou l'idée principale (lorsqu'elle est clairement énoncée).
Faire des inférences simples	-Combler les « lacunes » relatives au sens en déduisant des informations à partir du texte -Exiger peu d'efforts des lecteurs ou se fait de façon spontanée	-Déduire qu'un évènement en a entrainé un autre -Tirer des conclusions quant à l'idée principale dans une série d'arguments -Déterminer le référent d'un pronom -Mettre en évidence les généralisations formulées dans le texte -Décrire la relation entre deux personnages
Interpréter et combiner des idées et des informations	-Acquérir une compréhension plus approfondie du texte en combinant les connaissances antérieures et les informations présentées dans le texte	-Dégager le message ou thème général d'un texte -Envisager une séquence d'actes différente pour les personnages -Mettre en évidence les points communs et les différences dans les informations contenues dans le texte -Dégager l'atmosphère ou le ton d'une histoire -Interpréter les applications possibles des informations dans le monde réel
Examiner et évaluer le contenu, le langage et les éléments textuels	-Prendre du recul par rapport au texte, afin de porter un regard critique sur son contenu, le langage utilisé ou les éléments textuels -Comparer la représentation d'un mot selon l'auteur à sa propre compréhension ou à des informations en provenance d'autres sources -Réfléchir à la clarté de l'expression du sens, en faisant appel à ses propres connaissances sur le genre en question, la structure ou les conventions linguistiques	-Évaluer le degré de probabilité que les évènements dépeints se produisent dans la vie réelle - Décrire la méthode utilisée par l'auteur pour inventer un dénouement inattendu - Déterminer dans quelle mesure les informations présentées dans le texte sont complètes ou claires -Déterminer le point de vue d'un auteur sur le thème central

2.1.5. Synthèse

Le cadre de référence du test PIRLS 2011 ne s'appuie pas sur un modèle théorique unique en lecture. Par contre, il adopte une visée interactive et constructive en lecture. Dans ce cas, la compréhension en lecture se réalise par l'interaction entre trois composantes : le texte, le lecteur et le contexte dans lequel le lecteur constitue l'acteur principal. Le lecteur doit mobiliser un ensemble de connaissances linguistiques et culturelles, des stratégies cognitives et métacognitives pour construire du sens. Les fondements théoriques de cette visée interactive et constructive en lecture peuvent donc se retrouver dans les modèles interactifs en lecture (Rumelhart, 1977), le modèle contemporain en lecture (Irwin, 1986; Giasson, 1991), la théorie du schéma (Alderson et Pearson, 1984, 1988) et le modèle de construction et intégration en lecture (Kintsch, 1988). Cette visée interactive et constructive en lecture se traduit en deux objectifs et quatre processus de la compréhension en lecture qui font l'objet des évaluations du test PIRLS 2011.

Ainsi, dans les lignes précédentes, nous avons décrit le cadre de référence du test PIRLS 2011 avec la définition de la littératie et de la compréhension de l'écrit, les deux objectifs et quatre processus de la compréhension en lecture proposés dans le PIRLS. Nous avons établi les liens explicites entre ce cadre de référence et les modèles et théorie en lecture. Étant donné que le test PIRLS s'intéresse à quatre processus de la compréhension en lecture, nous avons examiné quelques exemples des taxonomies d'habiletés en lecture afin de faire sortir des ressemblances avec ces quatre processus de la compréhension. Ces quatre processus sont décrits à l'intérieur des tâches demandées qui leur sont liées. Les grandes idées de cette partie sur le cadre de référence du test PIRLS 2011 peuvent être schématisées dans la figure 4. Alors, comment cette visée interactive et constructive en lecture, avec ses deux objectifs et quatre processus de compréhension ainsi que ces tâches identifiées, pourrait-elle être utilisée afin de dresser les profils plus fins et détaillés sur les forces et les faiblesses cognitives des élèves à travers des rapports diagnostiques? Les éléments de réponse à cette question vont être exploités dans les parties suivantes de notre cadre conceptuel.

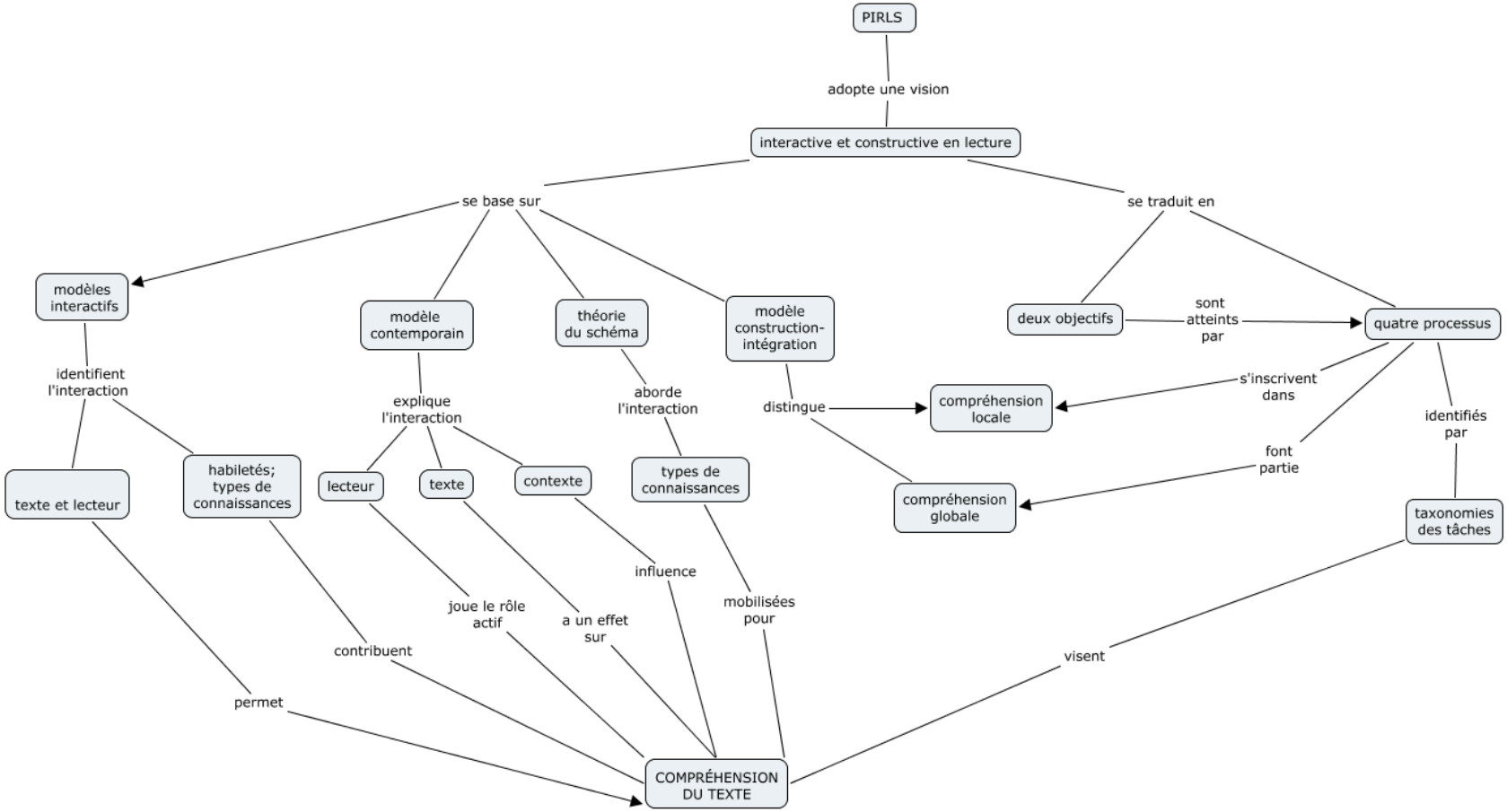


Figure 4. Liens entre les modèles théoriques en lecture et le test PIRLS 2011

2.2. Approche diagnostique cognitive

L'approche diagnostique cognitive (ADC) a été développée pendant les années 1980 avec la combinaison de la psychologie cognitive et de la psychométrie (Leighton et Gierl, 2007; Lee et Sawaki, 2009). Elle vise à poser un diagnostic sur des variables latentes discrètes représentant des processus cognitifs (Nichols, 1994; Loye, 2008). Sous la lumière de cette approche, une variété de modèles psychométriques a été développée afin de fournir des informations très fines sur les points forts et les faiblesses cognitives des apprenants (Alderson, 2005; Buck et Tatsuoka, 1998; Buck, Tatsuoka et Kostin, 1997; Gao et Rogers, 2007; Jang, 2005, 2009; Lee et Sawaki, 2009; von Davier, 2008). Dans le domaine des langues, les recherches sur l'application de ces modèles à des tests standardisés à grande échelle tels que le TOEFL, le TOEFL iBT, le TOEIC et le METLAB ont été réalisées. Les rétroactions détaillées sur le degré de maitrise des connaissances et des habiletés obtenues à la suite de ces modélisations des données permettent aux apprenants et aux enseignants de prendre des mesures de remédiation appropriées (Lee et Sawaki, 2009). Les lignes suivantes servent à décrire plus en détail cette approche diagnostique cognitive.

2.2.1. Définition de l'approche diagnostique cognitive

L'approche diagnostique cognitive a été développée en se basant sur deux composantes principales: (a) l'analyse du contenu des items afin d'identifier les attributs cognitifs concernés; (b) les modèles psychométriques représentant les relations entre ces items et ces attributs (Yang et Embretson, 2007; Lee et Sawaki, 2009). *Les attributs* renvoient à un ensemble d'habiletés, de connaissances, de capacités ou de processus mentaux et de stratégies qu'un individu mobilise pour répondre correctement à un item (Lee et Sawaki, 2009; Buck et Tatsuoka, 1998; Leighton et Gierl, 2007). L'analyse du contenu des items est souvent réalisée par les experts du domaine et une matrice Q est développée pour désigner les relations entre les items et les attributs. Cette matrice prend souvent forme d'un tableau de spécification pour déterminer si un attribut est nécessaire pour répondre correctement à un item. Les modèles psychométriques sont utilisés pour

estimer des paramètres informant sur l'état de maitrise des connaissances et des habiletés. Ces modèles sont appelés les modèles diagnostiques cognitifs (Gierl et al., 2000; Rupp et al., 2006), les modèles psychométriques cognitifs (Gao et Rogers, 2007), les modèles de classification diagnostique (Rupp et Temblin, 2011) ou encore les modèles psychométriques diagnostiques cognitifs (Fu et Li, 2007). Dans le cadre de ce projet, nous allons utiliser « modèles de classification diagnostique » (MCD) pour désigner ces modèles psychométriques, car il s'agit de l'appellation la plus utilisée dans les écrits en français sur le sujet.

En pratique, le diagnostic dans cette approche diagnostique cognitive est réalisé de deux façons. La première consiste à analyser les données des tests à grande échelle existants, qui n'ont pas été conçus spécifiquement pour une visée diagnostique, en recourant à un modèle cognitif sous-jacent dans l'espoir d'extraire des informations plus riches sur la maitrise ou non-maitrise des habiletés des apprenants. La deuxième consiste à concevoir un test à finalité diagnostique puis à modéliser les données issues de ce test pour faire sortir des informations diagnostiques (Dibello, Roussos et Stout, 2007). Notre recherche s'inscrit dans la première façon de faire. Dans le domaine de l'évaluation des compétences langagières, elle comporte les quatre étapes suivantes :

-**La définition des attributs**: cette étape vise à identifier un ensemble d'attributs liés au test et les relations entre les attributs et les items du test. L'identification des attributs peut être réalisée en se basant sur plusieurs sources, à savoir l'analyse du contenu, la spécification des tests, les théories du contenu du domaine, l'analyse du contenu des items et des protocoles du processus de passage des tests des apprenants. Il est important que la définition des attributs se fonde sur les résultats des recherches empiriques portant sur le processus mental ou les stratégies cognitives sous-jacentes aux habiletés et connaissances que les apprenants utilisent pour résoudre des problèmes (Lee et Sawaki, 2009; Leighton et Gierl, 2007). Lors de l'identification des attributs, l'attention devrait être prêtée au nombre adéquat d'habiletés, mais aussi à l'interaction entre ces habiletés. Idéalement, il ne faut pas que le nombre d'attributs soit trop grand ni trop petit pour qu'ils puissent être statistiquement supportables, ne pas perdre la capacité diagnostique et faciliter l'interprétabilité (Dibello; Roussos et Stout, 2007; Jang, 2005). Il est également important de garder en tête la nature interactive entre les habiletés

(conjonctive ou disjonctive), ce qui détermine le choix d'un modèle MCD approprié (compensatoire ou non compensatoire) (Dibello, Roussos et Stout, 2007).

-**La construction de la matrice Q** : après avoir identifié l'ensemble des attributs requis par le test, une matrice Q est construite pour représenter les relations entre les items et les attributs. La construction de la matrice Q peut se baser sur l'analyse du contenu du test ou sur la combinaison entre l'analyse du contenu et l'analyse des données empiriques issues par exemple d'une passation auprès d'un petit groupe de candidats (Tjoe et de la Torre, 2013a ; Tjoe et de la Torre, 2013b ; Loye et Lambert-Chan, 2015). Dans certains cas, les experts peuvent recourir à l'analyse des verbalisations à haute voix des candidats afin d'obtenir la matrice Q finale (Jang, 2005 ; Li et Suen, 2012, Li, 2011).

La notion de la matrice Q a été introduite pour la première fois par Tatsuoka en 1983 (Buck et Tasouka, 1983; Buck et Tatsuoka, 1997; Buck et Tatsuoka, 1998; Jang, 2005). Cette matrice prend souvent la forme d'indications binaires pour indiquer si un certain attribut est nécessaire ou non afin de répondre correctement à un item (Buck, 1990; Im et Corter, 2011). Le tableau 2 propose l'exemple d'une matrice Q (4x3) du test METLAB (Li, 2011) représentant les liens qui existent entre quatre items (I1, I2, I3; I4) et trois attributs : faire des inférences (A1); extraire des informations simples (A2) et connexion et synthèse (A3).

Tableau 2. Exemple d'une matrice Q du test METLAB (adapté de Li, 2011).

Items	Attribut 1 (Faire des inférences)	Attribut 2 (Extraire des informations simples)	Attribut 3 (Connexion et synthèse)
1	1	1	0
2	1	0	1
3	0	1	1
4	1	1	1

Dans ce tableau, nous supposons qu'il y a quatre items, et que les trois attributs A1 (Faire des inférences); A2 (Extraire des informations simples) et A3 (Connexion et synthèse) sont nécessaires pour y répondre correctement. Plus précisément, A1 est

essentiel pour résoudre l'item 1, 2 et 4 tandis que A2 est indispensable pour l'item 1, 3 et 4. A3 est demandé pour répondre correctement aux items 2, 3, et 4. Avec un test d'I (items) et K (attributs), nous allons obtenir une matrice Q de dimension I x K permettant de déterminer quels attributs le sujet doit maitriser afin de répondre correctement à chaque item, soit :

$$q_{ik} = \begin{cases} 1 & \text{si l'attribut k est nécessaire pour l'item i} \\ 0 & \text{sinon} \end{cases}$$

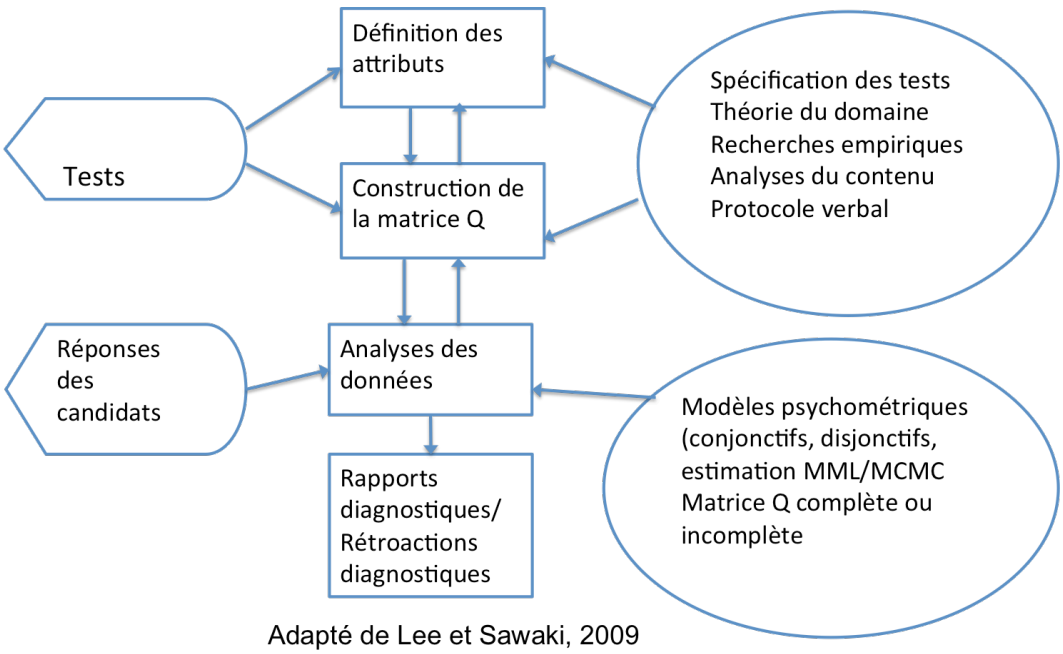
-**La modélisation des données:** La matrice Q élaborée est ensuite intégrée dans la modélisation des données issues des réponses aux items des candidats avec un MCD choisi. La procédure de modélisation des données peut être différente d'un MCD à un autre tout dépendant de la méthode d'estimation utilisée. Certains modèles utilisent les algorithmes de classification et se concentrent donc sur la classification des candidats sur les classes latentes à partir de leurs schémas d'attributs maitrisés. Les résultats de la modélisation fournissent les paramètres d'items et les paramètres des sujets. Les paramètres d'items permettent de porter un jugement sur le potentiel diagnostique du test et la qualité de la matrice Q, tandis que les paramètres de sujets fournissent les informations diagnostiques sur la maitrise ou non-maitrise des attributs des sujets (Loye, 2010).

Toutefois, peu importe la méthode d'estimation utilisée, il est important de vérifier la convergence des estimations des paramètres d'items et des sujets et d'évaluer l'ajustement du modèle aux données avec les statistiques associées. Si la convergence des paramètres et l'ajustement du MCD ne sont pas atteints, il faut réviser la procédure de la construction du modèle, l'identification des attributs et l'élaboration de la matrice Q. Une fois que les tests sont calibrés en utilisant la matrice Q spécifiée et le MCD approprié, les sujets peuvent être classifiés dans les profils de maitrise des attributs. Chaque candidat correspond à un vecteur α représentant l'état de sa maitrise de chaque attribut. À titre d'exemple, l'état de la maitrise de K attributs du sujet j est : $\alpha_j = (\alpha_{1j}, \alpha_{2j}, \ldots, \alpha_{Kj})$. Cet état de maitrise des attributs peut être modélisé comme une variable dichotomique selon :

$$q_{\alpha jk} = \begin{cases} 1 & \text{si le sujet j maitrise l'attribut k} \\ 0 & \text{sinon} \end{cases}$$

Les profils de maitrise des attributs des sujets peuvent être exprimés sous formes binaire (0=non-maitrise, 1=maitrise) ou multi-catégorielle (0=faible; 1=moyen; 3=élevé) (Lee et Sawaki, 2009).

-Les rétroactions diagnostiques : l'évaluation diagnostique des compétences langagières vise à fournir aux apprenants et aux enseignants des informations sur leurs points forts et leurs faiblesses. Ces informations sont obtenues à travers l'interprétation des résultats des modélisations des données des sujets en fonction de la matrice Q et portent sur les attributs qui sont liés à la compétence langagière. Ces résultats sont complexes et difficiles à interpréter, ils nécessitent donc d'être expliqués aux apprenants et aux enseignants (Yang et Embretson, 2007 ; Lee et Sawaki, 2009). Ainsi l'interprétation et la communication de ces informations diagnostiques cognitives sont des enjeux principaux de la présente étude. La figure 5 illustre les quatre étapes de cette approche diagnostique cognitive.



Adapté de Lee et Sawaki, 2009

Figure 5. Étapes de l'approche diagnostique cognitive

2.2.2. Études empiriques en lecture qui font appel au MCD

Dans cette partie, nous présenterons une synthèse des recherches en lecture réalisées sous la lumière de l'ADC. L'accent est mis sur les tests utilisés, les habiletés

évaluées, les MCD choisis pour les analyses des données ainsi que les rétroactions diagnostiques fournies aux apprenants et à leurs enseignants lorsque disponibles.

2.2.2.1. Définition des MCD

Malgré les appellations variées de ces modèles et le manque de consensus dans les définitions des MCD (Li, 2011), celle proposée par Rupp et Templin (2008, p.226) résume presque toutes les caractéristiques de ces modèles. Selon eux, les MCD se définissent comme suit :

> *Diagnostic classification models are probabilistic, confirmatory multidimensional latent-variable models with a simple or complex loading structure. They are suitable for modelling observable categorical response variables and contain unobservable (latent) categorical predictor variables. The predictor variables are combined in a compensatory and non-compensatory ways to generate latent classes. DCM enable multiple criterion-referenced interpretations and associated feedbacks for diagnostic purposes, which is typically provided at a relatively fine-grain size. This feedback can be, but does not have to be, based on a theory of response processing grounded in applied cognitive psychology. Some DCM are further able to handle complex samplings designs for items and respondents, as well as heterogeneity due to strategy use* (p.226).

Les MCD reposent sur le postulat que la performance des étudiants au test dépend de la maitrise ou la non-maitrise d'un ensemble d'attributs impossibles à observer directement (Tatsuoka, 1983, 1998; Rupp, 2007; Gierl, Leighton et Hunka, 2000; Milewski et Baron, 2002). Ces attributs sont reliés les uns aux autres par un ordre hiérarchique (Loye, 2008). C'est pour cette raison qu'ils sont utilisés pour estimer les probabilités de répondre correctement à un item à partir de la maitrise ou de la non-maitrise des attributs des étudiants.

Aujourd'hui, avec l'avancement de la psychologie cognitive et de l'informatique, ces modèles sont en pleine évolution (Rupp, 2007; Loye, 2008). Ils se distinguent l'un de l'autre par la structure des habiletés évaluées, la nature des items et l'interaction des attributs ainsi que les méthodes d'estimation des paramètres (Lee et Sawaki, 2009), ce qui différencie les modèles *compensatoires/disjonctifs* de ceux *non-*

compensatoires/conjonctifs (Marris, 1999). En effet, un modèle *non compensatoire* suppose que le sujet doit maitriser tous les attributs exigés pour un item afin de produire une réponse exacte (Rupp et Templin, 2008; Rupp, Templin et Henson, 2010). Le manque d'un attribut ne peut pas être compensé par un autre, ce qui est possible dans les modèles *compensatoires*. Certains modèles qui ont été utilisés dans les recherches en langue vont être décrits plus en détail dans les lignes qui suivent.

2.2.2.2. Le modèle de Rule-Space

Les premières recherches effectuées en lecture avec l'ADC sont celles de Buck, Tatsuoka et Kostin (1997) avec le test TOEIC; Buck et al., (1998) avec le SAT verbal; Kasai (1997) et Scott (1998) avec le test TOEFL selon le modèle de Rule-Space. Dans la recherche de Buck, Tatsuoka et Kostin (1997), l'analyse des données est faite avec 5000 étudiants japonais à partir de la section lecture du test TOEIC. Des 27 attributs initialement identifiés, 24 ont finalement été retenus, dont 16 attributs et 8 interactions. Ces attributs se basent principalement sur les taxonomies de sous-habiletés en lecture proposée par Grabe (1991) et sur les études empiriques menées par Feedle et Kostin (1993) qui se concentrent sur sept catégories différentes : (a) le croisement entre le texte et le choix correct; (b) la longueur des phrases; (c) la longueur des paragraphes; (d) l'organisation rhétorique; (e) l'utilisation de la négation; (f) l'utilisation des références et (g) la longueur des passages.

Avec ces attributs, les auteurs peuvent classifier 91% des candidats dans leurs profils de maitrise des habiletés et fournir les scores sur chaque habileté en termes de probabilité de maitrise. Ces scores sont ensuite analysés avec une régression multiple, dont les résultats suggèrent que les attributs peuvent expliquer 97% de la variation de la performance de 91% des candidats au test. Ainsi, le modèle de Rule-Space peut expliquer les performances des élèves sur les tâches complexes comme la lecture et leur fournir des informations diagnostiques (Buck, Tatsuoka et Kostin, 1997). Toutefois, les limites de cette étude résident dans la subjectivité des critères de sélection des attributs. En outre, l'analyse des données est réalisée avec une seule forme du test, ce qui pourrait être fait avec d'autres formats afin de pouvoir comparer les résultats obtenus (Buck, Tatsuoka et Kostin, 1997).

Le modèle Rule-Space a été appliqué au test TOEFL dans les recherches de Kasai (1997) et Scott (1998). Plus spécifiquement, la recherche de Kasai (1997) a mis en évidence un total de 27 attributs, dont 16 attributs principaux et 11 interactions. Ces attributs sont reliés aux : (a) connaissances du vocabulaire; (b) connaissances de l'organisation rhétorique; (c) connaissances de la grammaire; (d) traitement de l'information et au repérage de certaines parties du passage. Les résultats indiquent que ces attributs principaux et interactions peuvent classer plus de 80% des candidats dans les profils de maitrise des habiletés du test et expliquer 96% de leur performance au test (Kasai, 1997).

Avec ce même test, la recherche de Scott (1998) a obtenu les résultats plus ou moins similaires à celle de Kasai (1997) même si les attributs identifiés ne sont pas tout à fait identiques. Ces attributs sont élaborés en se basant sur cinq sources différentes: (a) la recherche de Alderson et ses collaborateurs (1991) qui met l'accent sur l'importance des types d'items; (b) celle de Buck et al., (1996) qui suggère un modèle pour les tests à choix multiples; (c) la théorie du schéma de Carrell (1984) que nous avons présentée dans les modèles cognitifs en lecture; (d) l'étude de Koda (1994) qui montre l'importance des connaissances du vocabulaire et de la syntaxe dans la lecture en langue seconde et (e) l'étude de Kasai (1996) qui suggère que le format du test exerce une influence sur la performance des candidats. Plus précisément, parmi 24 attributs identifiés pour cette recherche, seulement 13 attributs sont semblables à ceux de Kasai (1997) tandis que les attributs suivants sont différents: le repérage des informations nécessaires; les connaissances de l'organisation rhétorique; l'utilisation de l'option pour fournir la réponse correcte. Par contre, l'auteur a ajouté deux nouveaux attributs : l'identification des idées principales; le sens du vocabulaire et la référence. L'analyse des données du TOEFL avec le modèle de Rule-Space montre que le modèle peut classer 84% des candidats dans les niveaux de maitrise appropriés. Ces attributs peuvent aussi prédire 94% de leur performance (Scott, 1998). Ces résultats confirment donc le potentiel de diagnostic du test TOEFL relativement aux habiletés cognitives des candidats.

Le modèle de Rule-Space (RSM) (Tatsuoka, 1983, 1998, 2005, 2009; Im et Corter, 2011) présume qu'afin de répondre correctement à un item, le sujet devrait maîtriser tous les attributs nécessaires pour cet item particulier. En d'autres termes, le manque ou le

déficit d'un attribut ne peut pas être compensé par la maitrise d'autres attributs (Jang, 2005; Im et Corter, 2011). Il se base sur l'approche de la reconnaissance des schémas (pattern recognition approach) en mesurant la distance entre *les états de connaissances observés* des sujets et un ensemble de *schémas de réponses idéaux*. La classification des distances est ensuite estimée par un modèle de la théorie de réponses aux items (TRI) à deux paramètres, avec θ qui représente le niveau d'habileté estimé du sujet et ζ, un indice de type person-fit qui fait référence au degré selon lequel ils répondent incorrectement aux items plus faciles et correctement à ceux plus difficiles (Tatsuoka, 1985; 1990; Buck et Tatsuoka, 1998; Im et Corter, 2011).

Afin de déterminer le classement de chaque sujet, les couples (θ, ζ) sont représentés sur un plan cartésien à deux dimensions. Et les distances de Mahalanobis au carré entre ce couple (θ, ζ) et le centre de gravité correspondant aux schémas de réponses idéaux est calculé. Si la distance entre la position de l'état de connaissance observée d'un sujet est plus proche qu'une distance de coupure (cut off), ce centre de gravité est considéré comme l'état de connaissance du sujet (Im et Corter, 2011). La règle de décision de Bayes est ensuite appliquée pour déterminer l'erreur minimale de classer les sujets dans des états de connaissances possibles (Buck et Tatsuoka, 1998). Enfin, la probabilité de maitriser des attributs de chaque sujet est calculée par un modèle probabiliste en utilisant des attributs binaires dans les états de connaissances des sujets et les probabilités postérieures correspondantes.

Le modèle de Rule-Space comporte l'avantage de ne pas avoir beaucoup de paramètres à estimer, ce qui facilite l'interprétation. Celle-ci est réalisée par la visualisation graphique de la distance entre les états des connaissances observés des sujets et ceux idéaux (Loye, 2008). Toutefois, lorsque le nombre des schémas de réponse idéaux est grand, ces points idéaux ne peuvent pas être séparés clairement l'un de l'autre et l'erreur commise dans le classement peut être élevée (Buck et Tatsouka, 1998; Buck, Tatsuoka et Gostin, 1997; Kasai, 1996). En outre, une des limites dues à la nature du modèle de Rule-Space lui-même est qu'il peut fournir les scores sur les attributs seulement aux candidats qui sont classés avec succès dans leurs états de connaissances. Dans l'étude de Buck, Tatsuoka et Kostin (1997) par exemple, le modèle ne peut pas fournir les scores des attributs à 9% des candidats. Ainsi, si les chercheurs veulent utiliser

ce modèle pour des analyses diagnostiques, ils doivent trouver un moyen pour fournir les scores à ces candidats. Ces deux limites justifient les raisons pour lesquelles nous n'avons pas retenu ce modèle pour notre recherche.

2.2.2.3. Le modèle de Fusion

Dans les dernières années, le modèle de Fusion a été largement utilisé pour des analyses diagnostiques cognitives du test TOEFL iBT, dont un exemple est la recherche de Jang (2005). Avec les protocoles verbaux des étudiants, les experts et l'analyse du contenu des items, l'auteure a identifié neuf habiletés sous-jacentes en lecture pour réussir au test. L'analyse du test et des items se concentrent sur certaines variables qui peuvent prédire chez les candidats proposés des difficultés en compréhension en lecture dans les recherches précédentes telles que les niveaux de fréquence des mots (Drum et al., 1981; Freedle et Kostin, 1993); les longueurs de phrases, des paragraphes et des passages (Drum et al., 1981; Freedle et Kostin, 1993); les négations des phrases utilisées (Carpenter et Just, 1975); les structures de l'organisation rhétorique (Freedle et Kostin, 1993) et le nombre d'expressions référentielles (Abrahamsen et Shelton, 1989). Par contre, l'analyse des protocoles verbaux porte sur les habiletés et les stratégies permettant de répondre correctement aux items : la lecture expéditive; les connaissances lexicales et syntaxiques; la compréhension de base; l'inférence et la synthèse.

Jang (2009) analyse les données de 2703 candidats avec le modèle de Fusion afin d'estimer leur probabilité de maitrise des attributs. En se basant sur les protocoles verbaux des apprenants, les experts ont identifié neuf habiletés en lecture, à savoir : (1) Déduire le sens d'un mot ou d'une phrase du texte en utilisant les éléments fournis dans le texte; (2) Déterminer le sens d'un mot en utilisant les connaissances antérieures; (3) Comprendre les relations entre les parties du texte grâce aux connecteurs logiques; (4) Repérer des informations explicites; (5) Comprendre des informations implicites; (6) Faire des inférences; (7) Négation; (8) Résumer des idées principales; (9) Reconnaitre des idées ou des arguments contradictoires. Parmi les neuf habiletés identifiées, les habiletés 8, 1, 2 sont celles les plus maitrisées par les apprenants tandis que les habiletés 7 et 9 sont les moins maitrisées. Ce qui semble intéressant dans cette recherche est que la performance de certains étudiants a été évaluée avec le test avant et après avoir pris des

cours préparatoires au test. Les résultats montrent que ces étudiants ont amélioré de 12% leur probabilité de maitrise des habiletés après le cours et qu'environ 85% des étudiants peuvent améliorer leur performance sur les habiletés. Toutefois, une limite de cette étude est que le nombre d'attributs a été beaucoup diminué dans la Q matrice finale afin d'être théoriquement et statistiquement supportable (de 32 attributs à 16 et finalement à 9 attributs). Idéalement, le nombre d'attributs ne devrait pas être grandement réduit pour ne pas perdre l'utilité diagnostique et l'interprétabilité (Jang, 2005; Roussos, DiBello et Stout, 2007).

Ce modèle de Fusion a été également utilisé pour analyser des données du test METLAB dans l'étude de Li (2012) et Li et Suen (2011). Les attributs identifiés sont basés principalement sur les recherches de Gao (2006) et Jang (2005) pour le test TOEFL. Avec les protocoles verbaux et quatre experts, les auteurs ont initialement identifié six attributs pour les réduire finalement à quatre vu le nombre insuffisant d'items pour les attributs, à savoir : (a) le vocabulaire; (b) la syntaxe; (c) l'extraction des informations explicites; (d) la compréhension des informations implicites. Ainsi, si nous faisons référence au modèle de construction-intégration en lecture, à part des connaissances linguistiques telles que le vocabulaire et la syntaxe, l'attribut « extraire des informations explicites » représente la lecture locale tandis que l'attribut « comprendre des information implicites » fait partie de la compréhension globale. Les résultats suggèrent que le paramètre $\boldsymbol{\pi}$ représentant la probabilité qu'un candidat qui maitrise tous les attributs exigés pour l'item i va utiliser correctement ces attributs pour résoudre cet item est très grand ($\boldsymbol{\pi}$ =0,891). Les attributs sont globalement maitrisés par plus de 55% des sujets (74,4% pour le vocabulaire; 71,3% pour la syntaxe; 59,9% pour l'extraction des informations explicites et 67,7% pour la compréhension des informations implicites). Toutefois, la capacité diagnostique des items (r) varie entre 0,237 et 0,852, ce paramètre est encore faible chez certains items probablement parce que le test n'a pas été conçu spécifiquement pour une visée diagnostique (Li, 2012; Li et Suen, 2013). Ces paramètres vont être abordés plus en détail dans les lignes suivantes.

Développé à partir du modèle unifié de Diello et ses collaborateurs, le modèle de Fusion (Hartz, 2002; Roussos, Dibello et al., 2007; Jang, 2005, 2009; Li, 2011; Li et Suen, 2013) est un modèle multidimensionnel de la TRI qui suppose que le sujet doit

maitriser tous les attributs demandés pour un item afin d'y répondre correctement (Jang, 2005). Il s'inscrit donc dans la famille des modèles conjonctifs. À la différence du modèle de Rule-Space qui estime seulement deux paramètres, le modèle de Fusion est un modèle à trois paramètres qui peut s'appliquer à la fois aux données dichotomiques ou polychotomiques. Le paramètre $\boldsymbol{\pi}_i^*$ représente la probabilité de répondre correctement à un item lorsqu'un sujet maîtrise tous les attributs nécessaires. Ce paramètre peut être interprété comme le niveau de difficulté de la matrice Q pour l'item i, qui varie entre 0 et 1. Le paramètre r_{ik}^* permet de comparer la probabilité de répondre correctement à l'item i lorsque le sujet maîtrise l'attribut k et dans le cas de la non-maîtrise de l'attribut k (Jang, 2005). Autrement dit, ce paramètre peut être considéré comme un indicateur de la capacité diagnostique de l'item i pour l'attribut k et prend aussi les valeurs entre 0 et 1. Plus l'attribut k est nécessaire pour l'item 1, plus petite est la valeur de r_{ik}^*. Cette valeur est donc interprétée comme le paramètre de discrimination de l'item 1 pour l'attribut k (Li, 2011). Finalement, le paramètre c_i qui varie entre 0 et 3 représente l'exclusivité de la liste d'attributs dans la matrice Q permettant de vérifier si cette matrice Q peut contenir tous les attributs nécessaires (Loye, 2010; Jang, 2005).

L'avantage le plus important du modèle de Fusion est qu'il reconnait le caractère incomplet de la matrice Q en témoignant des attributs non pris en compte mais qui ont été utilisés par le sujet pour répondre correctement aux items (Roussos, Dibello et al., 2007; Li, 2011). Étant donné que nous ne pouvons pas connaitre tous les types de connaissances et d'habiletés sous-jacentes dans le processus de la compréhension en lecture de chaque sujet, il est impossible de pouvoir identifier tous les attributs cognitifs nécessaires pour répondre correctement à un item (Li, 2011). D'ailleurs, le modèle permet non seulement d'évaluer la performance du sujet sur chaque attribut, mais aussi la capacité diagnostique de chaque item et du test grâce au paramètre r_{ik}^*. Plus cette valeur est petite, plus grande est la capacité diagnostique de l'item (Roussos, Xu et Stout, 2003; Li et Suen, 2013). Le désavantage réside peut-être dans la difficulté de l'interprétation des résultats, car le modèle contient beaucoup de paramètres à estimer. D'ailleurs, si ces paramètres sont estimés par le MCMC, il est difficile d'atteindre et de juger la convergence (Sinharay, 2004; Li et Suen, 2013; Loye, 2008). C'est pour ces deux raisons que nous n'avons pas choisi le modèle de Fusion pour l'analyse des données.

2.2.2.4. D'autres modèles

Lee et Sawaki (2009) ont mené une étude avec le test TOEFL iBT pour diagnostiquer à la fois la performance en lecture et à l'écoute. À la différence de la recherche de Jang (2005), seulement 4 attributs ont été identifiés pour répondre correctement aux items du test, à savoir : (a) comprendre le sens du vocabulaire; (b) comprendre des informations spécifiques; (c) connexion des informations; (d) synthèse et organisation des informations. Ces quatre attributs peuvent donc être classifiés en connaissances linguistiques (a); en compréhension locale (b) et en compréhension globale (c et d). Toutefois, l'analyse des données se fait non seulement avec le modèle de Fusion, mais aussi avec le GDM et le modèle des classes latentes (LCM) permettant d'obtenir des similitudes et différences dans la classification des profils de maitrise des candidats avec ces trois modèles. Avec ce même modèle, GDM, Von Davier (2005) a également appliqué au test TOEFL iBT en analysant les formes A et B du test, qui traitent à la fois des données dichotomiques et polychotomiques. Avec les 4 habiletés identifiées, les résultats obtenus ont démontré l'applicabilité du GDM à des tests à grande échelle en langue.

Le GDM (von Davier, 2005; von Davier et Yamamoto, 2004; Lee et Sawaki, 2009) est un large cadre de référence analytique pour rassembler des modèles diagnostiques cognitifs qui sont précédemment disponibles et pour combiner les caractéristiques de la TRI, des modèles log-linéaires et des analyses des classes latentes (Lee et Sawaki, 2009). Ainsi, le GDM englobe plusieurs modèles qui peuvent être utilisés pour des analyses diagnostiques cognitives tels que les modèles de classification multiple des classes latentes de Maris (1985), la version compensatoire de Hartz (2002) et la version de crédit partiel de GDM (pGDM) (von Davier, 2005; Lee et Sawaki, 2009). Ce modèle de crédit partiel de GDM a été utilisé dans la recherche en compréhension orale (écoute) et en compréhension écrite (lecture) de Lee et Sawaki (2009) avec le test de TOEFL iBT ainsi que dans la recherche de von Davier (2005). Supposons qu'il y a k attributs, les probabilités de répondre correctement aux items pour la version logistique du pGDM peuvent être estimée par le $\boldsymbol{\beta}_{xi}$, le paramètre de seuil de la réponse x à l'item i ; i_k, le paramètre de pente de l'attribut k pour chaque catégorie de réponse non nulle ; ik, l'entrée de la matrice Q pour l'item i et l'attribut k et le a_k, la variable latente

multidimensionnelle, $\theta = (a_1, \ldots a_k)$ (Lee et Sawaki, 2009). Ces paramètres peuvent être estimés par le logiciel mdltm (von Davier, 2005) en utilisant la vraisemblance marginale maximum (Marginal maximum likelihood).

Cette version de crédit partiel du GDM est un modèle à la fois compensatoire et non compensatoire qui traite des données dichotomiques et polychotomiques. Dans le cas du modèle de Fusion, les données polychotomiques devraient être dichotomiquement recodées tandis qu'avec le GDM, nous pouvons travailler avec des données polychotomiques telles quelles. Le GDM offre donc plus d'options d'analyses avec les différents types de données (Lee et Sawaki, 2009). Toutefois, dans notre projet, il y a peu d'items qui sont polychotomiques de nature (seulement deux items), nous avons donc décidé de les coder dichotomiquement. Ainsi, l'avantage de travailler avec les données polychotomiques du GDM ne nous semble pas très utile dans notre cas.

D'ailleurs, la méthode d'attribut hiérarchique (AHM) a été récemment utilisée pour des analyses diagnostiques cognitives des tests en lecture, comme dans le cas du test SAT (Wang et Gierl, 2011). La recherche été réalisée en deux étapes : dans la première étape, trois modèles cognitifs ont été développés basés sur les balises théoriques en lecture telles que les taxonomies de Bloom (1956), le modèle de Van Dijk et Kintsch (1983) présentant les trois niveaux de la représentation du texte : la surface (le vocabulaire et la syntaxe); le « textbase » (faire des inférences et la compréhension locale) et le modèle de la situation qui fait référence aux interactions entre le texte explicite et les connaissances de base (background knowledge). Quelques recherches empiriques sur l'analyse des attributs du SAT ont été réalisées et les sept attributs suivants ont été identifiés : (a) comprendre le sens du vocabulaire; (b) comprendre le contenu, la forme et les fonctions des phrases; (c) comprendre la situation impliquée dans un texte; (d) comprendre le contenu, la forme et les fonctions des différentes parties du texte; (e) analyser les objectifs et des stratégies de l'auteur; (f) utiliser des connaissances de raisonnement et de la résolution des problèmes; (g) utiliser des stratégies métacognitives.

La deuxième étape de la recherche de Wang et Gierl (2011) consiste à faire des analyses psychométriques afin de déterminer le modèle qui s'ajuste le mieux aux données des 2000 candidats. Une fois le modèle choisi, les probabilités de maîtrise des attributs ont été estimées pour un échantillon de 15 candidats (Wang et Gierl, 2011). Avec les 9

attributs identifiés, le AHM peut prédire 84,4 % des scores totaux des candidats. Les rapports diagnostiques ont été développés par le AHM permettant de fournir aux candidats des informations détaillées sur les habiletés cognitives et le niveau de maitrise de ces habiletés. Toutefois, à la différence d'autres recherches menées avec les MCD, le AHM exige que les attributs soient classés hiérarchiquement selon leur degré de difficulté afin d'obtenir la consistance des résultats. Cet aspect restreint les possibilités avec un test à grande échelle qui n'a pas été conçu à visée diagnostique (Wang et Gierl, 2011), ce qui justifie la raison pour laquelle nous n'avons pas choisi ce modèle pour analyser nos données.

2.2.2.5. Les modèles de DINA et de G-DINA

En comparaison des modèles de Fusion, de Rule-Space ou de GDM, les modèles de DINA et G-DINA sont encore peu utilisés dans le domaine des langues. La seule étude est celle de Ravand, Berati et Widhiarso (2012) avec le test GET (General English Test), un test d'entrée à l'université en lecture en Iran. L'analyse des données a été réalisée avec le modèle de DINA auprès 1500 candidats au doctorat avec 5 attributs identifiés par les experts, à savoir : (a) le vocabulaire; (b) la syntaxe; (c) le repérage; (d) la connexion et synthèse et (e) les inférences. Les résultats suggèrent que les indices de pseudo-chance et d'étourderie des items sont trop élevés (0,36 et 0,38), ce qui fait que la capacité diagnostique des items est inférieure à 0,5. Le paramètre de pseudo-chance renvoie à la probabilité qu'un sujet peut répondre correctement à un item même s'il ne maitrise pas tous les attributs nécessaires tandis que le paramètre d'étourderie représente la probabilité qu'un sujet peut ne pas donner une bonne réponse même s'il maitrise tous les attributs demandés. Deux raisons ont été proposées pour expliquer ces paramètres élevés de pseudo-chance et d'étourderie: la nature conjonctive plutôt que compensatoire des habiletés, ce qui fait que les sujets ne sont pas nécessairement obligés de maitriser tous les attributs pour répondre correctement à l'item. La deuxième raison réside dans l'erreur de la spécification de la matrice Q, étant donné que le paramètre qui témoigne du caractère incomplet de la matrice Q n'est pas estimé avec le modèle DINA (Ravand, Berati et Widhiarso, 2012).

Le modèle DINA (Deterministic Inputs, Noisy and Gate model) (Cui, Gierl et Chang, 2012; de la Torre et Douglas, 2008; de la Torre, 2011; Junker et Sijtsma, 2001) divise les sujets en deux classes latentes pour chaque item: ceux qui maitrisent tous les attributs exigés pour un item (ξ_{ij} =1) et ceux qui ne les maitrisent pas (ξ_{ij} =0) (Cui, Gierl et Chang, 2012). Il s'agit d'un modèle non compensatoire qui suppose que le sujet doit maitriser tous les attributs nécessaires afin de répondre correctement aux items. Le modèle prend également en considération le fait que le sujet peut donner une mauvaise réponse même s'il maitrise tous les attributs nécessaires (Loye, 2010). Ainsi, il estime la probabilité de répondre correctement à un item avec deux paramètres: le paramètre de pseudo-chance (g_i) et le paramètre d'étourderie (s_i). Idéalement, ces deux paramètres de pseudo-chance et d'étourderie d'un item devraient être assez petits pour montrer que cet item a une grande capacité diagnostique. Les relations entre ces paramètres sont représentées comme suit:

$$P(X_{ij} = 1 \mid \xi_{ji}, s_j, g_j) = (1 - s_j)^{\xi_{ji}} g_j^{1-\xi_{ji}}$$

Ainsi, le modèle DINA estime la probabilité de répondre correctement à un item en fonction des probabilités des paramètres de pseudo-chance (g_ig$_i$) et d'étourderie (s_i) dépendant de deux classes latentes distinguées par le modèle. Plus spécifiquement, pour le groupe qui maîtrise tous les attributs, la probabilité de répondre correctement à un item est égale à 1-s_is$_i$, tandis que pour le groupe qui ne maitrise pas tous les attributs, cette probabilité est égale à g_i. Le tableau 3 résume donc ces probabilités selon les deux groupes latents.

Tableau 3. Probabilités de réponse dans le modèle de DINA (adapté de Rupp, Templin et Henson, 2010).

	X_{ij}=1 (Réponse correcte)	X_{ij}=0 (Réponse incorrecte)
ξ_{ij} =1 (Maitrise de tous les attributs)	1-s_i	s_i
ξ_{ij} =0 (Non-maitrise de tous les attributs)	g_i	1-g_i

L'avantage du modèle DINA réside dans sa simplicité tandis que sa principale limite est que la probabilité de bonnes réponses ne varie pas selon « le nombre et le type d'habiletés qui ne sont pas maitrisées » (Loye, 2010, p.86, Roussos et al., 2007). Afin de combler cette limite, le modèle G-DINA (generalized DINA) a été développé par de la Torre (2011). Le G-DINA peut être classé parmi les MCD généraux comme le GDM de von Davier (2005) qui ne tiennent pas compte de la relation restreinte comme conjonctive ou disjonctive des attributs afin de répondre correctement à un item (Ravand, Barati et Widhiarso, 2012). À la différence du modèle DINA, le modèle G-DINA assouplit l'hypothèse de la probabilité égale de réponses correctes lorsque le sujet ne maitrise pas tous les attributs qu'il faut pour cet item. Ainsi, au lieu de séparer les sujets en deux classes latentes pour chaque item comme dans le cas de DINA, le G-DINA partitionne les classes latentes en $2^{K_j^*}$ groupes latents, dont K_j^* est le nombre d'attributs demandés pour l'item j. Chaque groupe latent représente un vecteur d'attribut réduit α_{ij}^* qui obtient sa propre probabilité de réussite (de la Torre et Douglas, 2008).

Le G-DINA présume que même si les sujets ne peuvent pas maitriser tous les attributs qu'il faut pour un item (ξ_{ij} =0), les probabilités d'obtenir une réponse correcte peuvent varier. À titre d'exemple, pour un item nécessitant 3 attributs afin d'y répondre correctement, le sujet qui maîtrise deux attributs va avoir une plus grande probabilité de réussite que celui qui maitrise seulement un attribut. Ainsi, si le DINA peut différencier seulement deux groupes (ceux qui maitrisent tous les attributs et ceux qui ne maitrisent pas tous les attributs), le G-DINA peut différencier les sujets avec les différents niveaux de maitrise (Ravand, Barati et Widhiarso, 2012).

2.2.2.6. Synthèse et choix de DINA et G-DINA

Les lignes précédentes ont permis de faire une recension des écrits des recherches empiriques réalisées en lecture avec les MCD tels que le modèle Rule-Space, le modèle Fusion, le GDM, les modèles DINA et G-DINA ainsi qu'une description de ces modèles. Dans l'ensemble, les habiletés évaluées peuvent être regroupées en trois catégories principales : (a) les connaissances linguistiques; (b) la compréhension locale; et (c) la compréhension globale. Le tableau 4 présente une synthèse des habiletés évaluées dans

les recherches diagnostiques en lecture. Nous avons discuté également les avantages et les désavantages de chaque modèle, ce qui permet de faire émerger les raisons pour lesquelles les modèles DINA et G-DINA ont été choisis pour notre projet.

Le modèle de ***Rule-Space*** a été utilisé dans plusieurs recherche en langue, tant en compréhension de l'écrit qu'en compréhension orale (Buck, Tatsuoka et Kostin, 1997; Buck et al., 1998; Kansai, 1997; Scott, 1998) et on voit bien son applicabilité dans l'évaluation diagnostique cognitive des compétences langagières. Malgré la facilité de l'interprétation des résultats grâce à la visualisation graphique des états de connaissances, ce modèle ne peut fournir que les scores des attributs aux candidats qui sont classés avec succès dans leurs états de connaissances. D'ailleurs, le modèle peut commettre des erreurs dans le classement des états de connaissances lorsque le nombre de schémas de réponses est élevé. En comparaison du modèle Rule-Space, le modèle ***Fusion*** peut fournir plus d'informations sur le caractère incomplet de la matrice Q avec le paramètre résiduel c_i pour tenir compte du fait que tous les autres attributs ont été utilisés par le sujet mais qui n'ont pas été identifiés dans la matrice Q. Ceci engendre, toutefois, la difficulté dans l'interprétation et dans l'estimation étant donné que le modèle contient plus de paramètres. En outre, la convergence est plus difficile à atteindre étant donné que ces paramètres sont estimés avec le MCMC (Sinharay, 2004; Li et Suen, 2013). Le ***GDM***, quant à lui, offre la possibilité de travailler avec les items polychotomiques. Toutefois, cet avantage n'est pas très pertinent dans notre projet, car nous avons peu d'items qui sont polychotomiques de nature. Une des caractéristiques importantes de la ***méthode AHM*** est que les attributs doivent être classés hiérarchiquement afin d'obtenir la consistance des résultats. Cette caractéristique n'est donc pas convenable avec le contexte du test PIRLS dont la conception en lecture met davantage l'accent sur l'interaction entre les différentes stratégies cognitives lors de la compréhension en lecture et non pas sur une relation hiérarchique des attributs.

Lors du choix des modèles pour notre recherche, les modèles DINA et G-DINA attirent beaucoup notre attention pour plusieurs raisons. *Premièrement*, en comparaison avec d'autres modèles, les modèles DINA et G-DINA sont encore peu utilisés dans le domaine des langues jusqu'à présent. Ils sont, par contre, appliqués avec succès aux modélisations des données en mathématique ou issues de simulations (Cui, Gierl et

Chang, 2012; de la Torre, 2008; de la Torre et Douglas, 2004; de la Torre, 2011). *Deuxièmement*, il s'agit des modèles les plus simples, donc les plus restrictifs et interprétables dans la famille des MCD qui peuvent traiter les données dichotomiques (de la Torre et Douglas, 2008). En effet, selon DiBello, Russos et Stout (2007), lors du choix d'un MCD pour l'analyse des données, il faut tenir compte de deux caractéristiques: la faisabilité et la parcimonie. Ces caractéristiques sont liées à l'importance de garder les modèles aussi simples que possible en termes de paramètres et d'arriver à un ajustement adéquat des données afin de bien atteindre l'objectif de diagnostic. *Finalement*, le DINA et le G-DINA sont des modèles qui peuvent se compenser l'un l'autre, dans le sens où le modèle G-DINA peut combler une limite principale du modèle DINA qui renvoie au fait que la probabilité de bonnes réponses ne varie pas selon « le nombre et le type d'habiletés qui ne sont pas maitrisées » (Loye, 2010, p.86). Ainsi, dans un item, au lieu de classer les candidats seulement en deux groupes latents (répond correctement ou incorrectement à l'item), le modèle G-DINA permet de différencier les sujets en plusieurs niveaux de probabilités de répondre correctement à cet item même s'ils ne maitrisent pas tous les attributs nécessaires. Ces trois raisons justifient notre choix de ces modèles pour les analyses diagnostiques cognitives d'un test en lecture comme le PIRLS 2011.

Tableau 4. Synthèse des habiletés identifiées dans les recherches diagnostiques en lecture

Recherches	**Connaissances linguistiques**				**Compréhension locale**		**Compréhension globale**	
	Vocabulaire	**Syntaxe**	**Négation**	**Organisation rhétorique**	**Localisation des informations**	**Faire des inférences**	**Informations implicites**	**Synthèse et connexion**
Buck, Tatsuoka et Kostin (1997)	**X**	**X**	**X**	**X**				
Kasai (1997)	**X**	**X**		**X**	**X**			
Scott (1998)	**X**	**X**		**X**	**X**			
Jang (2009)	**X**	**X**				**X**	**X**	
Lee et Sawaki (2009)	**X**				**X**			**X**
Li (2011)	**X**	**X**			**X**		**X**	
Von Davier (2005)	**X**				**X**			**X**
Wang et Gierl (2007)	**X**				**X**			**X**
Ravand, Barati et Widhiarso (2012)	**X**	**X**			**X**	**X**		**X**

2.2.3. Limites de l'approche diagnostique cognitive

La première limite de l'ADC est due au manque de cadre de référence pour les évaluations diagnostiques dans le domaine des langues (Lee et Sawaki, 2009). C'est pour cette raison que jusqu'à présent, la plupart des recherches sont réalisées avec les tests à grande échelle existants qui ne sont pas conçus spécifiquement dans une visée diagnostique (Buck et Tatsuoka, 1998; Gao, 2006 ; Jang, 2005; Kasai, 1997; Lee et Sawaki, 2009). En effet, dans le contexte des évaluations à grande échelle, la définition des modèles cognitifs sous-jacents qui expliquent les habiletés en lecture présente encore certaines limites. Selon Leighton et Gierl, (2006), un modèle cognitif devrait fonctionner en se basant sur : (a) la construction des indices et des principes de l'évaluation qui visent à déterminer des tâches d'évaluation; (b) les types de connaissances mesurées; et (c) le lien entre les cibles de l'évaluation et les compétences mesurées, les différents niveaux de maitrise des habiletés.

Pourtant, dans le contexte des évaluations à grande échelle, cette définition ne semble pas tout à fait convenable, étant donné que peu de tests à grande échelle sont capables de diagnostiquer le processus de réflexion des apprenants, car ils sont développés avec des objectifs certificatifs et produisent des scores globaux plutôt que diagnostiques. Par ailleurs, ils ne sont pas conçus en se basant sur des théories ou des modèles d'apprentissage cognitifs. En réalité, la définition d'un modèle cognitif ne se base que sur une série des jugements des experts face aux tâches observées. Ces croyances ne peuvent pas être considérées comme un modèle cognitif, car il manque de consensus et de preuves empiriques. Ainsi, afin que ces modèles soient appliqués dans le contexte des évaluations à grande échelle, nous sommes obligés d'accepter une certaine limite dans sa définition (Leighton et Gierl, 2006).

Dans la plupart des recherches, les habiletés sont identifiées à partir des jugements des experts du domaine, la définition et la construction de la matrice Q restent encore très subjectives, ce qui constitue la deuxième limite de l'ADC. Comme les stratégies cognitives sous-jacentes peuvent être différentes entre les individus, il est important que cette matrice Q soit validée auprès des participants à travers des protocoles verbaux, ce qui permet d'éviter les erreurs lors de la spécification de la matrice Q (Lee et Sawaki, 2009).

Finalement, la recension des écrits sur les recherches empiriques réalisée en langues montre que les informations diagnostiques obtenues sont peu exploitées d'une manière adéquate (Lee et Sawaki, 2009). Étant donné que l'ADC peut fournir un profil détaillé avec des informations spécifiques sur le degré de maitrise des habiletés cognitives, ces informations devraient être intégrées dans des rapports diagnostiques compréhensibles et interprétables. Le tableau 5 présente une synthèse de ces recherches avec les tests utilisés, le nombre d'habiletés évaluées, les MCD choisis ainsi que l'élaboration et l'évaluation des rapports diagnostiques. Toutefois, parmi les onze recherches réalisées en lecture avec l'ADC, seulement deux visent à concevoir des rapports diagnostiques et une seule recherche a validé ces rapports auprès d'étudiants et d'enseignants. Cette réalité montre qu'il manque encore des recherches empiriques visant à développer et évaluer des rapports diagnostiques destinés aux enseignants et aux apprenants (Jang 2005; Roberts et Gierl, 2010). Notre recherche vise donc à combler ce manque. Alors, comment développer et évaluer des rapports diagnostiques compréhensibles et interprétables en tenant compte des caractères spécifiques des informations issues de l'ADC? Les éléments de réponse vont être trouvés dans la prochaine partie de notre cadre conceptuel.

2.3. Élaboration et évaluation des rapports diagnostiques

Les rapports élaborés à la suite d'une évaluation jouent un rôle extrêmement important en éducation, car ils sont considérés comme un moyen de communication entre les concepteurs des tests et les différents publics visés (Ryan, 2006; Roberts et Gierl, 2010). La communication et l'interprétation des résultats deviennent donc la responsabilité de tous les éducateurs, notamment ceux qui s'engagent directement dans la remédiation des difficultés des élèves et l'amélioration des stratégies d'enseignement, comme c'est le cas des enseignants (Klesch, 2010). Malgré le rôle important des rapports dans les programmes d'évaluation, la quantité des recherches sur l'élaboration et l'évaluation des rapports, notamment ceux qui sont destinés aux enseignants restent encore très modestes (Ryan, 2006; Roberts et Gierl, 2010) et la plupart de ces recherches ne tiennent pas compte du rôle actif du public visé lors du processus d'élaboration de ces rapports (Zapata-Rivera et Kartz, 2014; Vezzu, Vanwinkle et Zapata-Rivera, 2012). Ce point constitue le deuxième objectif général de notre recherche qui vise à combler cette lacune.

Tableau 5. Synthèse des recherches réalisées en lecture avec les MCD

Recherches	Tests utilisés	Nombre d'habiletés	MCD	Élaboration des rapports	Évaluation des rapports
Buck, Tatsuoka et Kostin (1997)	TOEIC	24	Rule-Space	Non	Non
Kasai (1997)	TOEFL	27	Rule-Space	Non	Non
Scott (1998)	TOEFL	24	Rule-Space	Non	Non
Svetina, Gorin et Tatsuoka (2011)	Test d'entrée au collège aux E-U	22	Rule-Space	Non	Non
Jang (2009)	TOEFL iBT	9	Fusion	Oui	Oui
Lee et Sawaki (2009)	TOEFL iBT	4	Fusion, GDM, LCA	Non	Non
Sawaki, Kim et Gentile (2009)	TOEFL iBT	4	Fusion	Non	Non
Li (2011)	METLAB	4	Fusion	Non	Non
Von Davier (2005)	TOEFL iBT	4	GDM	Non	Non
Wang et Gierl (2007)	SAT verbal	7	AHM	Oui	Non
Ravand, Barati et Widhiarso (2012)	GET	5	DINA	Non	Non

Ainsi, dans cette partie, nous présenterons, dans un premier temps, les différents fonctions et destinataires des rapports. Dans un deuxième temps, nous analyserons les cadres de référence d'élaboration et d'évaluation des rapports avec les quatre étapes principales: (a) identifier le public visé et analyser ses besoins; (b) documenter les formats des rapports qui existent; (c) développer les rapports; et (d) évaluer les rapports. Finalement, nous présenterons une synthèse sur les forces et les faiblesses des méthodes utilisées dans les recherches empiriques sur l'évaluation des rapports.

2.3.1. Fonctions et destinataires des rapports

Les rapports jouent plusieurs fonctions qui peuvent varier selon les objectifs des tests conçus et les destinataires visés. Selon Kolen (2006), les rapports élaborés sont l'une des composantes les plus visibles des programmes d'évaluation. Certains rapports peuvent refléter la position relative d'un candidat par rapport au groupe de référence, d'autres présentent la performance selon les normes établies par les concepteurs tandis qu'il y a des rapports qui renseignent seulement sur les résultats à une sous-partie d'un test. Dans certains cas, plusieurs rapports sont élaborés pour communiquer les résultats d'un même test, car ce test a été conçu pour plusieurs objectifs (Kolen, 2006, p.155).

Klesh (2010) a identifié les quatre fonctions suivantes des rapports: (a) fournir des résultats d'une évaluation sous forme de scores bruts, ou de pourcentages des réponses correctes ou incorrectes; (b) fournir le contexte des résultats en incluant des informations sur la performance des autres candidats ou sur les normes ou les critères évalués; (c) permettre la catégorisation des candidats sur une même échelle de mesure et (d) fournir des informations diagnostiques sur le déficit des connaissances ou d'habiletés des candidats à travers des réponses correctes ou incorrectes aux tests.

Ryan (2006), quant à lui, considère que les rapports sont conçus pour deux objectifs principaux. Premièrement, les rapports peuvent répondre à un objectif pédagogique qui vise à informer différents publics de l'apprentissage des élèves, de l'efficacité de l'enseignement, des impacts et de la valeur du curriculum. Le deuxième objectif fondamental des rapports est de fournir des résultats des programmes d'évaluations aux responsables locaux, provinciaux et nationaux. Avec ces deux objectifs, les destinataires des rapports sont essentiellement les apprenants, les parents, les enseignants, les

directeurs d'école, les administrateurs gouvernementaux et le grand public. D'après Zapata-Rivera et Kartz (2014) et Ryan (2006), l'efficacité de différents types de rapports reflète la connexion entre ce rapport, les publics visés ainsi que les objectifs anticipés. Un rapport qui peut être utile pour un groupe ne peut pas toujours répondre aux attentes d'un autre (Ryan, 2006). Ainsi, les rapports devraient être élaborés et évalués en tenant compte du public visé ainsi que des objectifs établis. Dans le cadre de ce projet, nous nous intéressons à concevoir des rapports diagnostiques destinés aux enseignants parce que ce sont les enseignants qui ont le moyen de mettre en place des approches pédagogiques pour tenter de remédier aux faiblesses identifiées chez leurs élèves. La partie suivante présentera donc les cadres de référence d'élaboration et d'évaluation des rapports destinés aux enseignants.

2.3.2. Cadres de référence d'élaboration et d'évaluation des rapports

Zapata-Rivera, VanWinkle et Zwick (2012), Vezzu, VanWinkle et Zapata-Rivera (2012), Zapata-Rivera (2011) ont développé un cadre de référence pour l'élaboration de rapports interactifs en mathématique dans le cadre du CBAL (Cognitively Based Assessment of, for, and as Learning) destinés aux enseignants et aux élèves. En s'inspirant des méthodologies dans les domaines de la conception en évaluation (assessment design) et de l'ingénierie des logiciels (Mislevy, Steinberg et Almond, 2003), ce cadre de référence comprend les quatre étapes suivantes: (a) collecter des informations sur l'évaluation des besoins; (b) concilier ces besoins avec les informations disponibles; (c) élaborer des rapports types; et (d) évaluer ces rapports avec les experts internes et externes. La figure 6 présente les relations entre ces étapes.

L'étape de *collecter des informations sur l'évaluation des besoins* consiste à rassembler les informations sur les besoins des personnes concernées, y compris les concepteurs qui s'occupent du contenu des rapports et les publics visés. Dans cette étape, nous pouvons utiliser des informations qui sont disponibles, comme les résultats des recherches empiriques qui ont été menées auprès du même public (Zapata-Rivera, VanWinkle et Zwick, 2012). Les informations obtenues dans cette phase peuvent fournir des idées initiales sur les attentes des enseignants sur les rapports.

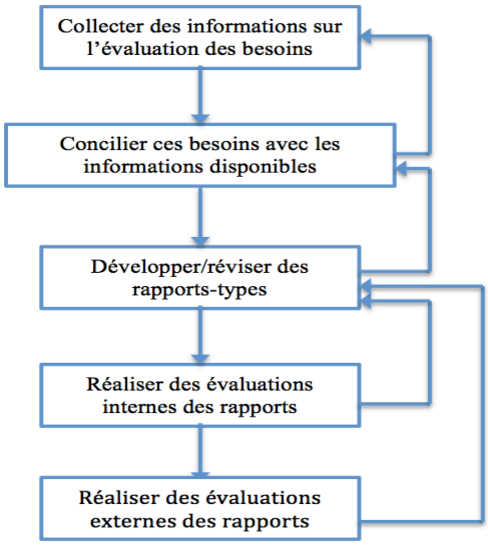


Figure 6. Cadre de référence du développement et de l'évaluation des rapports (adapté de Zapata-Rivera, VanWinkle et Zwick, 2012)

La conciliation des besoins des enseignants et les informations disponibles permet d'assurer la cohérence entre leurs attentes et les résultats qui devraient être présentés et communiqués dans les rapports, plus précisément, sur les types d'informations fournies ainsi que sur la manière de présenter ces informations (Zapata-Rivera, VanWinkle et Zwick, 2012). Les rapports types sont ensuite produits en respectant des principes de la représentation des rapports tels que proposés par Fast (2002), Hatie (2009), Goodman et Hambleton (2004) que nous allons aborder dans les lignes suivantes. Ces rapports sont finalement évalués par les experts internes sur certains critères comme le contenu, l'utilité et l'accessibilité et évalués par le public visé à travers des études qualitatives ou quantitatives permettant d'améliorer les rapports ou de proposer des recommandations sur l'élaboration de ces rapports en général (Zapata-Rivera, VanWinkle et Zwick, 2012).

En comparaison avec ce cadre de référence, celui proposé par Hambleton et Zenisky (2010) présente des points semblables (voir le tableau 6). Plus précisément, il contient sept étapes : (a) définir l'objectif des rapports; (b) identifier le public visé; (c) faire la revue de littérature sur les types de rapports; (d) développer des rapports; (e) collecter des données; (f) réviser et reproduire les rapports; et (g) entretenir les rapports. Toutefois, si dans le cadre de référence proposé par Zapata-Rivera, VanWinkle et Zwick

(2012), le public visé est identifié dès le début, cette étape sera réalisée après avoir défini l'objectif des rapports dans le cadre de référence de Hambleton et Zenisky (2010). En outre, l'étape de « faire la revue de littérature sur les types de rapports » est une phase séparée dans le cadre de référence de Hambleton et Zenisky (2010) tandis qu'elle est regroupée dans la phase de « élaborer des rapports types » dans celui de Zapata-Rivera, VanWinkle et Zwick (2012).

Tableau 6. Cadre de référence du développement et de l'évaluation des rapports (adapté de Hambleton et Zinsky, 2010).

Préparation initiale	Développement des rapports	Validation et révision des rapports
• Définir des objectifs • Évaluer des besoins • Identifier le public visé	Développement des rapports types	• Valider sur le terrain • Collecter des données • Réviser (Répéter au besoin)

Fast et Tucker (2001), dans une recherche sur l'élaboration et l'évaluation des rapports pour le Programme de Testing du Connecticut (Connecticut Mastery Testing Program) ont suivi une démarche comprenant quatre étapes. La première étape consiste à faire une revue de littérature sur les types de rapports qui existent dans l'État du Connecticut et les exigences de cet État sur les rapports. Dans la deuxième étape, la revue de littérature a été élargie sur les types de rapports qui existent dans d'autres États. Une troisième étape vise à mener une série d'entrevues avec les parents, les enseignants et les administrateurs. Dans cette phase, les différents groupes de participants sont invités à donner leurs commentaires sur les différents types de rapports. Plus précisément, les parents et les enseignants examinent les rapports destinés aux élèves tandis que les enseignants et les administrateurs donnent leurs rétroactions sur les rapports conçus pour les classes et les rapports diagnostiques. Finalement, les informations obtenues des

différentes sources vont être utilisées pour élaborer et réviser des rapports destinés aux différents publics visés.

Klesch (2010), en menant une recherche sur l'élaboration des rapports destinés aux enseignants, a aussi adopté une démarche comprenant quatre étapes: (a) élaborer des rapports types et des guides d'interprétation; (b) réaliser des entrevues chez les groupes focus (enseignants) afin d'obtenir des rétroactions sur les rapports types; (c) analyser les données et réviser les rapports; et (d) fournir des rétroactions aux étudiants. Il a également proposé un cadre de référence comprenant huit points. Toutefois, à la différence des cadres de référence précédents qui mettent l'accent sur les démarches, le cadre de référence de Ryan (2003) prête plus d'attention aux caractéristiques qui figurent dans les rapports, à savoir (a) le public visé; (b) les échelles ou les scores métriques; (c) les références pour l'interprétation; (d) le moment et l'endroit de la passation; (e) les unités de mesure; (f) les erreurs de mesure; (g) les modes de présentation; et (h) les moyens de communication.

En se basant sur le cadre de référence général de Ryan (2003) et celui de Jaeger (1998), Roberts et Gierl (2010) ont proposé leur propre cadre de référence pour développer des rapports à visée diagnostique. Malgré des visées différentes des rapports, ces cadres de référence adoptent une approche similaire qui prend en compte le public visé et leurs besoins et attentes spécifiques ainsi que les objectifs des rapports et utilise des résultats de recherches empiriques pour développer et évaluer des rapports selon un cycle itératif (Zapata-Rivera, VanWinkle et Zwick, 2012). Toutefois, le cadre de référence proposé par Roberts et Gierl (2010) met l'accent sur le processus de l'élaboration des rapports diagnostiques sans tenir compte de l'évaluation de ces rapports auprès du public visé.

En bref, dans les lignes précédentes, nous avons fait une synthèse des cadres de référence d'élaboration et d'évaluation des rapports avec des similitudes et des différences entre ceux-ci. D'une manière générale, le cadre de référence pour l'élaboration et l'évaluation des rapports devrait se composer des étapes principales suivantes: (a) identifier le public visé et analyser ses besoins; (b) documenter les recherches et les formats de rapports qui existent; (c) développer des rapports et (d)

évaluer des rapports. Ces étapes vont être décrites plus en détail dans les parties suivantes.

2.3.2.1. Identifier le public visé et analyser ses besoins

L'identification du public visé et l'analyse de ses besoins est une étape indispensable avant toute élaboration et évaluation des rapports, car les utilisateurs des rapports sont des acteurs centraux du processus, qui vont faire des inférences et prendre des décisions à partir des résultats communiqués (Zapata-Rivera et Kartz, 2014). Cette étape devrait être réalisée avec soin, étant donné que chaque public visé peut avoir ses propres forces et faiblesses et prendre ses propres décisions à partir de ces résultats. Ainsi, si le concepteur peut définir le public visé et ses besoins d'une manière adéquate, il ouvrira la possibilité de raffiner les rapports afin de satisfaire les caractéristiques spécifiques et uniques de ce public (Vezzu, VanWinkle et Zapata-Rivera, 2012; Zapata-Rivera et Kartz, 2014).

Dans une recherche sur l'analyse des publics, Zapata-Rivera et Kartz (2014) se basent principalement sur le modèle de l'analyse des publics de Flower (1985). Ainsi, l'analyse du public visé doit porter sur les trois aspects suivants : les besoins, les connaissances et les attitudes du public visé (Flower, 1985). *Les besoins* revoient aux objectifs et aux raisons pour lesquelles ce public visé lit les rapports à partir desquels ils peuvent faire des inférences. Pour les enseignants, leurs besoins consistent à guider leur enseignement face à des questions comme: *Quelle est la performance de la classe sur le test? Quelles sont les forces et les faiblesses de mes élèves? Comment un score d'un élève peut-il être comparé à celui des autres élèves? etc*. (Underwood et al., 2007). *Les connaissances* portent non seulement sur ce que le public visé a déjà connu, mais aussi sur ce qu'il faut lui fournir afin de mieux comprendre les rapports. Ainsi, les enseignants connaissent bien leurs élèves ainsi que leurs niveaux de performance en classe. Ils ont aussi de bonnes connaissances sur les objectifs et le contenu des évaluations ainsi que sur les échelles de mesure (Zapata-Rivera et Kartz, 2014). Toutefois, leurs connaissances sont encore limitées au niveau des erreurs de mesure ou de l'interprétation des résultats et du langage utilisé dans les rapports (Hambleton et Slater, 1997; Underwood et al., 2007; Zapata-Rivera et al., 2011; Zapata-Rivera et Kartz, 2014). Finalement, *les attitudes*

réfèrent aux sentiments ou biais qui peuvent influencer leurs interprétations des résultats des rapports (Flower, 1985; Vezzu, VanWinkle et Zapata-Rivera, 2012). En tenant compte des caractéristiques spécifiques des enseignants, l'identification du public visé et l'analyse de ses besoins permettent de prendre les décisions sur le design des rapports (Zapata-Rivera et Kartz, 2014).

2.3.2.2. Documenter les recherches sur l'élaboration et les formats des rapports

Dans cette partie, nous proposerons, une recension des écrits des recherches en élaboration de rapports de manière générale puis dans une approche diagnostique cognitive. L'accent est mis sur les faiblesses des rapports élaborés qui aboutissent à des recommandations pour développer des rapports compréhensibles et accessibles.

2.3.2.2.1. Des recherches en élaboration des rapports

Parmi les recherches portant sur l'élaboration des rapports, celle de Godman et Hambleton (2004) attire l'attention de nombreux chercheurs du domaine, car elle fournit un cadre de référence très précis et complet sur l'élaboration des rapports (Ryan, 2006) et devient une référence théorique de base pour plusieurs recherches ultérieures telles que celles de Ryan (2006), Roberts et Gierl (2010) et Klesch (2010). La revue de littérature de Goodman et Hambleton (2004) se base principalement sur les recherches précédentes de Hambleton (2002), Hambleton et Slater (1997), Impara (1991), et Jaeger (1998) dans le contexte de l'élaboration des rapports pour le NAEP. Les auteurs ont identifié un grand nombre de faiblesses existant dans les rapports tant sur la forme que sur le fond qui peuvent influencer l'interprétation des résultats par les utilisateurs.

Au niveau de la forme, les rapports développés représentent : (a) un manque de variété dans les formats de présentation des résultats; (b) une surcharge d'informations et une densité des données présentées et (c) un encombrement dans la présentation de certaines données afin de faciliter la visualisation ou l'accessibilité. Quant au fond, les rapports ont les faiblesses majeures suivantes: (a) le niveau élevé des connaissances statistiques qui peuvent causer de la confusion et intimider certains utilisateurs; (b) le manque de support pour expliquer des termes, des concepts, des symboles techniques, ce qui empêche la compréhension des rapports par les utilisateurs; (c) le manque

d'informations descriptives qui peuvent montrer l'importance ou la signification des résultats d'évaluation; et (d) la présence de calculs arithmétiques inutiles (Aschbacher et Herman, 1991; Goodman et Hambleton, 2004; Klesch, 2010; Roberts et Gierl, 2010; Ryan, 2006; Sinharay et al., 2010; Zenisky et al., 2009).

Dans leur propre recherche, Goodman et Hambleton (2004) ont examiné les rapports de 11 États aux États-Unis et de deux provinces au Canada, avec trois compagnies de testing. Par l'usage des groupes de discussion lors de l'élaboration et de la révision des rapports, la recherche suggère des principes de base à respecter lors de l'élaboration des rapports tant pour la forme que pour le fond qui seront détaillés dans la partie suivante.

Dans la recherche de Fast et Tucker (2001) sur l'élaboration des rapports pour le Programme de Testing du Connecticut, des commentaires ont été obtenus à partir de différents groupes de discussion constitués de parents, d'enseignants et d'administrateurs sur les différents types de rapports qui leur sont destinés. Les auteurs ont identifié les éléments suivants comme sources de défis lors de l'élaboration des rapports : (a) les caractéristiques du format; (b) les représentations graphiques; (c) les représentations numériques; (d) les informations normatives; (e) les détails et les spécificités des rapports; (f) les moyens de support; (g) le glossaire des termes; (h) les directives pour les informations supplémentaires. Le fait de valider ces rapports auprès de différents publics visés est essentiel parce que cela permet de comprendre les attentes et les besoins spécifiques de chaque groupe afin d'assurer l'efficacité de ces rapports (Ryan, 2006).

Les recherches plus récentes de Zapata-Rivera et VanWinkle (2010) et Zapata-Rivera, VanWinkle et Zwick (2012) s'intéressent à élaborer des rapports interactifs en ligne en mathématique destinés aux enseignants et aux élèves dans le cadre du CBAL (Cognitively Based Assessment of, for, and as Learning). Trois différents types de rapports ont été développés et validés qualitativement et quantitativement auprès des enseignants: pour chaque individu, pour toute la classe et pour les informations du test, avec tutoriel et sans tutoriel. Ces rapports contiennent des informations dites « traditionnelles » comme les échelles de mesure, le niveau de performance et les scores bruts, ainsi que le guide d'interprétation et les liens pour accéder à des sources supplémentaires (définition des habiletés, exemple des tâches, explications des termes

statistiques) et les recommandations à suivre pour les enseignants (Zapata-Rivera et VanWinkle, 2010; Zapata-Rivera, VanWinkle et Zwick, 2012). À la différence des recherches précédentes qui s'attardent à l'élaboration des rapports en version papier, les recherches de Zapata-Rivera et VanWinkle (2010) et Zapata-Rivera, VanWinkle et Zwick (2012) s'intéressent plutôt à l'interprétation des résultats de ces rapports interactifs en ligne. Leurs résultats suggèrent que les informations doivent être présentées d'une manière concise et facile à lire, parce que les longs paragraphes sont souvent ignorés. Les recherches ont souligné l'importance des guides d'interprétation et du glossaire, car les enseignants peuvent ainsi comprendre des concepts généraux comme la difficulté des items ou l'erreur de mesure. Les détails de la méthode de l'évaluation de ces rapports vont être présentés dans notre partie sur l'évaluation des rapports.

En résumé, les lignes précédentes servent à décrire quelques recherches en élaboration et évaluation des rapports. De manière générale, les recherches peuvent être regroupées en trois grandes catégories : (a) les recherches sur la manière d'élaborer des rapports destinés à un public spécifique (Fast 2002; Goodman et Hambleton, 2004, Hambleton et Slater, 1997; Hattie, 2009; Underwood, Zapata-Rivera et VanWinkle, 2007; Zapata-Rivera, VanWinkle et Zwick, 2012); (b) les recherches sur les aspects techniques des rapports comme les représentations quantitatives des données avec les graphiques (Tufte, 1983, 1996; Wainer, 1997, 2005) et (c) les recherches sur l'interprétation et l'évaluation des rapports (Klesh, 2010; Zapata-Rivera et VanWinkle, 2010; Zapata-Rivera, VanWinkle et Zwick, 2012). Des principes et des recommandations pour un rapport efficace ont été suggérés dans ces recherches dont nous discuterons dans la partie sur le développement des rapports.

2.3.2.2.2. Des recherches en élaboration des rapports diagnostiques

Dans une approche diagnostique cognitive, la revue de littérature en élaboration et évaluation des rapports diagnostiques montre qu'il existe encore très peu de recherches sur le sujet. Le premier rapport diagnostique est le « Score Plus Report » élaboré par le College Board pour communiquer les résultats diagnostiques en mathématique, en littératie critique et en écrit de PSAT/NMSQT avec le modèle de Rule-Space. Les informations diagnostiques sont présentées sous forme des trois habiletés qui doivent être

le plus améliorées pour chaque matière ainsi que des stratégies de remédiation proposées (Roberts et Gierl, 2010). Un format de rapport semblable au « Score Plus Report » destiné aux apprenants et aux enseignants a été conçu à partir de l'analyse des résultats du test TOELF iBT avec le modèle Fusion (Jang, 2005; 2009). Un des objectifs de l'étude est de vérifier si les rétroactions diagnostiques obtenues sont utiles pour l'amélioration des habiletés en lecture chez les apprenants (Jang, 2005; Li, 2012). En tenant compte des besoins spécifiques des apprenants, ce rapport contient quatre parties principales: (a) vérifier vos réponses; (b) améliorer vos habiletés; (c) interpréter les rapports diagnostiques; et (d) décrire les habiletés avec les exemples des questions posées (Jang, 2005). Ainsi, les informations diagnostiques fournies prennent la forme des probabilités de maitrise des habiletés, des indices de discrimination des items ainsi que du descripteur des habiletés (Roberts et Gierl, 2010).

À la différence de la recherche de Jang (2005; 2009), celle de Roberts et Gierl (2010) prête beaucoup d'attention aux éléments du design en offrant une revue de littérature très riche sur les cadres de référence d'élaboration des rapports et des principes plus spécifiques de présentation des informations textuelles et non textuelles. En se basant sur les cadres de référence de Jaeger (1998) et Ryan (2003), les rapports diagnostiques ont été développés à partir des résultats d'un test en mathématique (Algebra and Functions of Applying Mathematical Knowledge) modélisés avec la méthode AHM. Les résultats communiqués portent sur les profils de maitrise des habiletés des apprenants comprenant trois parties: (a) les directives pour lire le rapport ; (b) l'examen des réponses avec la performance pour chaque habileté; (c) un résumé de la performance de l'apprenant. Le rapport est accompagné d'un guide fournissant des informations supplémentaires sur la description des habiletés, les profils de maitrise des habiletés et des questions souvent posées ainsi que l'interprétation des résultats du rapport. Ce rapport peut être considéré comme un exemple à suivre pour ceux qui s'intéressent à élaborer des rapports dans une approche diagnostique cognitive. Toutefois, cette recherche comporte également quelques limites. Premièrement, l'étape de définir les besoins et les objectifs spécifiques du public visé n'a pas été prise en compte dans cette recherche. Deuxièmement, la recherche se limite à l'élaboration des rapports sans les évaluer. Ainsi, il est souhaitable que de futures recherches soient menées afin de vérifier

l'efficacité de ces rapports dans l'enseignement et l'apprentissage (Roberts et Gierl, 2010).

La même méthode AHM a été appliquée dans la recherche de Wang et Gierl (2012) pour analyser des données du test en lecture SAT. Avec les résultats obtenus, les auteurs ont fourni l'exemple d'un rapport diagnostique de la performance des candidats. Le rapport présente les informations sur le score total de chaque candidat ainsi qu'une description de sa performance, de ses forces et de ses faiblesses et lui suggère des aspects à améliorer. Toutefois, ces informations sont présentées seulement sous forme de texte, ce qui peut attirer moins l'attention des candidats. Comme dans l'étude de Robert et Gierl (2010), cette recherche ne s'arrête qu'à la présentation des informations diagnostiques obtenues sans les valider auprès du public visé. Elle ne tient pas compte non plus des principes de la présentation des rapports. Ainsi, des recherches futures devraient porter sur ces aspects (Wang et Gierl, 2012).

En résumé, la partie précédente est consacrée à présenter une synthèse des recherches en élaboration des rapports diagnostiques avec leurs forces et les faiblesses. Une des faiblesses majeures de ces recherches est qu'elles s'intéressent plutôt à l'élaboration qu'à l'évaluation des rapports. À part la recherche de Roberts et Gierl (2010) qui se base sur les cadres de référence d'élaboration des rapports suggérés à partir de la revue de littérature, les autres ne suivent pas un cadre de référence particulier pour l'élaboration des rapports. Plus spécifiquement, celles de Jang (2005 et 2009) et de Wang et Gierl (2012) ne prêtent pas attention aux caractères de la présentation des informations. En outre, les recherches de Roberts et Gierl (2010) et de Wang et Gierl (2012) ne tiennent pas compte des besoins spécifiques et des objectifs du public visé.

Dans une ADC, les concepteurs devraient donc exploiter de nouveaux types d'informations diagnostiques telles que les descriptions des habiletés et la maitrise de ces habiletés autres que les scores totaux et les percentiles issus des évaluations à grande échelle (Roberts et Gierl, 2010). Ainsi, comment intégrer des informations diagnostiques détaillées, mais complexes aux informations techniques sophistiquées issues de l'ADC et les communiquer à un public non professionnel d'une manière compréhensible et efficace? (Roberts et Gierl, 2010). Le processus de développement des rapports

diagnostiques avec les critères principaux fournit des éléments de réponses afin de relever ce défi.

2.3.2.3. Développer des rapports

Plusieurs guides d'élaboration des rapports efficaces ont émergé dans la littérature (Roberts et Gierl, 2010) et varient peu selon les programmes d'évaluation et les provinces (Godman et Hambleton, 2004). À titre d'exemple, Aschbacher et Herman (1991) ont proposé un guide d'élaboration des rapports pour le domaine de la psychologie et de la communication. Leurs principes sont regroupés en cinq points: (a) connaitre le public et l'objectif visé; (b) garder le rapport simple; (c) être clair, précis, compréhensible et équilibré; (d) utiliser les techniques pour attirer l'attention des utilisateurs; et (e) adapter le format des rapports aux objectifs visés. Ryan (2006) a également suggéré des principes pour l'élaboration des rapports dans le cas de la province de la Caroline du Sud. En se basant sur les recherches empiriques, les auteurs tels que Goodman et Hambleton (2004), Hambleton et Slater (1997), Jaeger (1998), Wainer et al. (1999), Ryan (2006), Aschbacher et Herman (1991), Zenisky, Hambleton, Sireci (2009), Sinharay, Puhan et Haberman (2010), Roberts et Gierl (2010), et Klesch (2010) ont proposé un ensemble de principes afin d'assurer une bonne qualité des rapports que nous allons regrouper en 4 critères: **Accessibilité, Utilité, Lisibilité et Validité.** Le tableau 7 présente une synthèse de tous ces principes.

Tableau 7. Synthèse des critères et principes pour l'élaboration des rapports

Critères	Contenus	Références
Accessibilité	(a) capacité de comprendre le langage et la terminologie utilisée dans le rapport; (b) importance accordée au public visé; (c) variété des types de rapports aux différents publics; (e) personnalisation des rapports; (f) mécanisme de production des rapports selon les types d'informations, l'objectif et le public	Klesch (2010); Hambleton (2002); Goodman et Hambleton (2004); Hambleton et Slater (1997); Ryan (2006; Zenisky, Hambleton, Sireci (2009)
Utilité	(a) large éventail d'informations; (b) accent mis sur les informations pertinentes; (c) guide d'interprétation des concepts importants; (d) explication des niveaux de maitrise (élémentaire, intermédiaire, avancé); (e) explication des objectifs des tests, des niveaux d'attentes; (f) description des habiletés et des connaissances acquises et non acquises des élèves; (g) des informations diagnostiques sous forme des résultats des sous-habiletés	Klesch (2010); Hambleton (2002); Goodman et Hambleton (2004); Hambleton et Slater (1997); Ryan (2006); Zenisky, Hambleton, Sireci (2009)
Lisibilité	(a) format lisible, facile à comprendre; (b) absence de jargons statistiques et verbaux; (c) présentation claire, simple et pas encombrée; (d) variété des formes de présentation avec les légendes; (d) utilisation des couleurs selon le niveau de la performance	Klesch (2010); Hambleton (2002); Goodman et Hambleton (2004); Hambleton et Slater (1997); Ryan (2006); Zenisky, Hambleton, Sireci (2009)
Validité	(a) validation des rapports afin d'identifier les points forts et les faiblesses; (b) sous-scores devraient fournir des informations nécessaires autres que les scores totaux des tests; (c) sous-scores doivent être comparables entre les individus ou les institutions pour pouvoir faire des comparaisons; (d) présentation des résultats en comparaison avec différents groupes; (e) fournir des informations sur les erreurs de mesure de la performance des élèves; (f) ne présenter que des informations diagnostiques qui sont fiables	Wainer, Hambleton et Meara (1999); Godman et Hambleton (2004); Sinharay, Puhan et Haberman (2010); Ryan (2006); Roberts et Gierl (2010); Klesh (2010)

Lors de l'élaboration des rapports, un premier aspect important à considérer est le mode de présentation des résultats. Dans un rapport, les informations présentées peuvent être sous formes textuelles ou graphiques. Roberts et Gierl (2010) ont d'ailleurs proposé une revue de littérature sur les principes de présentations des informations textuelles et graphiques où l'attention devrait être accordée à : (a) la structure interne du texte; (b) la structure externe du texte. La structure interne du texte comprend des techniques d'organisation et de séquences permettant de fournir un cadre de référence interne du texte pour comprendre son contenu. Par exemple, un paragraphe doit contenir au moins une phrase qui exprime son idée-clé et une autre phrase pour conclure le problème annoncé. La structure externe comprend des techniques comme l'accessibilité de la structure, la typographie et la disposition des textes.

L'accessibilité de la structure du texte sera plus facile s'il y a une table des matières, un glossaire, des objectifs et un sommaire (Petterson, 2002; Roberts et Gierl, 2010). En ce qui concerne la *typographie*, les différents types de caractères (italiques, gras) et des couleurs différentes peuvent être utilisés afin d'attirer l'attention des utilisateurs sur les mots ou idées importantes et afin de créer le contraste (Horton, 1991). L'utilisation de la disposition horizontale ou verticale du texte permet de renforcer la structure hiérarchique du document au niveau visuel (Roberts et Gierl, 2010).

Lors de la communication des informations quantitatives comme les scores, les moyennes, les percentiles, les erreurs de mesure sous forme de tableaux et de graphiques, il faut respecter les quatre principes principaux suivants : (a) utiliser le contraste pour exprimer les informations importantes; (b) utiliser les redondances des indices visuels pour accentuer les informations présentées (couleurs différentes); (c) utiliser la proximité pour regrouper les éléments similaires ensemble; et (d) utiliser un alignement commun pour marquer la structure visuelle des informations (Cleveland et Mcgill, 1985; Koslyn, 1994; Roberts et Gierl, 2010). Le choix de présenter les informations sous forme d'un tableau ou d'un graphique dépend des objectifs de la présentation des données et du public visé. L'utilisation d'un tableau est recommandée lorsque nous présentons une petite quantité de données et qu'il faut fournir des chiffres exacts pour les comparaisons (Tufte, 1996) tandis que les graphiques peuvent mieux communiquer les changements et les tendances des données (Shah, Mayer et Hegarty, 1999). Wainer (1997) et Roberts et

Gierl (2010) ont cité les principes suivants lors de la création d'un tableau : (a) arrondir les chiffres à 2 décimales; (b) utiliser la moyenne des lignes et des colonnes pour fournir un focus visuel et un sommaire; (c) utiliser les lignes plutôt que les colonnes afin de faciliter les comparaisons; (d) classer les lignes et les colonnes de manière appropriée; et (e) utiliser les espaces blancs pour regrouper les images et diriger les yeux. Le graphique et le texte doivent être intégrés ensemble dans une même page afin de faciliter l'interprétation de l'information (Roberts et Gierl, 2010). La figure 7 résume les éléments à suivre lors de la présentation des informations textuelles et non textuelles.

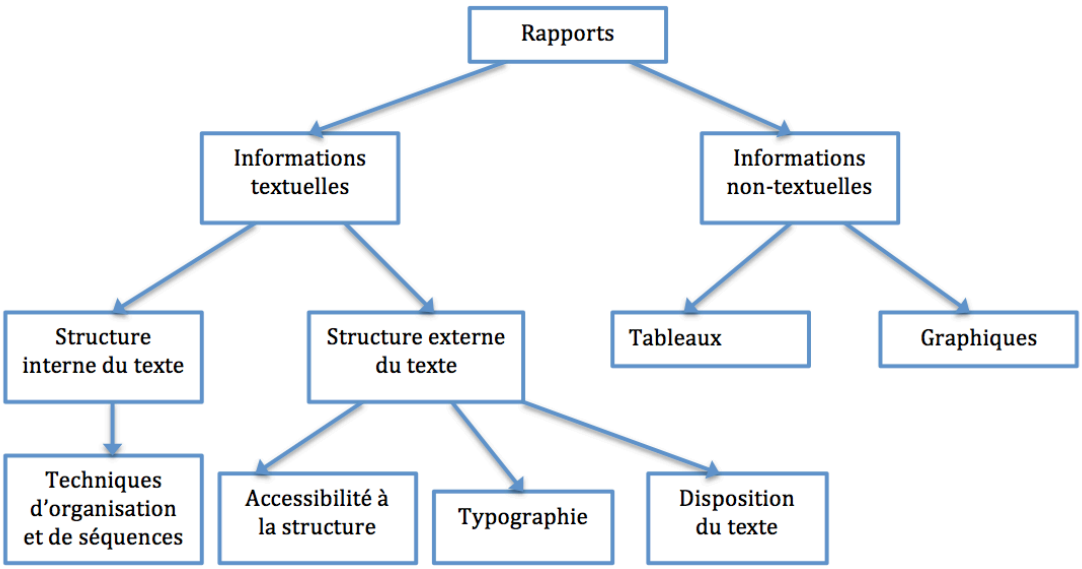


Figure 7. Synthèse des principes de la présentation des informations dans un rapport (synthétisé de Roberts et Gierl, 2010).

Roberts et Gierl (2010) ont proposé un cadre de référence pour l'élaboration de rapports diagnostiques avec la méthode AHM en sachant que des éléments proposés dans ce cadre peuvent être transférables à d'autres MCD. Les auteurs ont posé deux questions principales : *Quelles informations obtenues de la méthode AHM peuvent être présentées dans un rapport ? Comment présenter ces informations diagnostiques d'une manière*

efficace? Ainsi, la première question porte sur les éléments de fond des rapports tandis que la deuxième porte sur la forme des rapports diagnostiques.

En ce qui concerne le contenu des rapports diagnostiques, ils doivent contenir les principaux éléments suivants: (a) les objectifs du rapport; (b) une description des habiletés évaluées; (c) la performance pour chaque habileté; (d) un résumé de la maitrise des habiletés; et (e) un guide d'interprétation des résultats. Les rapports sont organisés en trois parties: la première partie consiste à fournir des informations sur les scores totaux obtenus ainsi que les directives pour lire le rapport. Les informations normatives peuvent être accompagnées des scores totaux permettant aux utilisateurs d'obtenir des informations plus claires de leurs résultats. Ces informations semblent également pertinentes dans le contexte diagnostique cognitif parce que des apprenants, bien qu'ils obtiennent les mêmes scores, peuvent obtenir des profils de maitrise des habiletés différents. Les directives pour lire les rapports et un rappel sur le guide d'interprétation doivent fournir un appui nécessaire pour comprendre le rapport (Roberts et Gierl, 2010).

La deuxième partie contient des informations diagnostiques spécifiques telles que les scores diagnostiques ainsi que le niveau de maitrise des attributs. Les informations spécifiques sur la maitrise de chaque attribut peuvent accompagner des réponses des candidats aux items, les réponses correctes ainsi qu'un sommaire des réponses. Le niveau de maitrise de chaque habileté peut être présenté sous forme de graphiques selon deux catégories dichotomiques (maitrise ou non-maitrise); trois catégories (non-maitrise; maitrise partielle ou maitrise) ou encore des probabilités.

La dernière partie fournit un résumé sur la performance des élèves sur toutes les habiletés. Ces informations textuelles permettent de comprendre l'essentiel du rapport sans une lecture complète ou dans le cas où l'utilisateur a des difficultés avec les graphiques. Les forces et les faiblesses cognitives ainsi que les recommandations pour améliorer les faiblesses peuvent être communiquées directement aux candidats dans cette partie (Roberts et Gierl, 2010).

Le guide d'interprétation est un élément important du rapport diagnostique qui est bien souvent présenté dans une page séparée. Ce guide doit contenir les aspects suivants : (a) une description très détaillée des habiletés évaluées; (b) les définitions des attributs et comment ils sont reliés au test; comment les scores diagnostiques ont été obtenus et (c)

les directives pour interpréter les résultats du rapport. Ces directives peuvent être présentées sous forme de questions et de réponses (Roberts et Gierl, 2010).

Quant à la forme des rapports, les principes sont similaires à ceux que nous avons présentés précédemment pour les informations textuelles et non textuelles comme les graphiques et les tableaux. Parmi les trois parties du rapport développé par Roberts et Gierl (2010), la première partie et la troisième partie du rapport contiennent des informations dans un format textuel, tandis que des graphiques en barres sont utilisés pour la deuxième partie portant sur le niveau de maitrise de chaque habileté. Afin de faciliter la visualisation, les différentes parties devraient être bien séparées l'une de l'autre. Le tableau 8 propose une synthèse des éléments à retenir pour l'élaboration des rapports avec la méthode AHM tant sur la forme que sur le fond. Étant donné que la plupart des éléments sont applicables pour les autres MCD, ces éléments sont très utiles pour présenter des résultats diagnostiques issus des modèles DINA et G-DINA.

En résumé, dans cette partie, nous avons présenté une revue de littérature sur les principes à respecter lors de l'élaboration des rapports. Ces principes sont regroupés en quatre critères principaux : l'accessibilité, la lisibilité, l'utilité et la validité. Nous avons également présenté les principes plus spécifiques lors de la présentation des informations textuelles et non textuelles dans un rapport et nous les avons transférés au contexte d'un rapport diagnostique en analysant un exemple de rapports développés avec la méthode AHM (Roberts et Gierl, 2010). Toutefois, le processus de développement des rapports est complété seulement si ces rapports sont évalués auprès du public visé. Ainsi, la prochaine partie est réservée à décrire les méthodes utilisées dans les recherches empiriques sur l'évaluation de ces rapports.

2.3.2.4. Évaluer des rapports

Les recherches portant sur l'élaboration et l'évaluation des rapports indiquent que les éducateurs rencontrent des difficultés dans la compréhension des termes statistiques et des graphiques utilisés pour communiquer les résultats dans les rapports (Hambleton et Slater, 1997; Zapata-Rivera, VanWinkle, et Zwick, 2010; Zapata-Rivera, Vezzu et VanWinkle, 2013; Zwick et al., 2008). Face à ce problème, de nombreuses stratégies ont été suggérées, dont l'évaluation de ces rapports par des experts et des membres du public

visé afin de détecter les forces et les faiblesses de ces rapports puis de les réviser avant de les publier en grand nombre (Zapata-Rivera, Vezzu et VanWinkle, 2013; VanWinkle, Vezzu et Zapata-Rivera, 2011; Vezzu, VanWinkle et Zapata-Rivera, 2012). Dans l'ensemble, l'intérêt des recherches en évaluation des rapports porte particulièrement sur: (a) les perceptions des enseignants sur la représentation textuelle et des graphiques et (b) la compréhension ou l'interprétation des résultats. Nous allons proposer une synthèse de ces recherches dans les lignes suivantes.

Tableau 8. Cadre de référence d'élaboration des rapports diagnostiques avec la méthode AHM
(adapté de Roberts et Gierl, 2010)

Caractéristiques des rapports	**Éléments obtenus avec AHM**
Forme des résultats présentés	
Échelle	Probabilités de maitrise des attributs; scores totaux
Référence pour interprétation	Modèle cognitif; critères de référence
Unité d'évaluation	Niveau d'attribut; niveau de modèle cognitif
Public visé	Élèves; parents; enseignants
Erreur de mesure	Fiabilité d'attributs
Mode de présentation des résultats	
Numérique	Probabilité et fiabilité des attributs
Graphique	Probabilité des attributs; classification des profils; modèle cognitif
Narratif	Probabilité des attributs; classification des profils; modèle cognitif; résumé de la performance
Modes de présentation	Papier, interactif
Application des principes du design	Contraste; redondance; proximité; structure du texte; représentation des données quantitatives

Dans le cadre du projet CBAL, Zapata-Rivera et ses collaborateurs (2012, 2013, 2014) ont mené plusieurs recherches en élaboration et évaluation des rapports en mathématique interactifs en ligne. Plus précisément, Zwick, Zapata-Rivera et Hegarty (2014) ont validé les rapports auprès 148 enseignants et 98 étudiants en psychologie. Leur objectif principal vise à déterminer: (a) les formats de représentations graphiques et verbales qui facilitent la compréhension; (b) les méthodes de représentations préférées des enseignants et des étudiants et le lien entre leur compréhension et leur préférence; (c) les facteurs sociodémographiques qui influencent la compréhension et la préférence des rapports et (d) les mauvaises conceptions qui caractérisent leurs réponses sur l'erreur de mesure. Quatre versions alternatives de rapports ont été développées qui correspondent à quatre conditions expérimentales de la recherche.

Les informations sur ces rapports ont été recueillies avec trois questionnaires: un questionnaire sur la compréhension des rapports, un sur la préférence des enseignants et des étudiants et un dernier sur leur « background ». Le questionnaire sur la compréhension se compose de questions à choix multiple et de questions ouvertes portant sur la compréhension et les explications des participants sur les représentations des scores des candidats. À titre d'exemple, déterminer le vrai score du candidat parmi les trois choix proposés. Le questionnaire de préférence porte sur l'évaluation de l'utilité, de la compréhensibilité des représentations des rapports et les suggestions pour les améliorer. Par exemple, les participants sont invités à choisir un format de représentations graphiques qui leur semble le plus facile à interpréter parmi les deux formats proposés. Cette question est suivie par une question ouverte sur les raisons qui justifient leur choix (Zwick, Zapata-Rivera et Hegarty, 2014). Finalement, le questionnaire sociodémographique ainsi que sur les connaissances en mesure, en statistiques et en interprétation des rapports permettant de vérifier l'interaction entre les connaissances antérieures et les caractéristiques de représentations des rapports (Zwick, Zapata-Rivera et Hegarty, 2014).

Des analyses d'ANOVA à deux facteurs ont été réalisées pour vérifier les différences entre les scores de la compréhension des rapports et les variables sociodémographiques et les quatre conditions expérimentales ainsi que leurs interactions. Les résultats suggèrent toutefois qu'il n'y a pas de différence significative entre ces

conditions. Une des suggestions consiste à améliorer le texte et les graphiques et à les combiner afin de faciliter la compréhension. Quant aux relations entre la compréhension, la familiarité avec les statistiques et le nombre d'années d'expérience des enseignants, les résultats obtenus ne sont pas statistiquement significatifs. Les résultats des analyses d'ANOVA à deux facteurs ne suggèrent aucune différence statistiquement significative entre les scores de compréhension et les préférences. Autrement dit, les scores de la compréhension des rapports ne varient pas selon les formats de rapports de préférence.

Une méthode similaire a été appliquée dans une autre recherche de Zapata-Rivera, Vezzu et VanWinkle (2013) auprès de 125 enseignants au secondaire. Toutefois, cette fois-ci, au lieu de proposer quatre conditions expérimentales, trois formats de rapports ont été élaborés: le premier rapport utilise une boite à moustaches, la deuxième comporte des icônes empilées et le dernier rapport est sous forme d'une courbe de distribution des scores. Comme dans la recherche précédente, trois questionnaires portant sur la compréhension, la préférence et les informations sociodémographiques destinés aux enseignants ont été administrés en ligne. Ces questionnaires ont été validés auprès de six enseignants au secondaire afin de vérifier l'accessibilité, la lisibilité et les problèmes de navigation. Il existe des différences entre les versions des questionnaires de cette recherche par rapport à la précédente. À titre d'exemple, le questionnaire sur la compréhension comprend 14 QCM et pas de question ouverte. Elles ont été développées en se basant sur les aspects suivants: les percentiles, la moyenne, la valeur la plus petite et celle la plus grande. Quant au questionnaire sur la préférence, à part des questions sur le format de représentation préféré et sur celui qui facilite le plus la compréhension, les questions portent également sur les perceptions des versions que les autres enseignants vont comprendre le mieux.

Toutefois, à la différence de la recherche précédente, les analyses d'ANOVA montrent qu'il y a une différence statistiquement significative entre les scores de compréhension et les trois formats de rapports. Plus précisément, la différence de compréhension se trouve entre la boite à moustaches et les icônes empilées et entre la boite à moustaches et la courbe de distribution des scores. Les analyses des difficultés des items pour chaque format de rapport ont été réalisées, cependant les résultats suggèrent que les enseignants trouvent plus difficile d'interpréter des boites à moustaches. C'est

pour cette raison que ce format de rapport a été le moins choisi (11,4%) par les enseignants en termes de préférence en comparaison avec les icônes empilées (39,8%) et la courbe de la distribution des scores (48,9%).

Dans les deux autres recherches de Zapata-Rivera et VanWinkle (2010) et Zapata-Rivera, VanWinkle et Zwick (2012), l'objectif principal vise à vérifier l'interprétation des résultats de rapports interactifs en ligne. Plus précisément, dans la première recherche qualitative sur l'utilité des rapports en mathématique du CBAL auprès 12 enseignants de la 8^{e} à 12^{e} année des écoles des provinces de New Jersey et Pennsylvanie, les résultats montrent que, dans l'ensemble, les enseignants interagissent d'une manière positive avec les rapports en ligne. Quant à la deuxième recherche quantitative auprès de 147 enseignants, trois rapports types du CBAL ont été développés avec deux conditions (avec ou sans tutoriel). Après avoir examiné les rapports, ces enseignants sont invités à remplir un questionnaire sur la compréhension des informations présentées dans les rapports. Malheureusement, le contenu de ces instruments n'a pas été présenté en détail dans leurs rapports de recherche. Bien que les scores totaux ne varient pas statistiquement entre les deux conditions expérimentales (avec ou sans tutoriels), les réponses aux différentes questions montrent que 94 % des enseignants peuvent reconnaitre l'utilité de chaque type de rapports. Toutefois, une proportion significative d'enseignants considère que certains objectifs proposés dans les rapports ne sont pas valides. Cette étude montre également que les enseignants ont des difficultés à comprendre des concepts statistiques fondamentaux comme : la fiabilité (43%); les percentiles (50%); les vrais scores (50%); et les échelles de mesure (42%) (Zapata-Rivera, VanWinkle et Zwick, 2012). Ces résultats semblent donc cohérents avec les constats proposés dans la recherche de Goodman et Hambleton (2004). Des suggestions sur la révision et l'évaluation des rapports ont été proposées, surtout au niveau de l'amélioration des informations sur les statistiques; de la limitation d'utiliser des termes techniques et des explorations pour utiliser des formats de représentations graphiques alternatives. Ces recherches confirment donc l'importance d'adapter la conception des rapports selon les différents besoins des acteurs du milieu et la nécessité d'évaluer ces rapports auprès des publics visés (Hambleton et Slater, 1997; Wainer, Hambleton et Meara, 1999; Zapata-Rivera et Zwick, 2011).

Si les recherches de Zapata-Rivera et ses collaborateurs s'intéressent à élaborer et évaluer les rapports interactifs en mathématique dans le cadre du programme CBAL, l'étude de Zenisky, Hambleton et Sireci (2009) vise à concevoir et évaluer les rapports pour présenter les résultats du NAEP. Bien que cette recherche vise les mêmes objectifs que celles précédentes, la méthode adoptée est différente car elle met l'accent sur les interactions directes entre les participants, les chercheurs et les formats de représentation des rapports. Ainsi, les objectifs spécifiques de cette recherche visent à examiner à quel niveau les acteurs locaux et provinciaux sont familiers avec les méthodes courantes de la représentation des résultats du NAEP, les types d'inférences qu'ils peuvent faire à partir de ces représentations et les suggestions pour améliorer le design.

Des rapports en mathématique et en lecture avec les différents formats de représentation des données ont été développés et présentés à deux groupes de discussion, dont un groupe pour évaluer les rapports en mathématique et l'autre pour ceux en lecture. Lorsque chaque format de représentation est projeté à l'écran, les participants sont invités à réfléchir sur ce format pendant quelques minutes et à répondre ensuite aux questions sur les représentations posées par un intervenant de l'entrevue. Les questions varient, de la nature des informations (les scores moyens des élèves en 8e année en mathématique en 2005, par exemple) à leurs perceptions sur les représentations (par exemple, les éléments qui causent la confusion et qui ne sont pas clairs dans la représentation). Les résultats obtenus grâce à ces deux groupes de discussion montrent que les participants rencontrent des difficultés dans l'interprétation des échelles des scores sans les informations contextuelles supplémentaires. Les participants s'intéressent à savoir comment les niveaux de maitrise présentés dans le NAEP correspondent à des catégories de niveau utilisées dans chaque province. Plusieurs participants pensent que les rapports seront plus clairs si les différences de la performance des sous-groupes (genre et ethnie, par exemple) sont illustrées dans les représentations. En outre, ces acteurs du milieu rencontrent encore des difficultés dans l'interprétation des rapports de NAEP même s'ils ont suivi des formations en méthodes quantitatives. Après cette recherche, deux types de connaissances ont été suggérées comme nécessaires afin de faciliter la compréhension des acteurs du milieu: (a) une grande familiarité avec les scores du test et les jargons statistiques (erreur standard, échelle des scores, etc.) et (b) la familiarité avec les termes

communs de NAEP et les mécanismes de présentation des résultats (des niveaux de maitrise, des outils interactifs en ligne) (Zenisky, Hambleton et Sireci, 2009).

S'inscrivant dans la même perspective, la recherche de Klech (2010) vise à détecter des concepts qui sont difficiles à comprendre dans les rapports. Pour ce faire, trois rapports types ont été conçus et distribués à trois groupes de participants: 16 pédagogues; 6 professionnels provenant des programmes d'évaluation et 6 étudiants au doctorat. Les participants ont examiné ces rapports et répondu à un questionnaire sur les informations présentées dans le rapport dans un groupe de discussion. Le questionnaire comprend une dizaine de questions sur les informations qui figurent sur les rapports comme la date de passation du test, le statut du passage, les objectifs de l'apprentissage, mais aussi sur la compréhension des résultats de la performance. Plus précisément, les questions sur la compréhension du rapport A portent sur un diagramme en cylindre et la performance à chaque objectif d'apprentissage. Pour le rapport B, des questions spécifiques sur la performance des candidats en lien avec la médiane et le pourcentage des réponses correctes du candidat. Pour le rapport C, les questions tournent autour de l'intervalle de confiance et son interprétation ainsi que la médiane et le pourcentage des réponses correctes du candidat (Klesch, 2010). Les résultats suggèrent que le groupe de professeurs et le groupe des professionnels ont les mêmes constats face à ces rapports : (a) il n'y a pas de connexion claire entre les scores bruts des performances et les échelles présentés; et (b) il faut diriger les candidats vers les matériels et les sources disponibles sur les programmes d'évaluation et le contenu présenté dans le test. Suite à ces résultats, la recherche a proposé un certain nombre de recommandations pour l'élaboration et l'évaluation des rapports dont certains principes ont été présentés dans le tableau 6 portant sur les critères à respecter lors de l'élaboration des rapports.

À la différence des recherches précédentes qui prêtent l'attention aux représentations graphiques ou à la compréhension des rapports, l'étude de Impara et ses collaborateurs (1991) met l'accent sur l'importance des informations explicatives dans la compréhension des résultats par les enseignants. Ainsi, le rapport développé se divise en deux parties qui correspondent à deux conditions expérimentales. La première moitié du rapport est accompagnée des informations interprétatives tandis que l'autre moitié du rapport ne les contient pas. Le questionnaire a été administré à 369 enseignants ayant de 1

à 34 années d'expérience. L'instrument contient 17 questions sur l'évaluation des rapports et 7 questions sociodémographiques. Parmi les 17 questions sur le rapport, il y a 5 questions concernant la nomenclature du rapport comme la signification des rangs percentiles nationaux. Cinq autres questions demandent que les enseignants localisent des informations spécifiques sur le rapport et les 7 questions qui restent portent sur l'interprétation des résultats du test.

Des analyses d'ANOVA à trois facteurs ont été effectuées entre les scores de la compréhension et les variables indépendantes comme la présence ou l'absence des guides d'interprétation, les cours antérieurs en mesure et le grade. Les différences statistiquement significatives se trouvent entre la présence ou l'absence des guides d'interprétation et entre la participation à des classes en mesure et non. Les résultats suggèrent également que les guides d'interprétation aident les enseignants à mieux répondre aux questions de compréhension. La preuve est que seulement 11% des enseignants ne répondent pas correctement à un item avec le guide alors que 49,5% des enseignants n'arrivent pas à y répondre sans le guide (Impara et al., 1991). Cette recherche confirme donc la nécessité de fournir des guides d'interprétation aux enseignants.

Dans une visée diagnostique, jusqu'à présent, la recherche de Jang (2005) est la seule qui a validé les rapports diagnostiques conçus auprès des enseignants et des apprenants afin de vérifier l'utilité des informations diagnostiques reçues à travers les rapports pour l'enseignement et l'apprentissage. Ces informations sont obtenues grâce à deux enquêtes avec des questions ouvertes. Un questionnaire a été administré au début de la session d'une classe préparatoire au test TOEFL après que les candidats aient reçu les rapports diagnostiques sur le test d'entrée. Les questions portent sur (a) les informations les plus utiles et les moins utiles dans le rapport; (b) ce que les candidats pensent de leur maitrise des habiletés et leur degré d'accord avec ces résultats; (c) la compréhensibilité des descripteurs des habiletés et (d) les suggestions pour améliorer les rapports. Le deuxième questionnaire est distribué aux participants à la fin de la session afin d'examiner les activités que les candidats ont faites en dehors de la classe afin d'améliorer leurs faiblesses en lecture et les habiletés qu'ils veulent améliorer.

Tous les enseignants ayant participé à l'étude constatent que les rétroactions diagnostiques sont utiles pour la reconnaissance des forces et des faiblesses cognitives

des étudiants, ce qui confirme la possibilité d'appliquer les MCD aux tests en langue à grande échelle qui existent afin de fournir des informations diagnostiques via des rapports détaillés (Jang, 2005; Li, 2012; Li et Suen, 2011; Lee et Sawaki, 2009). Toutefois, la recherche met plus d'accent sur l'utilité des informations diagnostiques à l'amélioration des habiletés plutôt que sur les évaluations relatives aux caractères spécifiques du design du rapport, ce qui constitue une des limites de cette recherche.

En guise de résumé, nous avons présenté une synthèse sur les recherches empiriques en évaluation des rapports auprès des enseignants en mettant l'accent sur les instruments de collecte des informations. Dans l'ensemble, les intérêts des recherches pointent dans deux directions principales: (1) les perceptions des enseignants sur des formats de représentations graphiques des rapports et (2) la compréhension et l'interprétation des résultats. Parmi les huit recherches recensées, quatre ont adopté la méthode quantitative avec des questionnaires comprenant des questions à choix multiples et des questions ouvertes dont le contenu se trouve dans le tableau 9. Les quatre autres recherches ont été réalisées avec une méthode qualitative consistant en un panel d'experts. Ainsi, nous discuterons les forces et les faiblesses des recherches réalisées et proposerons des suggestions pour notre propre recherche dans la partie suivante.

2.3.3. Synthèse

La section 2.3 vise à fournir des balises théoriques et empiriques sur l'élaboration et l'évaluation des rapports. En nous basant principalement sur les cadres de référence de Zapata-Rivera, VanWinkle et Zwick (2012) et de Hambleton et Zinsky (2010), nous avons décrit les quatre étapes principales du processus d'élaboration et d'évaluation des rapports: (a) identifier le public visé et analyser ses besoins; (b) documenter les recherches et les formats de rapports qui existent; (c) développer des rapports et (d) évaluer des rapports. Ces cadres de référence vont au-delà des recherches précédentes de Tuffe (1983, 1996); Fast (2002); Hambleton et Slatter (1997); Hatie (2009) et Wainer et al., (1997, 2005) parce qu'ils mettent l'accent sur le processus de la conception du rapport au lieu de présenter simplement les directives ou règlements qui guident le design d'un rapport. En d'autres termes, ces cadres de référence nous suggèrent des activités spécifiques qui nous conduisent à une conception plus cohérente et efficace des rapports (Zapata-Rivera et Kartz, 2014).

Tableau 9. Synthèse de la méthodologie des recherches en évaluation des rapports

Recherches	Méthode	N	Instruments	Dimensions
Zwick, Zapata-Rivera et Hegraty (2014)	Quantitative	148 enseignants et 98 étudiants en psychologie	-quatre formats de rapports -trois questionnaires sur la compréhension, la préférence et les questions sociodémographiques	-représentations graphiques et verbales -méthodes de représentations préférées -facteurs liés à la compréhension et la préférence -fausses conceptions liées à l'erreur de mesure
Zapata-Rivera, Vezzu et VanWinkle (2013)	Quantitative	125 enseignants au secondaire	-trois formats de rapports -trois questionnaires sur la compréhension, la préférence et les questions sociodémographiques	-efficacité de trois formats de représentation des rapports: la boite à moustaches, des icônes empilées et la courbe de la distribution des scores
Zapata-Rivera et VanWinkle (2010)	Qualitative	12 enseignants de 8e à 12e année NJ et PA	-questions sur l'interaction avec les rapports	-utilité des rapports interactifs
Zapata-Rivera, VanWinkle et Zwick (2012)	Quantitative	147 enseignants Niveau ?	-trois rapports développés avec ou sans tutoriel -questionnaire sur la compréhension	- interprétations des résultats des rapports
Zenisky, Hambleton et Sireci (2009)	Qualitative	Deux groupes d'éducateurs Niveau ?	- rapports en mathématique et en lecture élaborés avec des formats de représentation différents -questions d'entrevue	-familiarité avec les méthodes courantes de la représentation des résultats du NAEP -types d'inférences faites à partir de ces représentations -suggestions pour améliorer le design
Klesch (2010)	Qualitative	16 professeurs; 6 professionnels des programmes d'évaluation et 6 étudiants au doctorat.	-trois rapports élaborés -une dizaine de questions d'entrevue ouvertes dans un groupe de discussion	-informations sur les rapports: date de passation; objectifs d'apprentissage; le statut du passage -résultats des performances: diagramme en cylindre; la médiane et les pourcentages de réponses correctes
Impara et al., (1991)	Quantitative	369 enseignants Niveau ?	-questionnaires comprenant 17 items et 7 items descriptifs	-nomenclature du rapport -informations spécifiques du rapport -interprétation des résultats
Jang (2005)	Qualitative	27 étudiants d'une classe préparatoire du test TOEFL	- questions d'entrevues ouvertes	-utilité des informations diagnostiques; la maitrise des habiletés; la compréhensibilité des descripteurs et des suggestions pour améliorer les rapports; les activités pour améliorer leurs habiletés en lecture

L'accent est mis sur les étapes du développement des rapports avec les quatre critères principaux: *l'accessibilité, la lisibilité, l'utilité et la validité* et l'étape de l'évaluation des rapports avec les éléments méthodologiques, à savoir la méthode adoptée, les conditions expérimentales, les instruments de collecte des données ainsi que les dimensions de ces instruments. Dans l'ensemble, les dimensions de ces instruments peuvent être regroupées en (a) la compréhension et l'interprétation des représentations graphiques; (b) les préférences des formats de représentations graphiques; (c) l'efficacité et l'utilité des représentations; (d) le contenu spécifique des rapports; (e) les suggestions pour améliorer le design des rapports et (f) les informations sociodémographiques.

En général, les résultats des recherches en évaluation des rapports suggèrent que les enseignants rencontrent encore des difficultés dans l'interprétation des notions statistiques telles que la fiabilité, les percentiles, les vrais scores et les échelles de mesure (Zapata-Rivera, VanWinkle et Zwick, 2012; Zenisky, Hambleton et Sireci, 2009; Hambleton et Slater, 1997; Koretz et Diebert, 1993). De plus, la compréhension des rapports dépend de nombreux facteurs tels que les formats de représentation (Zapata-Rivera, Vezzu et VanWinkle, 2013) et la présence ou l'absence des guides d'interprétation permettant de faire des inférences (Impara et al., 1991; Zenisky, Hambleton et Sireci, 2009) ainsi que la familiarité avec les cours en statistiques (Hambleton et Slater, 1997; Impara, Divine et al., 1991). En effet, les enseignants rencontrent plus de difficultés lors de l'interprétation de diagrammes de boites à moustaches que les icônes empilées ou la courbe de distribution des scores (Zapata-Rivera, Vezzu et VanWinkle, 2013) et des rapports sans guide d'interprétation ou informations supplémentaires sur le contexte (Impara et al., 1991; Zenisky, Hambleton et Sireci, 2009). En outre, les participants ayant une plus grande familiarité avec les statistiques ont tendance à avoir une plus grande compréhension des scores (Zwick, Zapata-Rivera et Hegarty, 2014). Ces recherches pointent vers un besoin spécifique des enseignants relativement à des informations additionnelles sur les concepts de base en statistique afin de comprendre le contenu présenté dans les rapports à travers des guides d'interprétation (Zapata-Rivera, VanWinkle et Zwick, 2010; Zwick et al., 2008). Elles ont confirmé la nécessité de valider des rapports auprès du public visé afin de les rendre compréhensibles et accessibles. Ceci constitue une des forces majeures des recherches en évaluation des rapports.

Une autre force de ces recherches réside dans les formats des rapports proposés et des instruments de collecte de données, notamment avec les rapports interactifs et les questionnaires d'évaluation administrés en ligne. En effet, avec les rapports interactifs, les candidats peuvent posséder la flexibilité d'avoir plus d'options lors de l'assemblage des rapports et de les ajuster au besoin (Klesch, 2010). En ce qui concerne les questionnaires en ligne, les participants peuvent avoir des interactions directes et spontanées avec le format de rapport à évaluer selon la question posée. Ceci constitue un avantage important qui facilitera le processus de collecte des informations et les résultats obtenus. Quant au format des questions posées dans les recherches quantitatives, à part des questions à choix multiples, le fait de poser quelques questions ouvertes sur des rapports permet d'obtenir des informations plus larges sur les caractéristiques souhaitées par les participants ainsi que sur des explications à certains termes techniques (Zwick, Zapata-Rivera et Hegarty, 2014). Toutefois, le fait de laisser plus de liberté dans les réponses cause un problème lors du codage, étant donné qu'une réponse peut faire partie de plus d'une catégorie (Zwick, Zapata-Rivera et Hegarty, 2014), ce qui constitue une des faiblesses des recherches qualitatives. En outre, bien que les rencontres en groupe de discussion puissent stimuler les discussions entre les participants et les intervenants sur les différents formats de représentation des graphiques et les amener à répondre aux questions d'interprétation d'une manière plus collaborative (Zenisky, Hambleton et Sireci, 2009), elles peuvent engendrer certains biais dans les réponses à cause de l'effet de l'ordre de présentation des formats (Klesch, 2010). À titre d'exemple, avec 3 formats de rapports élaborés, si le format A est toujours présenté en premier lieu pour être évalué, les participants vont avoir la tendance à le juger meilleur ou vice versa (Klesch, 2010). À notre avis, cet élément peut être évité avec des questionnaires administrés en ligne où l'ordre de l'affichage des rapports à évaluer peut être réglé aléatoirement.

Une autre a faiblesse des recherches se trouve dans le contenu des rapports à évaluer. Les recherches prêtent plus d'attention à évaluer les représentations graphiques que les parties textuelles telles que la définition des objectifs d'apprentissage et les guides d'interprétation. Malgré l'importance des guides d'interprétation dans la compréhension des résultats (Impara et al., 1991; Roberts et Gierl, 2010), cet aspect a été abordé

seulement dans une des huit recherches présentées, ce qui constitue un élément important pour notre recherche.

Lors de l'élaboration et l'évaluation des rapports diagnostiques, il faut prendre en considération des caractères liés aux objectifs visés, et à la nature des résultats obtenus des modélisations. Ainsi, le contenu des rapports diagnostiques devrait contenir essentiellement (a) les objectifs du rapport ; (b) une description des habiletés évaluées; (c) la performance pour chaque habileté; (d) un résumé de la maitrise des habiletés; et (e) un guide d'interprétation des résultats. L'évaluation des rapports devrait couvrir nécessairement tous ces éléments tant textuels que des représentations graphiques. D'ailleurs, étant donné que les résultats communiqués ne sont pas des scores exacts, mais plutôt des niveaux de maitrise ou des probabilités et des profils de maitrise des habiletés, les questions posées sur la compréhension devraient porter sur ce degré de maitrise des habiletés ou les profils, mais pas sur les scores exacts tels quels.

Dans l'ADC, un constat général tiré à partir d'un nombre modeste des recherches sur le sujet porte sur le fait qu'elles s'intéressent soit à l'élaboration des rapports, soit à l'évaluation des rapports, mais pas à ces deux processus, alors qu'ils devraient être intégrés l'un à l'autre afin d'assurer la cohérence et la fiabilité du processus. En outre, si une seule recherche en évaluation des rapports diagnostiques existe, elle a pour objectif principal de vérifier l'utilité des informations diagnostiques dans l'enseignement et l'apprentissage en lecture, mais pas d'examiner les caractéristiques de la représentation des résultats. Ceci montre donc un déficit important des recherches en évaluation des rapports diagnostiques destinés aux enseignants. Ainsi, notre recherche vise à combler cette lacune, lorsque nous nous intéressons à élaborer des rapports diagnostiques destinés aux enseignants et à les faire évaluer par le public visé. Reste à voir comment cet objectif général peut être transféré en des objectifs et des questions spécifiques, ce qui fait l'objet de la partie suivante.

2.4. Objectifs et questions spécifiques de la recherche

Malgré la demande croissante des évaluations à grande échelle et la forte pression pour rendre ces évaluations les plus informatives possibles sur les forces et les faiblesses des élèves, la plupart des évaluations à grande échelle fournissent les résultats des élèves

seulement sous forme des scores globaux qui témoignent de leur performance. Ainsi, les résultats obtenus ne nous renseignent pas sur les informations spécifiques liées aux forces et faiblesses des élèves leur permettant d'évaluer et d'améliorer leurs habiletés cognitives et ne guident pas les enseignants dans la conception des interventions pédagogiques appropriées (Cui et al., 2012). C'est le cas du test PIRLS 2011, alors qu'il a été conçu avec deux objectifs et quatre processus de compréhension en lecture, il a donc le potentiel de fournir des informations diagnostiques plus détaillées sur les forces et les faiblesses des élèves que les scores totaux du test.

Ainsi, notre recherche vise, dans un premier temps, à modéliser les données du test PIRLS 2011 des élèves en 4^{e} année à visée diagnostique. Cet objectif général se compose de deux objectifs spécifiques qui consistent à: (a) vérifier la capacité diagnostique du test PIRLS 2011 et (b) fournir des profils de maitrise des habiletés aux élèves du Canada ayant participé au test PIRLS 2011. Nous avons donc posé les deux questions de recherche spécifiques suivantes:

1. Quelle est la qualité diagnostique des items du PIRLS 2011?

Les modélisations avec les modèles de DINA et G-DINA nous fournissent, dans un premier temps, les valeurs de six statistiques qui servent à évaluer les ajustements relatifs et absolus de ces modèles afin de déterminer celui qui s'ajuste mieux aux données. Par la suite, les résultats des modélisations nous indiquent les indices de pseudo-chance et d'étourderie de 35 items du PIRLS 2011 nous permettant de tirer des conclusions sur la qualité diagnostique de ces items.

2. Quels sont les profils de maitrise des habiletés des élèves canadiens ayant participé au PIRLS 2011?

Les résultats des modélisations nous renseignent sur des profils de maitrise et non-maitrise des habiletés de chaque élève sous forme des probabilités qui varient entre 0 et 1. Ils nous fournissent également des probabilités de maitrise de chaque habileté des élèves ainsi que celles de chaque classe latente. À partir de ces résultats, nous pouvons donc déterminer le profil de maitrise et non-maitrise des habiletés le plus courant de 4762 élèves au Canada ayant passé le PIRLS 2011 ainsi que l'habileté la plus ou la moins maitrisée.

Avec les profils de maitrise et de non-maitrise des habiletés obtenus à partir des modélisations, la finalité de l'ADC est de fournir aux apprenants des rétroactions formatives sur leurs forces et leurs faiblesses à travers des rapports diagnostiques compréhensibles et interprétables (Roberts et Gierl, 2010). À la différence des rapports issus des tests à grande échelle qui communiquent seulement les scores totaux ou des rangs percentiles, le défi des rapports diagnostiques réside dans l'intégration des informations substantives et techniques correspondant aux besoins du public visé et des informations sophistiquées provenant des modélisations de l'approche diagnostique cognitive (Roberts et Gierl, 2010). Cependant, la revue de littérature sur le sujet montre qu'il existe très peu de recherches empiriques sur l'élaboration et l'évaluation de ces rapports diagnostiques.

À partir de ces constats, le deuxième objectif général de notre recherche consiste à élaborer et à évaluer des rapports diagnostiques issus des modélisations destinées aux enseignants. Ainsi, cet objectif général peut se transférer en deux objectifs spécifiques qui visent à : (a) élaborer des rapports à visée diagnostique destinés aux enseignants et (b) évaluer des rapports diagnostiques auprès du public visé. Afin d'atteindre ces deux objectifs spécifiques, nous avons posé les deux questions spécifiques suivantes:

3. Comment élaborer des rapports à visée diagnostique destinés aux enseignants à partir des résultats obtenus des modélisations?

Les éléments de réponse à cette question sont obtenus à partir du travail d'un panel d'experts que nous avons recruté de différents milieux éducatifs pour l'élaboration des rapports diagnostiques. Ainsi, les discussions avec les membres du panel d'experts nous permettent de déterminer un profil-type ainsi que les trois formats de représentation les plus appropriés pour communiquer des résultats diagnostiques. Ces trois formats de présentation proposés doivent être variés et répondre à quatre grands critères de l'élaboration d'un rapport retenus de la littérature, à savoir l'accessibilité, la lisibilité, l'utilité et la validité. Ces trois formats de rapports proposés par les experts sont ensuite évalués par des enseignants au primaire, des conseillers pédagogiques et des orthopédagogues et les résultats obtenus servent à répondre à la dernière question de recherche qui est la suivante:

4. Quelles sont les perceptions des enseignants au primaire, des conseillers pédagogiques et des orthopédagogues sur les formats de rapports diagnostiques élaborés?

Les résultats en lien avec cette question obtenus suite à l'administration d'un questionnaire auprès des enseignants au primaire, des conseillers pédagogiques et des orthopédagogues nous amènent à décrire leur préférence, l'évaluation sur la qualité des rapports ainsi que leur compréhension des informations diagnostiques présentées. Ces résultats nous permettent de connaitre leurs perceptions sur les trois formats de rapports diagnostiques en lien avec des variables contextuelles révélées dans notre questionnaire.

CHAPITRE 3. MÉTHODOLOGIE

À titre de rappel, notre recherche vise deux objectifs principaux, soit à modéliser les données du test PIRLS 2011 à visée diagnostique et à élaborer et à évaluer des rapports diagnostiques, destinés aux enseignants, issus des résultats obtenus des modélisations. Plus spécifiquement, nous allons, dans un premier temps, vérifier la capacité diagnostique de l'épreuve PIRLS 2011 afin de tirer les conclusions sur la qualité diagnostique des items et sur les ajustements des modèles de DINA et de G-DINA aux données. Dans un deuxième temps, nous viserons à fournir des profils de maitrise et de non-maitrise des habiletés des élèves. Troisièmement, nous nous intéressons à proposer les formats de représentation les plus appropriés pour communiquer efficacement les résultats diagnostiques. Finalement, nous allons valider les rapports diagnostiques auprès du public visé afin d'évaluer la préférence des enseignants au primaire, des orthopédagogues et des conseillers pédagogiques, l'accessibilité, la lisibilité, l'utilité et la validité et la compréhension des rapports diagnostiques élaborés. Ainsi, nous allons organiser notre chapitre de méthodologie selon les deux objectifs généraux de notre recherche. Chaque objectif de recherche correspond à une phase avec ses sources de données, ses participants et ses instruments qui méritent d'être détaillés.

3.1. Phase 1: Modéliser les données du test PIRLS 2011 des élèves en 4e année au Canada dans une visée diagnostique

Pour répondre au premier objectif de la recherche, nous décrivons l'épreuve de PIRLS 2011, la base de données, les participants ainsi que la procédure de l'élaboration de la matrice Q et de la modélisation des données avec les modèles de DINA et de G-DINA.

3.1.1. Épreuve PIRLS 2011

Le PIRLS 2011 s'intéresse aux trois facettes suivantes de la lecture: (a) les deux buts de la lecture: lire pour l'expérience littéraire et lire pour acquérir et utiliser des informations; (b) les quatre processus cognitifs de la compréhension: se concentrer sur les informations énoncées de façon explicite et les extraire du texte; faire des inférences

simples; interpréter et combiner les idées et des informations; et examiner et évaluer le contenu, le langage et les éléments textuels; et (c) les comportements et les attitudes à l'égard de la compréhension de l'écrit (Labrecque et al., 2012; Mullis et al., 2009). Dans le cadre de ce projet de thèse, nous nous intéressons seulement aux deux premières facettes, soit les deux objectifs et les quatre processus cognitifs en lecture qui font l'objet de l'épreuve PIRLS 2011. Étant donné que le PIRLS s'intéresse à dresser un portrait exhaustif de la compréhension de l'écrit des élèves partout dans le monde, les extraits des textes et les items ont été choisis en fonction du cadre conceptuel ainsi que des deux objectifs de la lecture et quatre processus visés (Mullis et al., 2009; Labrecque et al., 2012). La répartition de questions selon les deux objectifs et quatre processus en lecture sont présentés dans le tableau 10.

Tableau 10. Pourcentages attribués aux deux objectifs et quatre processus en lecture du PIRLS 2011 (adapté de Labrecque et al., 2012; Mullis et al., 2009)

Objectifs de la lecture	
Expérience littéraire	50%
Acquérir et utiliser des informations	50%
Processus de compréhension	
Se concentrer sur les informations explicites et les extraire du texte	20%
Faire des inférences simples	30%
Interpréter et combiner des idées et des informations	30%
Examiner et évaluer le contenu, le langage et les éléments textuels	20%

En 2011, il y a au total 10 extraits des textes conçus pour l'évaluation du PIRLS, dont cinq textes littéraires et cinq textes informatifs. Parmi ces 10 extraits, six proviennent des épreuves précédentes tandis que les quatre extraits restants ont été nouvellement élaborés pour le PIRLS 2011 (Labrecque et al., 2012). Au total, il y a 135 items, dont de 13 à 16 items pour chaque extrait qui sont répartis de façon quasi égale entre des questions à choix multiples et des questions à court développement (Labrecque

et al., 2012). Les extraits et les items ont été divisés en 10 Blocs de 40 min chacun et ont été ensuite organisés en 13 cahiers selon un arrangement systématique. Les 13 cahiers se distinguent donc par la combinaison des extraits de textes ainsi que des items associés.

En raison de l'accessibilité aux contenus des items et des réponses des élèves, dans le cadre de ce projet, nous avons analysé les données des élèves en 4e année au Canada ayant pris le cahier 13 du PIRLS 2011. Le test est en anglais et se compose de 35 items, répartis en deux parties. La première partie correspond au premier extrait et contient 16 items tandis que la deuxième partie au deuxième extrait et comprend 19 items. Il y a au total 15 questions à choix multiples (QCM) à un point chacun. Chaque question à choix multiples comprend quatre choix de réponse qui sont rédigés de *«* façon concise afin de minimiser la charge de la lecture », dont une seule réponse est correcte, les trois autres réponses incorrectes sont plausibles, mais pas trompeuses (Labrecque et al., 2012, p. 17). Ainsi, dans le cadre du PIRLS, ce type de question est principalement utilisé pour évaluer les processus de compréhension qui ne demandent pas d'évaluation ou d'interprétation complexe. Les vingt items qui restent sont des items à réponse construite à deux points (19 items) et à trois points (1 item). L'attribution des points à ces questions à réponse construite dépend de la profondeur de la compréhension requise. Cette forme de question à court ou long développement vise à solliciter l'interaction entre le lecteur, le texte et le contexte et sert principalement à évaluer le processus de l'interprétation qui exige la mobilisation des connaissances et des expériences antérieures des élèves (Labrecque et al., 2012). Le tableau 11 résume la répartition des items du cahier 13 du PIRLS selon les objectifs de la lecture et les quatre processus de compréhension.

Tableau 11. Répartition des items du cahier 13 du PIRLS selon les deux objectifs de la lecture et les quatre processus de compréhension.

Extrait	Objectifs de la lecture	Processus de compréhension	Items
Enemy pie	**Expérience littéraire**	Se concentrer sur les informations explicites et les extraire du texte	3; 6; 7
		Faire des inférences simples	2; 5; 8; 9; 10; 11
		Interpréter et combiner des idées et des informations	4; 12; 14; 15
		Examiner et évaluer le contenu, le langage et les éléments textuels	1; 13; 16
The giant tooth mystery	**Acquérir et utiliser des informations**	Se concentrer sur les informations explicites et les extraire du texte	17; 19; 22; 29
		Faire des inférences simples	18; 21; 23; 27; 35
		Interpréter et combiner des idées et des informations	20; 24; 25; 26; 28; 31; 32; 33; 34
		Examiner et évaluer le contenu, le langage et les éléments textuels	30

3.1.2. Données

Au Canada, un total 23206 élèves provenant d'environ 1000 écoles ont participé au PIRLS 2011; de ce nombre, environ 16500 élèves ont fait le test en anglais alors qu'environ 6500 élèves ont participé au test en français (Labrecque et al., 2012). Ces élèves ont été répartis en 13 cahiers de tests, tel que présentés au tableau 12. Avec l'accessibilité au contenu des items et aux extraits de tests, dans le cadre de ce projet, nous avons décidé d'analyser les données des 4805 élèves canadiens ayant fait le test en anglais avec le cahier 13, ce qui représente 20,7% du total des participants. L'échantillon est composé à 49,3% de filles et 50,7% de garçons.

Tableau 12. Répartition des élèves au Canada selon les cahiers de tests du PIRLS 2011

Cahiers	Effectifs	Pourcentages
1	1528	6,6
2	1523	6,6
3	1522	6,6
4	1546	6,7
5	1523	6,6
6	1531	6,6
7	1528	6,6
8	1533	6,6
9	1542	6,6
10	1532	6,6
11	1551	6,7
12	1542	6,6
13	**4805**	**20,7**

Parmi les 4805 élèves, 43 participants ont manqué une des deux parties du test. Nous avons donc décidé de les éliminer de notre base de données. De plus, dans la base de données originale, les réponses des élèves sont enregistrées en quatre catégories: (0) réponse incorrecte; (1) réponse partiellement correcte; (2) réponse presque correcte; (3) réponse correcte. Afin de pouvoir les modéliser avec le DINA et G-DINA, nous avons dû coder ces réponses dichotomiquement (0 et 1). Plus précisément, les réponses aux QCM sont codées comme 1 (réponses correctes) ou comme 0 (réponses incorrectes). Pour les questions à développement, les réponses de 0 et 1 point sont codées comme 0 tandis que celles de 2 et 3 points sont codées comme 1. Les données manquantes ou les réponses incomplètes sont considérées comme mauvaises réponses et codées comme 0. Ainsi, finalement, nous obtenons une base de données dichotomiques des réponses de 4762 élèves aux 35 items. Les modélisations ont été réalisées avec ces deux bases de données afin de choisir celle qui s'ajuste le mieux aux modèles DINA et G-DINA.

3.1.3. Participants

L'élaboration de la matrice Q constitue une étape cruciale dans la modélisation des données dans une approche diagnostique cognitive. Ainsi, un panel d'experts doit

identifier les habiletés et les connaissances nécessaires pour répondre correctement aux items. Selon Loye (2008), pour être considérés comme experts, les participants doivent à la fois être des spécialistes diplômés du domaine disciplinaire concerné et bien connaitre le public visé par le test. Ainsi, nous avons choisi nos participants en nous basant sur trois critères: (a) avoir de bonnes connaissances en langue et (b) avoir des expériences dans l'enseignement des langues et (c) avoir analysé les données des tests de langues à grande échelle à visée diagnostique. Avec ces trois critères, nous avons formé un panel d'experts comprenant trois participants, dont la chercheuse pour l'élaboration de la matrice Q pour notre recherche. Ayant de très bonnes connaissances dans le domaine des langues, le premier expert a eu des expériences dans l'élaboration de la matrice Q pour le test PISA, un test très similaire à PIRLS dans une visée diagnostique tandis que le deuxième est très expérimenté dans l'identification des tables de spécification pour le test à grande échelle TCF, un test du français utilisé pour les immigrants au Québec. Le troisième expert est la chercheuse elle-même, qui a une formation en didactique des langues et de très bonnes connaissances en approche diagnostique cognitive. Dans ce cas, leurs tâches consistent à identifier les habiletés et les connaissances sous-jacentes pour répondre correctement aux items de ces deux tests. Leurs expériences acquises correspondent tout à fait à nos attentes pour l'élaboration de la matrice Q de notre projet de recherche.

3.1.4.1. Élaboration de la matrice Q

Deux matrices Q ont été élaborées pour notre recherche. Dans un premier temps, nous avons élaboré la première matrice Q (Q1) en nous basant sur les quatre processus de compréhension identifiés dans le cadre de référence du PIRLS 2011. L'intérêt est de vérifier si ces processus de compréhension pourraient être utilisés comme attributs pour des modélisations à visée diagnostique. Cette première matrice a été créée par la chercheuse elle-même en se basant sur le cadre de référence de PIRLS 2011 (Labrecque et al., 2012; Mullis et al., 2009).

Dans un deuxième temps, les trois experts ont identifié chacun de leur côté une liste d'attributs cognitifs sous-jacents pour répondre correctement aux items. Selon Loye (2008), le recours aux experts de manière individuelle lors de l'élaboration des attributs permet de tirer parti de leur expertise en évitant l'influence des autres experts. Afin de

faciliter le processus de l'élaboration de la matrice Q, des rencontres individuelles entre la chercheuse et les experts ont eu lieu. Lors de ces rencontres, la chercheuse doit fournir à chaque expert l'ensemble des informations concernant l'épreuve PILRS 2011, à savoir: (a) le contenu des deux passages, des items et des corrigés; (b) le cadre de référence de l'élaboration du PIRLS; (c) les modèles théoriques en lecture; (d) les informations sur les paramètres des items tel que présentées dans le rapport du CMEC (2012) et (e) les instructions et les attentes concernant les tâches demandées. Après cette première rencontre, chaque expert a identifié les attributs d'une manière indépendante en gardant des traces des arguments ou commentaires sur lesquels se base son jugement. Une seconde rencontre entre chaque expert et la chercheuse nous a permis de réviser les attributs et de répondre aux questions.

Les résultats obtenus à partir des attributs proposés par les trois experts pour chaque item ont été analysés à trois niveaux. Dans un premier temps, les analyses visent le niveau d'items. La statistique de Kappa de Fleiss (1971) a été calculée pour mesurer le degré d'accord inter-experts. Ainsi, pour chaque item, seulement les attributs obtenant plus de 50% de taux d'accord de trois experts ont été retenus comme essentiels pour l'item. Dans un deuxième temps, les attributs sélectionnés pour chaque item ont été compilés pour former une matrice Q2_initiale pour le PIRLS 2011. Finalement, cette matrice Q2_initiale a été examinée et raffinée encore une fois par nos experts afin d'arriver à une matrice Q2_finale pour les modélisations avec DINA et G-DINA. Bien que des directives strictes ne soient pas proposées, Hartz (2002) suggère que chaque attribut doit être mesuré par au moins trois items et défini au sens large. Ainsi, les attributs qui ne sont pas associés à au moins trois items sont combinés soit à un attribut similaire ou éliminé de la matrice Q2_finale (Kim, 2011).

3.1.4.2. Estimation des paramètres et évaluation de l'ajustement de DINA et G-DINA

Après avoir identifié les attributs et élaboré la matrice Q2_finale, l'étape suivante consiste à estimer les paramètres des items et des sujets. En principe, les modèles reposent sur une estimation non bayésienne, partiellement bayésienne ou totalement bayésienne selon la portion des paramètres estimés imposée à une distribution préalable

(a priori) (Dibello, Roussos et Stout, 2007). Dans le premier cas, tous les paramètres sont fixés sans la distribution préalable, une partie des paramètres est imposée à une distribution a priori pour le deuxième cas et dans le dernier cas, tous les paramètres sont attribués à une distribution préalable. Dans le cadre du modèle DINA et G-DINA, la procédure MLE (Maximum Likelihood Estimation) avec les deux méthodes extensions JMLE (Joint Maximum Likelihood Estimation) (de la Torre, 2009) et MMLE (Marginal Maximum Likelihood Estimation) (de la Torre, 2009 ; Huebner et Wang, 2011) sont souvent utilisées pour estimer les paramètres. Ces paramètres sont calculés soit avec l'algorithme MCMC (Markov Chain Monte Carlo) (de la Torre et Douglas, 2008 ; de la Torre, 2009) ou EM (Expectation-maximization) (de la Torre, 2009 ; Huebner et Wang, 2011). Dans le cadre de ce projet, sous la permission de de la Torre (2009) d'accéder aux codes écrits avec OxEdit qui est un logiciel gratuit, nous avons utilisé la procédure d'estimation MLE avec l'algorithme EM pour estimer les deux paramètres de pseudo-chance et d'étourderie. De plus, la procédure de MLE est la méthode la plus couramment utilisée pour estimer les paramètres et déterminer l'erreur de l'estimation, ce qui mérite donc d'être présenté plus en détail.

3.1.4.2.1. Estimation des paramètres avec MLE et EM

L'idée fondamentale de cette méthode d'estimation du maximum de vraisemblance (MLE) est qu'un bon choix de l'estimateur d'un paramètre d'intérêt est celui qui permet aux données observées d'apparaitre avec la plus grande occurrence possible (Huebner et Wang, 2011). Pour ce faire, nous devons établir une fonction de vraisemblance pour les paramètres d'intérêt, prendre son logarithme, et ensuite calculer la dérivée partielle de la fonction de log-vraisemblance pour chaque paramètre en définissant le résultat des équations égales à 0. Nous allons ensuite résoudre ces équations pour trouver les valeurs des paramètres. Étant donné que le DINA est une distribution conditionnelle des réponses des sujets X_{ij} et avec son vecteur d'habileté α_i, la vraisemblance conditionnelle de X_i peut être exprimée comme suit:

$$L(\boldsymbol{X}_i|\boldsymbol{\alpha}_i) = \prod_{j=1}^{J} P_j(\boldsymbol{\alpha}_i)^{X_{ij}} [1 - P_j(\boldsymbol{\alpha}_i)]^{1-X_{ij}}$$

dans lequel, $P_j(\alpha_i)$ est la probabilité du sujet j de répondre correctement à l'item i (de la Torre, 2009). Selon Huebner et Wang (2011) et de la Torre (2009), si les vecteurs d'habiletés sont connus, il sera facile de calculer les paramètres $(g_j; s_j)$. Toutefois, en réalité, ces vecteurs sont inconnus, les deux paramètres et le vecteur d'habiletés doivent être estimés en même temps avec JMLE, ce qui rend inconsistants les résultats des paramètres et des vecteurs d'habiletés (de la Torre, 2009; Huebner et Wang, 2011). À cause du manque d'appui théorique sur la consistance des résultats, cette méthode, JMLE, est peu utilisée actuellement (Dibello, Roussos et Stout, 2007). Une méthode alternative est d'utiliser le MMLE en fournissant la distribution du vecteur d'habiletés des sujets et l'intégration sur cette distribution (de la Torre, 2009 ; Dibello, Roussos et Stout, 2007 ; Huebner et Wang, 2011 ; Mislevy et Wilson, 1996). Dans ce cas, la vraisemblance marginale de X peut être estimée comme suit :

$$L(\boldsymbol{X}) = \prod_{i=1}^{I} L(\boldsymbol{X}_i) = \prod_{i=1}^{I} \sum_{l=1}^{L} L(\boldsymbol{X}_i|\boldsymbol{\alpha}_l) p(\boldsymbol{\alpha}_l)$$

où : L(X_i) est la vraisemblance marginale du vecteur de réponse du sujet j; P(α_l) est la probabilité antérieure du vecteur d'habiletés α_l et L=2^K, K est le nombre d'attributs (habiletés) (de la Torre, 2009). Comme les paramètres aléatoires du vecteur d'habiletés ont été éliminés de la fonction de la vraisemblance, l'estimation des paramètres d'items ne dépend plus de l'estimation du vecteur d'habiletés de chaque sujet, mais plutôt de la somme pondérée de la vraisemblance de 2^K schémas d'attributs (de la Torre, 2009 ; Huebner et Wang, 2011). Cette méthode MMLE peut être utilisée pour estimer les paramètres de DINA et G-DINA en utilisant l'algorithme EM (de la Torre, 2009 ; Huebner et Wang, 2011).

Dans le DINA et G-DINA, l'estimation des paramètres par EM comprend trois étapes. La première étape consiste à définir les valeurs initiales pour les paramètres de pseudo-chance (g_j) et d'étourderie (s_j). La deuxième étape vise à estimer les nouveaux paramètres: $I_{jl}^{(0)}$ (le nombre espéré des sujets qui manquent au moins d'un attribut demandé pour l'item j); $R_{jl}^{(0)}$(le nombre de sujets espéré parmi le $I_{jl}^{(0)}$ qui répondent correctement à l'item j); $I_{jl}^{(1)}$ et $R_{jl}^{(1)}$ (ont la même interprétation, mais pour le groupe qui

maîtrisent tous les attributs pour l'item j). Ces paramètres vont être calculés en se basant sur les valeurs courantes de pseudo-chance (g) et d'étourderie (s). L'étape 3 permet de trouver les vraies valeurs de *g* et s en replaçant les valeurs estimées de ces 4 nouveaux paramètres dans les équations. Les étapes 2 et 3 sont répétées jusqu'à ce que la convergence soit atteinte (de la Torre, 2009).

Bien que le MMLE puisse résoudre le problème de l'inconsistance dans l'estimation des paramètres d'items, elle a des difficultés dans l'estimation du vecteur d'habiletés des sujets ou sa procédure va être lente lorsque le nombre d'attributs est grand (Huebner et Wang, 2011). Toutefois, en comparaison avec le MCMC, le EM est plus efficace dans l'estimation des paramètres des MCD (Dibello, Roussos et Stout, 2007). Par contre, le EM a tendance à être plus difficile à étendre à des nouveaux modèles ou à des modèles avec de nouveaux paramètres que le MCMC (Dibello, Roussos et Stout, 2007). Peu importe la méthode d'estimation utilisée, sa finalité est de vérifier la convergence et la non-convergence (DiBello, Roussos et Stout, 2007). Si la non-convergence se produit, il faut réviser le processus de la construction du modèle ou d'élaboration de la matrice Q.

3.1.4.2.2. Évaluation de l'ajustement des modèles aux données

Une des étapes importantes des analyses diagnostiques cognitives des MCD est l'évaluation de l'ajustement des modèles aux données. Cette étape permet de vérifier la cohérence essentielle entre le modèle théorique sous-jacent et les données observées pour suggérer des améliorations au modèle et au processus de générer les données (DiBello, Roussos et Stout, 2007; Sinharay et Almond, 2007). Plus concrètement, cette étape se concentre sur l'évaluation du degré de l'ajustement entre le modèle estimé (valeurs prédites) et les valeurs observées.

Dans l'évaluation de l'ajustement de MCD, nous pouvons distinguer l'évaluation de *l'ajustement relatif* de celle de *l'ajustement absolu*. Avec la disponibilité des différents MCD, il est important de choisir le modèle le plus approprié grâce à l'évaluation de l'ajustement relatif. Ainsi, l'évaluation de *l'ajustement relatif* des MCD fait référence au processus de la sélection du modèle le plus approprié parmi un ensemble des modèles concurrents (Chen, de la Torre, Zhang, 2013). Dans le cadre de ce projet, les trois

statistiques suivantes sont utilisées pour l'évaluation de l'ajustement relatif de DINA et G-DINA :

(1) -2 log-likelihood

$$-2LL = 2\ln(ML)$$

(2) Akaike's information criterion (AIC)

$$AIC = -2LL + 2P$$

(3) Bayesian information criterion (BIC)

$$BIC = -2LL + P\ln(N)$$

Où: ML est le maximum de vraisemblance des paramètres d'items; P est le nombre de paramètres du modèle; L est le nombre total des schémas d'attributs et N est la taille de l'échantillon. Pour chacune de ces trois statistiques, le modèle avec la valeur la plus petite va être sélectionné parmi un ensemble des modèles concurrents (Chen, de la Torre, Zhang, 2013).

L'évaluation de *l'ajustement absolu* des MCD réfère au processus de déterminer si les modèles sont ajustés aux données de manière adéquate. Ainsi, les trois statistiques suivantes sont utilisées pour évaluer l'ajustement absolu des modèles DINA et G-DINA aux données :

(1) le résidu entre la proportion d'items corrects prédite et observée ;

(2) la corrélation transformée de Fisher prédite de chaque paire d'items (appelée corrélation transformée) et

(3) le résidu entre le ratio log-odds observé et prédit de chaque paire d'items (Chen, de la Torre, Zhang, 2013).

Avec ces trois statistiques, un grand nombre de schémas d'attributs sont échantillonnés à partir de la distribution postérieure des attributs. Les schémas d'attributs généralisés et les paramètres estimés peuvent être utilisés pour générer les réponses prédites aux items. La différence entre les réponses observées et celles prédites devrait être 0 si le modèle est ajusté aux données de manière adéquate. Afin d'utiliser ces trois statistiques, nous devons calculer leurs erreurs standardisées (SE) permettant de dériver les scores Z de ces trois statistiques afin de vérifier si les résidus sont statistiquement différents de 0. Le rejet de n'importe quel score z signifie que le modèle ne s'ajuste pas à

un item ou à une paire d'items d'une manière adéquate (Chen, de la Torre, Zhang, 2013). Ainsi, cette étape d'évaluation de l'ajustement absolu permet de détecter les erreurs de spécification de la matrice Q, soit la sur-spécification ou la sous-spécification des attributs.

En résumé, les lignes précédentes servent à décrire les méthodes d'estimation des paramètres de DINA et G-DINA ainsi que l'évaluation de l'ajustement relatif et absolu de ces modèles. Dans ce projet, nous avons utilisé la méthode d'estimation MLE avec EM, car il s'agit d'une méthode couramment utilisée. Bien que les codes soient disponibles dans le logiciel R, nous avons décidé de choisir le logiciel OxEdit, étant donné que le logiciel est très facile à utiliser et qu'il ne demande pas des connaissances très poussées en langage du codage. De plus, avec la permission du professeur Jimmy de la Torre, nous avons eu accès aux codes écrits pour ce logiciel. Nous avons également présenté les six statistiques développées pour l'évaluation de l'ajustement relatif et absolu de ces modèles, dont trois pour évaluer l'ajustement absolu et trois pour l'ajustement relatif. Ces statistiques vont être utilisées pour interpréter nos résultats sur l'ajustement des modèles aux données afin de tirer des conclusions nécessaires.

Avec ces deux matrices Q élaborées (Q1 et Q2_finale), nous avons modélisé les données avec les modèles DINA et G-DINA. L'objectif est de déterminer la matrice Q qui s'ajuste mieux à ces deux modèles. Celle qui s'ajuste le plus va être retenue pour fournir les résultats détaillés sur les profils de maitrise et de non-maitrise des habiletés des élèves. La partie suivante vise à décrire l'évaluation des profils de la maitrise et non-maitrise des habiletés des candidats.

3.1.4.2.3. Évaluation des profils de maitrise et non-maitrise des habiletés

Le but des MCD est de classifier les sujets dans des classes latentes en offrant un profil d'habiletés correspondant à leur degré de maitrise et non-maitrise des attributs. Avec le modèle DINA et G-DINA, nous pouvons obtenir les probabilités a posteriori de l'ensemble des classes latentes. Le tableau 13 présente un exemple de cinq classes latentes avec leurs probabilités associées de chaque groupe. Les classes latentes représentent les groupes des sujets qui obtiennent les mêmes profils de maitrise des attributs. Ainsi, en référant à ce tableau, 20,56 % des sujets font partie du profil (00000),

c'est-à-dire qu'ils ne maitrisent aucune de cinq habiletés. Par contre, 13,85 % des sujets ont le profil (11111) et maitrisent les cinq habiletés.

Tableau 13. Exemple d'un extrait des classes latentes et leurs probabilités (adapté de Ravand, Barati et Widhiarso, 2012)

Classes latentes	Probabilités
00000	0,2056
00100	0,0348
10011	0,0575
11110	0, 226
11111	0,1385

Les résultats des modélisations avec DINA et G-DINA donnent aussi la probabilité de maitrise pour chacune des habiletés. Le tableau 14 donne un exemple de 5 attributs avec leurs probabilités de maitrise. Ainsi, la probabilité de maitrise de l'attribut 1 est 0,546, c'est-à-dire que 54,6 % des sujets ont maitrisé l'habileté 1. L'habileté la mieux maitrisée est la cinquième, qui l'est par 71,9 % des sujets.

Tableau 14. Exemple de 5 attributs avec leurs probabilités de maitrise (adapté de Ravand, Barati et Widhiarso, 2012)

Attributs	A1	A2	A3	A4	A5
Probabilités	0,546	0,451	0,552	0,565	0,719

Finalement, le DINA et G-DINA peuvent fournir également le profil de maitrise individuel de chaque sujet. Ce profil peut être sous forme de probabilité ou de pourcentage pour chaque habileté et la valeur peut être arrondie à 0 ou 1 avec le point de coupure de 0,5 pour le DINA et G-DINA. La figure 8 présente un exemple d'un profil de maitrise des habiletés d'un sujet issu du test METLAB (Li, 2011) dans laquelle la ligne rouge verticale représente le point de coupure de 0,5. Ainsi, la probabilité de maitrise de ce sujet pour l'habileté 2 (syntaxe) est 0,03, donc inférieur à 0,5, ce sujet ne maitrise pas suffisamment cette habileté. Par contre, il maitrise les habiletés 1 (vocabulaire), 3 (extraire des informations explicites) et 4 (comprendre des informations implicites), ce qui aboutit au profil de maitrise (1011) pour ce sujet.

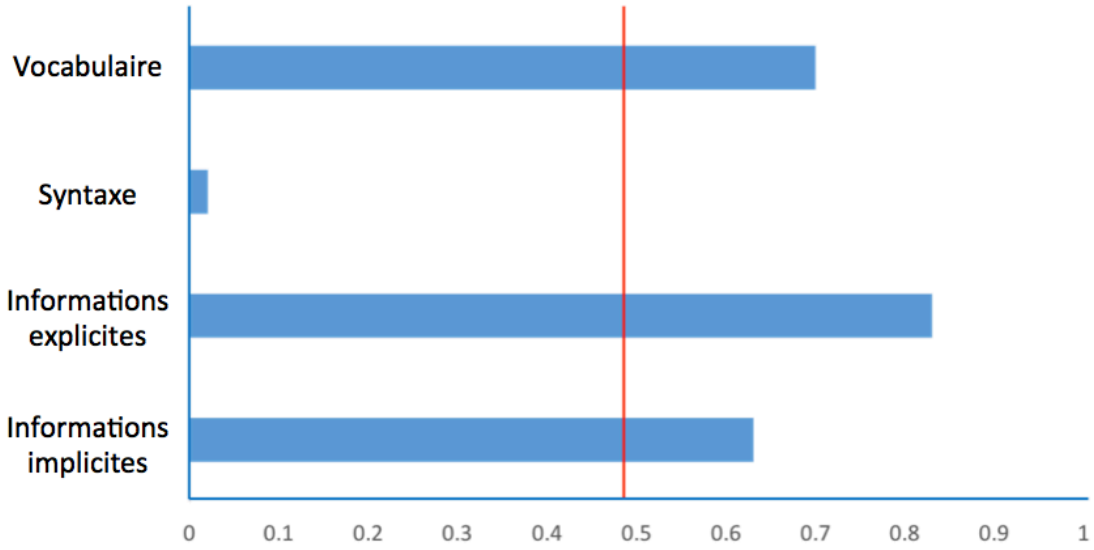


Figure 8. Exemple d'un profil de maitrise des habiletés d'un sujet issu du test METLAB (adapté de Li, 2012).

Le fait de fournir des profils détaillés sur le degré de maitrise et non-maitrise des habiletés autres que les scores ou les rangs totaux permet aux élèves et aux enseignants d'identifier leurs forces et leurs faiblesses et de guider l'apprentissage des premiers et l'enseignement des seconds. Ainsi, l'ADC a attiré l'attention de plusieurs chercheurs du domaine de langues, notamment en lecture, sur l'application de certains modèles de classification diagnostique à des tests de performance à grande échelle afin d'extraire des informations plus fines sur la structure des connaissances linguistiques et les habiletés cognitives (Lee et Sawaki, 2009). Ces recherches ont donc confirmé l'applicabilité des MCD et le potentiel diagnostique des tests à grande échelle dans le domaine des langues, ce qui nous a fourni des balises théoriques et empiriques importantes pour notre recherche avec l'épreuve PIRLS 2011. Le tableau 15 propose un résumé des éléments méthodologiques pour la première question générale de recherche

Tableau 15. Résumé des éléments méthodologiques pour la première question générale de recherche

Objectif général	**Objectifs spécifiques**	**Déroulement**	**Participants**	**Données**	**Analyses**
Modéliser les données du test PIRLS 2011 des élèves en 4e année à des fins diagnostiques	Vérifier la capacité diagnostique du test PIRLS 2011	Élaborer deux matrices Q	Trois experts	Contenu de 35 items et deux extraits de textes; les corrigés; le cadre de référence du PIRLS; les modèles cognitifs et les paramètres d'items	-Analyse du contenu des items et des extraits des textes -Analyse des arguments et des commentaires qui se basent les jugements des experts -Analyse du degré d'accord inter-juges afin de créer la matrice Q finale
		Modéliser les données avec DINA et G-DINA avec le logiciel Ox	4762 élèves en 4 au Canada ayant pris le cahier 13 du PIRLS 2011	- Données des réponses de 4762 élèves qui sont codées dichotomiquement -Deux matrices Q	-Évaluation de l'ajustement relatif et absolu des modèles aux données -Évaluation de la capacité diagnostique des items avec les indices de pseudo-chance et d'étourderie
	Fournir des profils de maitrise des habiletés aux élèves du Canada ayant participé au test PIRLS 2011	Évaluer les profils de maitrise et non-maitrise des habiletés des classes latentes et de l'individu	4762 élèves en 4e au Canada ayant pris le cahier 13 du PIRLS 2011	Les résultats obtenus suite aux modélisations des données avec DINA et G-DINA	-Analyse de la classification des classes latentes -Les probabilités de maitrise de chaque habileté -Les profils de maitrise de chaque individu

3.2. Phase 2: Élaborer et évaluer des rapports diagnostiques destinés aux enseignants

Le but ultime des analyses diagnostiques cognitives est de fournir aux apprenants des rétroactions détaillées sur leurs forces et leurs faiblesses cognitives à travers des rapports diagnostiques compréhensibles et interprétables. Ceci constitue le deuxième objectif général de la recherche. Ainsi, la partie suivante sert à décrire la procédure d'élaboration des rapports diagnostiques avec un panel d'experts, la façon dont ces rapports ont été évalués par des enseignants au primaire, des conseillers pédagogiques et des orthopédagogues, l'élaboration du questionnaire conçu à cet effet ainsi que les stratégies d'analyses des données envisagées. Une synthèse permet de résumer des éléments essentiels à retenir pour la méthodologie de notre recherche en lien avec l'objectif 2.

3.2.1. Élaboration des rapports diagnostiques

Dans le contexte de ce projet, l'élaboration des rapports diagnostiques a été effectuée selon deux étapes. Dans un premier temps, nous avons conçu un protocole du design des rapports diagnostiques en nous basant sur la revue de la littérature. Dans un deuxième temps, ce protocole a été présenté et discuté avec le panel d'experts afin de choisir le profil-type et les formats des rapports les plus appropriés à concevoir et à valider auprès des enseignants au primaire, des conseillers pédagogiques et des orthopédagogues.

3.2.1.1. Préparation du protocole du design des rapports diagnostiques

L'élaboration des rapports diagnostiques est considérée comme une étape principale afin de répondre au deuxième objectif général de notre recherche. Les informations sur le design des rapports ont été rassemblées à partir de la revue de littérature sur l'élaboration des rapports en général et des rapports diagnostiques en particulier. Plus précisément, nous nous basons principalement sur les principes d'élaboration des rapports proposés par différents auteurs (Goodman et Hambleton (2004), Hambleton et Slater (1997), Jaeger (1998), Wainer et al. (1999), Ryan (2006), Aschbacher et Herman (1991), De Vito et Konieg (2001), Zenisky, Hambleton, Sireci (2009), Sinharay, Puhan et Haberman (2010),

Roberts et Gierl (2010), Klesch (2010)) en s'assurant des quatre critères principaux, à savoir l'accessibilité, l'utilité, la lisibilité et la validité. Quant au contenu, les rapports diagnostiques doivent couvrir nécessairement les informations sur les objectifs des rapports, les descriptions des habiletés évaluées, les probabilités de maitrise pour chaque habileté, un résumé du profil de maitrise et de non-maitrise des habiletés et un guide d'interprétation des résultats d'après l'exemple d'un rapport diagnostique proposé par Roberts et Gierl (2010) avec la méthode AHM. Les formats de présentation des résultats doivent être variés avec des graphiques, des tableaux et des éléments textuels et sont élaborés avec ou sans couleur.

À la suite de la consultation avec les experts, nous avons décidé d'élaborer trois formats de rapports à partir d'un profil type choisi dans les résultats des modélisations du PIRLS 2011 de la phase 1. Lorsque nous travaillons sur un même profil, les comparaisons entre les différents formats de présentation deviennent donc plus pertinentes étant donné que les caractéristiques liées au profil sont constantes. Ce profil-type a été sélectionné selon les probabilités de maitrise des habiletés des classes latentes et doit contenir à la fois les habiletés maitrisées et non maitrisées. Pour ce faire, sept profils-types dont les probabilités de maitrise des habiletés les plus grandes ont été présentés au panel d'experts afin d'en choisir un pour l'élaboration des rapports.

Les rapports ont été développés dans l'intention qu'ils puissent être consultés en ligne ou imprimés, c'est pour cette raison que le nombre de pages doit se limiter à deux afin qu'ils puissent être imprimés au recto et verso d'une feuille de papier (Roberts et Gierl, 2010). Ainsi, la première page du rapport contient trois parties essentielles: (1) l'identification de l'élève et de l'enseignant et les directives pour lire le rapport; (2) les détails des résultats au test et le profil de maitrise et (3) une description du profil de l'élève et la proposition des pistes d'intervention. La deuxième page porte sur les informations supplémentaires comme la description détaillée des habiletés et le guide d'interprétation. Les figures 9 et 10 permettent de schématiser le tronc commun de trois différents formats du profil type ainsi que le guide d'interprétation. Ces figures font partie du protocole utilisé pour travailler avec le panel d'experts.

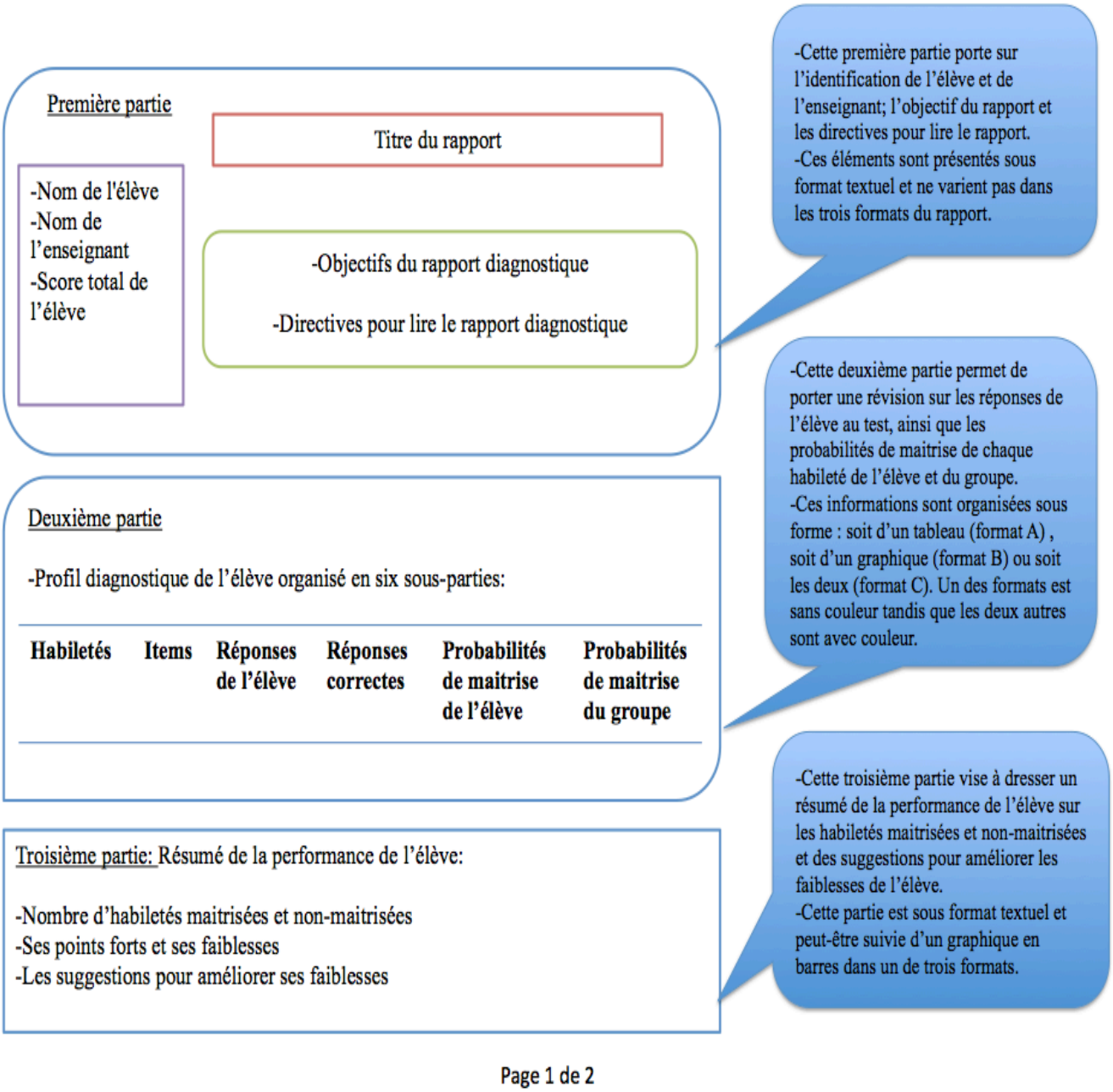


Figure 9. Schéma de la première page du rapport diagnostique

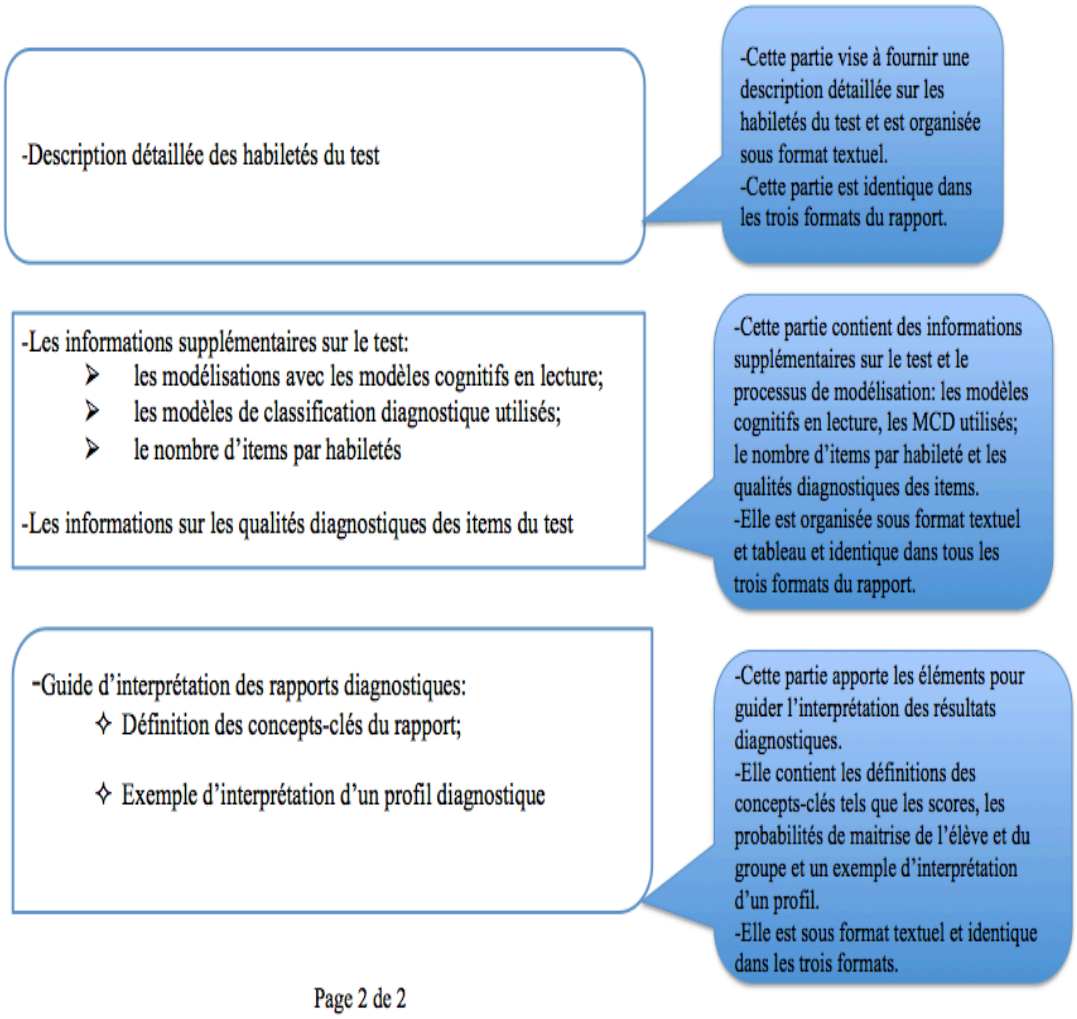


Figure 10. Schéma du guide d'interprétation du rapport diagnostique

3.2.1.2. Élaboration des rapports diagnostiques avec le panel d'experts

Un panel d'experts comprenant cinq membres a été formé pour l'élaboration des rapports. Les membres du panel d'experts doivent répondre au moins à un des critères suivants: (a) avoir une très bonne compréhension en mesure et évaluation, notamment l'approche diagnostique cognitive; (b) avoir une l'expertise dans le domaine de langues et (c) avoir de l'expérience en enseignement au primaire. Notre panel d'experts comprend donc: (a) une enseignante au primaire; (b) une conseillère pédagogique; (c) la responsable du service de mesure et évaluation d'une firme de testing; (d) une experte en mesure et évaluation et (e) la chercheuse.

Quatre rencontres en groupe ont été organisées à cet effet. Lors de la première rencontre, nous avons présenté les objectifs de la recherche, l'approche diagnostique cognitive ainsi que les tâches assumées par le panel d'experts. Nous leur avons également fourni des informations sur le test PIRLS 2011, ses objectifs ainsi que les quatre processus de compréhension en lecture identifiés dans le cadre de référence. Comme support, nous avons eu recours à deux extraits du test, les questions avec les habiletés en lecture annotées pour chaque question ainsi que les corrigés. Étant donné que le test est en anglais, les habiletés ont été initialement définies en anglais, nous avons décidé de reformuler ces habiletés dans un langage familier de la salle de classe lorsque nécessaire en nous basant sur les documents ministériels du MELS et du CMEC.

Ensuite, sept exemples de profils correspondant aux classes latentes ayant les probabilités de maitrise les plus élevées ainsi que leurs tableaux récapitulatifs de réponse au test ont été également présentés aux experts. Dans les profils-type, les probabilités de maitrise des habiletés ont été présentées sous forme des diagrammes en barres tandis que les réponses au test ont été présentées sous forme des tableaux en intégrant les réponses correctes et incorrectes de l'élève avec les habiletés identifiées pour chaque question. Les membres du panel d'experts ont analysé les profils-type avec ces tableaux de réponse pour en identifier un qui ressemble le plus facilement à un élève-type dans la classe. Ces profils-type sont présentés dans le chapitre 4. Le profil-type retenu a fait l'objet de l'élaboration de trois formats de rapports.

L'objectif principal de la deuxième rencontre était d'apporter des propositions sur le contenu et sur la forme des rapports diagnostiques selon le protocole élaboré. Ainsi, dans un premier temps, nous avons examiné quelques exemples de rapports élaborés à visée diagnostique et non-diagnostique suggérés dans les recherches empiriques sur le sujet afin d'avoir un regard global sur ce qui existe. Ensuite, le protocole du design des rapports a été fourni aux membres du panel d'experts afin qu'ils puissent proposer des éléments sur le contenu et la forme de présentation des rapports.

À la suite de leurs propositions, les premières versions de ces trois formats de rapports ont été développées. Lors de la troisième rencontre, les experts ont porté leurs commentaires sur ces formats élaborés. Comme support, nous avons préparé une fiche d'évaluation des rapports diagnostiques destinés aux membres du panel d'experts. Cette fiche contient une liste de vérification portant sur les différents éléments du rapport avec leurs commentaires ou suggestions (tableau 5.2 de l'annexe 5). Les questions posées aux experts portent nécessairement sur les quatre critères d'un bon rapport, à savoir l'accessibilité, l'utilité, la lisibilité et la validité des informations diagnostiques présentées dans les trois formats. Le tableau 16 présente donc quelques exemples des questions posées aux experts. Les discussions avec les experts ont permis non seulement d'améliorer les versions finales des rapports, mais aussi d'identifier des pistes pour l'élaboration du questionnaire administré aux enseignants pour l'évaluation externe de ces rapports diagnostiques dans la phase 3. Ainsi, cette étape a abouti à la production de trois formats finaux de rapports diagnostiques qui font ensuite l'objet d'une validation auprès d'enseignants au primaire, d'orthopédagogues et de conseillers pédagogiques. Étant donné que les formats de rapports élaborés font partie des résultats du travail avec le panel d'experts, nous les avons décrits plus en détail dans le chapitre 4. La quatrième rencontre a été entièrement consacrée à l'évaluation des versions finales des rapports et à l'élaboration du questionnaire pour l'évaluation de ces trois formats de rapports que nous abordons dans la partie suivante.

3.2.2. Évaluation des rapports diagnostiques par des enseignants, des orthopédagogues et des conseillers pédagogiques

Les trois formats ont été ensuite évalués par des enseignants au primaire, des orthopédagogues et des conseillers pédagogiques en utilisant un questionnaire administré en ligne. La partie suivante sert à décrire le questionnaire, les participants et les stratégies d'analyse des données envisagées.

3.2.2.1. Questionnaire

Un questionnaire (voir annexe 6) a été conçu afin de déterminer lequel des formats de représentation des résultats diagnostiques est le plus approprié ainsi que les perceptions des enseignants, des orthopédagogues et des conseillers pédagogiques sur les rapports diagnostiques élaborés. Le questionnaire porte nécessairement sur quatre aspects: (a) la préférence des enseignants sur les formats de représentation des résultats diagnostiques; (b) l'évaluation des participants sur les trois formats de rapports élaborés et (c) leur compréhension sur les trois formats de rapports et (d) des informations sociodémographiques.

En ce qui concerne la préférence des enseignants relativement au format des rapports, nous avons adapté les quatre questions utilisées dans la recherche de Zapata-Rivera, Vezzu et VanWinkle (2013). Les participants sont invités à choisir un format de rapports parmi les trois formats affichés à leur écran. La partie relative à l'évaluation des enseignants sur les trois formats de rapports diagnostiques élaborés se compose des questions sur les quatre critères; l'accessibilité, l'utilité, la lisibilité et la validité. L'évaluation porte sur les trois parties du rapport ainsi que sur le guide d'interprétation. En outre, nous avons utilisé certaines questions adaptées de la recherche de Vezzu, VanWinkle et Zapata (2012) sur les représentations graphiques et les reformulations des éléments textuels. Nous avons y ajouté des questions afin de vérifier la compréhension des enseignants sur les rapports diagnostiques élaborés. Ces questions d'évaluation sur les quatre critères ont été construites avec une échelle de type Likert à quatre points selon le degré d'accord ou désaccord tandis que les questions de compréhension sont à choix multiple avec une seule réponse correcte. Finalement, les questions sociodémographiques

portent, par exemple, sur le genre, la tranche d'âge, les années d'expérience et la formation des participants.

3.2.2.2. Participants

Ces trois formats de rapports et le questionnaire administrés en ligne ont été pré-validés auprès de deux conseillères pédagogiques et d'un expert en mesure et évaluation. L'objectif de cette étape de pré-validation est de vérifier la compréhension des questions afin de détecter les problèmes dans la reformulation des questions, l'accès aux rapports et au questionnaire ainsi que la navigation. Les commentaires et les suggestions des experts face à cette version préliminaire nous ont permis d'améliorer notre version finale du questionnaire.

Le questionnaire, dans lequel ont été inclus les trois formats de rapport, a été administré en ligne avec Survey Monkey auprès de 151 enseignants au primaire, conseillers pédagogiques et orthopédagogues. Le temps estimé pour répondre au questionnaire est d'environ 30 min. L'intérêt est de savoir si la préférence, l'évaluation des participants ainsi que leur compréhension des rapports diffèrent selon ces groupes et les formats de rapports. Pour les quatre questions qui portent sur la préférence, les trois formats de rapports ont été affichés afin que les participants puissent choisir un format de leur préférence. Quant aux questions de l'évaluation et de la compréhension des rapports, seulement un de trois formats a été assigné aléatoirement par Survey Monkey à chaque participant de manière à assurer une exposition équitable.

Nous avons fait la demande de l'approbation éthique au Comité plurifacultaire d'éthique de la recherche (CPÉR) de l'Université de Montréal et avons obtenu le certificat éthique pour notre collecte de données (annexe 7). Le consentement des participants a été obtenu en ligne sur Survey Monkey. Plus précisément, avant de répondre au questionnaire, les participants ont été invités à lire la lettre d'information et de consentement présentée en ligne et à donner leur consentement ou non-consentement en cochant Oui ou Non à une question sur le sujet (voir annexe 6). Ceux qui voulaient recevoir une copie écrite de la lettre de consentement avec la signature de la chercheuse ont été invités à inscrire leurs courriels. Par contre, ceux qui ont donné le non-consentement ont été éliminés de notre base de données lors du traitement des données.

Comme les trois formats et le questionnaire ont été administrés en ligne, les participants ont toute la liberté de les évaluer à l'endroit et au moment de leur choix selon leurs disponibilités.

3.2.2.3. Analyse des données

Les réponses des participants en ce qui concerne leur compréhension ont été exportées directement dans un fichier SPSS pour des analyses statistiques. Nous avons ensuite codé les réponses aux questions de compréhension en deux catégories: réponse correcte et réponse incorrecte afin de calculer les scores totaux de la compréhension des participants. Chaque réponse correcte vaut 1 point tandis la réponse incorrecte équivaut à 0 point.

L'analyse des perceptions des enseignants au primaire, des orthopédagogues et des conseillers pédagogiques sur les rapports diagnostiques élaborés portent sur les points suivants: (a) leur préférence; (b) leur évaluation des quatre critères d'un bon rapport et (c) leur compréhension des rapports diagnostiques élaborés. Ainsi, afin de choisir le format le plus efficace de représentation des résultats diagnostiques des participants, nous avons, dans un premier temps, réalisé des analyses descriptives pour les quatre questions sur la préférence. Les résultats obtenus sous forme de pourcentages ou effectifs nous permettent de déterminer le format le plus choisi. Dans un deuxième temps, nous nous intéressons à savoir si la préférence des formats de représentation est liée à des éléments sociodémographiques tels que le sexe, l'ancienneté, la tranche d'âge, le diplôme obtenu et le domaine de formation, le suivi ou non des cours en méthodes quantitatives, nous avons fait des tests de Khi-deux pour étudier ces liens de dépendance.

En ce qui concerne l'évaluation des rapports, nous avons, dans un premier temps, fait des analyses factorielles avec les 32 questions de cette partie afin de retenir le regroupement des facteurs qui correspondent aux différentes sous-parties du rapport que les participants doivent évaluer. Le fait de regrouper les questions dans des facteurs nous permet de comprendre plus en profondeur les perceptions des participants sur la qualité de chaque partie du rapport. Des variables composites ont ensuite été calculées à partir des facteurs retenus. Nous avons ensuite fait des analyses descriptives et inférentielles de l'évaluation des rapports en tenant compte des variables sociodémographiques. Nous

avons recouru à des tests inférentiels tels que le test t pour les échantillons indépendants, des analyses de ANOVA à un facteur afin de vérifier si les perceptions des participants quant à l'évaluation des rapports varient selon le format de rapport évalué, le genre, la tranche d'âge, l'ancienneté, le diplôme obtenu, le domaine de formation, le suivi ou non des cours en méthodes quantitatives ainsi que leur préférence.

En ce qui concerne la compréhension des rapports, dans un premier temps, nous avons fait des analyses descriptives avec la variable composite qui regroupe les 9 questions. Deuxièmement, nous avons fait des tests de Khi-deux pour voir les liens entre la compréhension de chaque question avec les variables sociodémographiques. Nous avons également recouru à des tests non-paramétriques de Mann-Whitney pour étudier si l'évaluation de la qualité des rapports varie selon la compréhension des rapports. Le tableau 16 résume les éléments importants de la méthodologie afin de répondre au deuxième objectif général de notre recherche.

Tableau 16. Résumé des éléments méthodologiques pour le deuxième objectif général de la recherche

Objectif général	Objectifs spécifiques	Déroulement	Participants	Outils et données	Analyses
Élaborer et évaluer des rapports diagnostiques destinés aux enseignants	Choisir les formats de représentations les plus appropriés pour communiquer les résultats diagnostiques d'une manière efficace	-Élaboration de 3 formats de rapports diagnostiques -Validation interne auprès du panel d'experts -Conception du questionnaire pour l'évaluation des rapports	-5 experts	-Revue de littérature sur l'élaboration des rapports diagnostiques et des recherches empiriques -Réponses des experts aux questions posées sur les quatre critères	-Analyses descriptives des réponses et des commentaires des experts
	- Évaluer des rapports diagnostiques auprès du public visé	-Pré-validation externe de trois formats de rapports et du questionnaire	-2 conseillères pédagogiques et un expert en mesure et évaluation	Réponses des participants sur la compréhension des questions, les problèmes d'accès et de navigation	- Analyses descriptives des réponses et des commentaires des participants
		-Validation externe auprès des participants avec le questionnaire -Évaluation de la préférence, des qualités des rapports et de la compréhension des participants des rapports élaborés.	-151 enseignants au primaire, orthopédagogues et conseillers pédagogiques	Réponses des participants sur leur préférence, l'évaluation de la qualité et la compréhension des rapports ainsi que les données sociodémographiques	-Analyses descriptives -Analyses inférentielles: test chi 2; test t pour échantillons indépendants, ANOVA à un facteur et test non-paramétrique de Mann-Whitney

3.3. Synthèse de la méthodologie

Notre recherche vise à modéliser dans une visée diagnostique les données de 4762 élèves au Canada ayant fait le test PIRLS 2011, puis à élaborer et à évaluer des rapports diagnostiques destinés aux enseignants. Plus précisément, nous visons 4 objectifs spécifiques: (a) vérifier la capacité diagnostique de l'épreuve PIRLS 2011: (b) fournir des profils de maitrise et non-maitrise des habiletés aux élèves; (c) proposer trois formats de rapports diagnostiques les plus appropriés et (d) évaluer des rapports diagnostiques par des enseignants au primaire, des orthopédagogues et des conseillers pédagogiques. Ainsi, nous avons organisé notre chapitre selon les deux phases correspondant à ces deux objectifs généraux de la recherche.

Pour répondre au premier objectif spécifique de recherche, nous avons d'abord élaboré deux matrices Q (Q1 et Q2_finale). La Q1 a été créée par la chercheuse tandis qu'un panel d'experts comprenant trois membres a été formé pour élaborer la Q2 en analysant le contenu des items et des passages. La base de données contenant les réponses aux 35 items de l'épreuve PIRLS a été codée dichotomiquement pour les modélisations. Par la suite, nous avons modélisé des données avec les modèles DINA et G-DINA afin d'estimer les paramètres de pseudo-chance et d'étourderie des items, d'évaluer les ajustements relatifs et absolus des modèles aux données et d'évaluer les profils de maitrise et de non-maitrise des habiletés des élèves. Avec ces deux bases de données et deux matrices Q, nous avons une matrice à laquelle ces deux modèles DINA et G-DINA s'ajustent le mieux pour fournir des profils de maitrise des habiletés aux élèves (objectif spécifique 2). Ces profils font l'objet du contenu des rapports diagnostiques à élaborer par un panel d'experts comprenant cinq membres et à évaluer auprès d´enseignants au primaire, des orthopédagogues et des conseillers pédagogiques, ce qui constitue nos troisième et quatrième objectifs spécifiques de la recherche.

Afin de répondre au troisième objectif qui vise à choisir des formats de présentation les plus appropriés des rapports, un panel d'experts comprenant cinq membres a été formé. À la suite des résultats obtenus des modélisations, nous avons choisi sept profils parmi les plus fréquents et les avons présentés au panel d'experts afin qu'ils retiennent un profil-type pour la conception des rapports diagnostiques. Quatre

rencontres en groupe ont été organisées afin que les experts puissent donner leurs propositions sur le contenu et les formats des rapports à partir du protocole conçu par la chercheuse. Ces rapports diagnostiques contiennent quatre parties principales: (A) une partie d'introduction sur l'identification de l'élève; les directives pour les rapports; (B) la représentation des résultats; (C) un résumé du profil de l'élève et (D) un guide d'interprétation des résultats. En général, les parties A, C, D sont identiques dans les trois formats des rapports. Ce qui diffère est la partie B avec les différents formats de représentation des résultats qui pourraient être sous forme d'un tableau, d'un graphique et des deux.

Un questionnaire a été conçu en nous basant sur la description de quatre critères principaux d'un bon rapport et sur les recherches de Zapata-Rivera, Vezzu et VanWinkle (2013) et de Vezzu, VanWinkle et Zapata (2012). Ce questionnaire, avec des questions sur une échelle de type Likert et des QCM, porte sur quatre dimensions: (a) la préférence des participants sur le format de présentation des résultats; (b) l'évaluation de la qualité des rapports, (c) leur compréhension et (d) les informations sociodémographiques. Le questionnaire a été ensuite administré en ligne et pré-validé auprès de deux conseillères pédagogiques et d'un expert en mesure et évaluation afin de vérifier leur compréhension aux questions posées et de détecter les problèmes d'accès et de navigation en ligne. Finalement, nous avons administré le questionnaire à 151 enseignants du primaire, orthopédagogues et conseillers pédagogiques pour l'évaluation externe des rapports. Chaque participant est invité à choisir un format de préférence parmi les trois formats affichés à l'écran, et à évaluer ensuite seulement un des trois formats assignés aléatoirement par l'ordinateur afin d'assurer une exposition équitable des trois formats. Des analyses descriptives et inférentielles ont été réalisées afin de fournir des éléments de réponses à la 4e question spécifique de notre recherche sur les perceptions des participants sur les formats de rapports élaborés.

CHAPITRE 4. RÉSULTATS

Dans ce chapitre, nous présentons d'abord les résultats du processus de l'élaboration des matrices Q par les experts, à savoir la matrice élaborée à partir du cadre de référence du PIRLS 2011, les matrices individuelles originales élaborées par les experts ainsi que la matrice finale proposée pour des modélisations.

Par la suite, grâce au logiciel Ox, nous effectuons des modélisations des données avec les deux matrices Q permettant d'évaluer des ajustements relatifs et absolus des modèles de DINA et G-DINA aux données, d'estimer des paramètres d'items et d'identifier les profils de maitrise des habiletés les plus courants des élèves.

Finalement, nous exposons des résultats du processus de l'élaboration des rapports avec le panel d'experts et de l'évaluation des formats de rapports par des enseignants au primaire, des orthopédagogues et des conseillers pédagogiques. Ces résultats permettent de nous renseigner sur les perceptions des enseignants au primaire, des orthopédagogues et des conseillers pédagogiques quant à leur préférence, leur évaluation de la qualité ainsi qu'à leur compréhension de ces rapports.

4.1. Élaboration des matrices Q

4.1.1. Matrice élaborée à partir du cadre de référence du PIRLS 2011 (Q1)

À titre de rappel, le test PIRLS 2011 vise deux objectifs en lecture, soit: (1) lire pour l'expérience littéraire et (2) lire pour acquérir des informations, puis quatre processus de compréhension en lecture: soit (1) se concentrer sur les informations énoncées de façon explicite et les extraire du texte, (2) faire des inférences simples, (3) interpréter et combiner les idées et des informations, et (4) examiner et évaluer le contenu, le langage et les éléments textuels. Chaque livret du PIRLS se compose de deux extraits qui correspondent chacun à un objectif de lecture différent. De plus, dans les extraits, chaque item évalue un processus de compréhension spécifique parmi les quatre processus de compréhension proposés. La description détaillée de ces processus de compréhension en lecture a été abordée dans la section 2.1.2.2. Le tableau 17 résume la description des attributs ainsi que la répartition des items par attribut.

Tableau 17. Description des attributs et la répartition des items par attribut.

Attribut	Définition	Nombre d'items
Examiner et évaluer le contenu, le langage et les éléments textuels (A1)	Porter un regard critique sur son contenu, le langage utilisé ou les éléments textuels et réfléchir à la clarté de l'expression du sens, en faisant appel à ses propres connaissances sur le genre en question, la structure ou les conventions linguistiques.	4
Faire des inférences simples (A2)	Combler les « lacunes » relatives au sens en déduisant des informations à partir du texte.	11
Se concentrer sur les informations explicites et les extraire du texte (A3)	Comprendre et repérer les informations énoncées de façon explicite et faire le lien avec la question posée.	7
Interpréter et combiner des idées et des informations (A4)	Acquérir une compréhension plus approfondie du texte en combinant les connaissances antérieures et les informations présentées dans le texte.	13

En nous basant sur le cadre de référence du PIRLS 2011, le contenu des extraits et des items, nous avons élaboré la matrice Q1. Étant donné que chaque item évalue un seul processus de compréhension, chacun des 35 items correspond à un seul attribut. Le tableau 18 présente donc la matrice Q1. Dans cette matrice Q1, nous constatons une répartition inégale des attributs selon les items et les extraits. Par exemple, pour l'attribut A1, il n'y a que 4 items au total (11,43% des items), dont 3 pour l'extrait 2 *«* Ennemy pie » et 1 pour l'extrait 2 *«* The giant tooth mystery ». Par contre, l'attribut A4 a été identifié par le plus d'items: 4 items pour l'extrait 1 et 9 items pour l'extrait 2, donc un total de 13 items (37,14%) pour cet attribut. Quant à l'attribut A2, il y a au total 11 items (31,43%), dont 6 items pour l'extrait 1 et 5 pour l'extrait 2. Finalement, sept items au total (20%) sont en lien avec l'attribut A3, avec une répartition de 3 items pour l'extrait 1 et 4 items pour l'extrait 2. Malgré cette répartition des items par attribut, chaque attribut a été identifié par au moins trois items, ce qui répond bien au critère sur le seul minimal de questions par attribut pour l'élaboration de la matrice Q proposée par Hartz, Roussos, et Stout (2002).

Tableau 18. Matrice Q1 élaborée à partir du cadre de référence du PIRLS 2011.

Items	**A1** Examiner et évaluer le contenu, le langage et les éléments textuels	**A2** Faire des inférences simples	**A3** Se concentrer sur les informations explicites et les extraire du texte	**A4** Interpréter et combiner des idées et des informations
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1
5	0	1	0	0
6	0	0	1	0
7	0	0	1	0
8	0	1	0	0
9	0	1	0	0
10	0	1	0	0
11	0	1	0	0
12	0	0	0	1
13	1	0	0	0
14	0	0	0	1
15	0	0	0	1
16	1	0	0	0
17	0	0	1	0
18	0	1	0	0
19	0	0	1	0
20	0	0	0	1
21	0	1	0	0
22	0	0	1	0
23	0	1	0	0
24	0	0	0	1
25	0	0	0	1
26	0	0	0	1
27	0	1	0	0
28	0	0	0	1
29	0	0	1	0
30	1	0	0	0
31	0	0	0	1
32	0	0	0	1
33	0	0	0	1
34	0	0	0	1
35	0	1	0	0

4.1.2. Matrices individuelles originales élaborées par les experts

Avec les critères de sélection des experts mentionnés dans le chapitre 3, nous avons formé un panel comprenant trois membres. Le premier expert est une professeure en évaluation des langues qui a des expériences dans l'élaboration de la matrice Q avec le test PISA à visée diagnostique. Le deuxième expert est la chercheuse avec de très bonnes connaissances en approche diagnostique cognitive et une formation en didactique des langues. Finalement, le troisième expert est un doctorant en mesure et évaluation qui a beaucoup d'expériences dans l'identification des tables de spécification des tests de langues à grande échelle. Le processus de l'élaboration de la matrice Q été divisée en deux étapes. Dans un premier temps, la chercheuse a travaillé en collaboration avec l'expert 1 lors de son stage à Rutgers pour identifier une liste d'attributs jugés nécessaires pour l'épreuve PIRLS 2011. En analysant le cadre de référence du PIRLS 2011, le contenu de deux extraits, des questions et des corrigés ainsi que la matrice Q1, l'expert 1 a proposé d'identifier cinq attributs qu'il juge nécessaires pour le test, tandis que l'expert 2 voulait garder les quatre processus de compréhension en lecture tels quels comme attributs. Toutefois, selon ce dernier, certaines questions sollicitent plus d'un processus de compréhension pour répondre à l'item. Nous avons dû analyser plus en détail la description de chaque processus de compréhension selon les tâches demandées. Dans le processus de compréhension « Interpréter et combiner des idées et des informations », cela exige des élèves de (a) dégager le message ou le thème général d'un texte; (b) envisager une séquence d'actes pour les personnages; (c) mettre en évidence les points communs et les différences des informations du texte; (d) dégager l'atmosphère et le ton d'une histoire et (e) interpréter les applications possibles des informations dans le monde réel. Selon nous, ces cinq types de tâches demandées nécessitent à la fois chez les élèves, une compréhension globale du test et une interprétation des idées dans leurs propres mots. De plus, ce processus de compréhension a été identifié pour le plus grand nombre d'items du test PIRLS 2011 selon la matrice Q1; nous avons donc décidé de le distinguer en deux attributs séparés, c'est-à-dire la compréhension globale et l'interprétation.

Quant au processus « examiner et évaluer le contenu, le langage et les éléments textuels », les élèves doivent: (a) prendre un recul par rapport au texte afin de porter un regard critique sur son contenu, le langage utilisé ainsi que les éléments textuels; (b)

comparer la signification d'un mot exprimée par l'auteur à sa propre compréhension ou aux informations en provenance d'autres sources et (c) réfléchir à la clarté de l'expression du sens, en faisant appel à ses propres connaissances sur le genre en question, la structure ou les conventions linguistiques. Ce processus de compréhension demande donc une compréhension globale du texte et aussi la capacité des élèves à reformuler les idées du texte selon leurs propres mots, ce qui est lié à l'habileté à maitriser le vocabulaire et la syntaxe. Avec ces analyses, les deux experts se sont mis d'accord pour estimer que les cinq attributs suivants devraient être retenus pour le test: (1) *Repérer des informations explicites; (2) Compréhension globale; (3) Interpréter; (4) Faire des inférences simples et (5) Vocabulaire et syntaxe*. Le tableau 19 présente la définition plus détaillée de ces cinq attributs.

Tableau 19. Définitions détaillées des attributs

Attribut	**Définition**
(A1) *Repérer des informations explicites*	Localiser et reconnaitre des informations explicites exprimées dans le texte pour répondre aux questions
(A2) *Compréhension globale*	Former une compréhension globale d'un paragraphe ou de l'ensemble du texte
(A3) *Interprétation*	Clarifier le sens des idées ou des configurations complexes et interpréter des relations
(A4) *Faire des inférences*	Comprendre les informations qui ne sont pas explicitement exprimées en faisant des inférences ou des prédictions
(A5) *Vocabulaire et syntaxe*	Exprimer des idées dans une grammaire correcte et compréhensible de l'anglais écrit

Lors de la 2e étape, avec la liste de ces cinq attributs, les trois experts ont identifié des attributs pour chaque item selon les directives et les documents de référence fournis par la chercheuse. Une copie de la matrice Q1 a été également fournie aux experts. Les experts ont accompli leurs tâches individuellement et il n'y avait pas de séances de formation organisées à cet effet. Nous leur avons demandé de choisir plus d'un attribut

par item s'ils le jugent nécessaire. Dans ce cas, ils devraient classer ces attributs selon l'ordre d'importance. De plus, si les experts pensent que leurs attributs identifiés pour chaque item sont différents du processus de compréhension proposé par le cadre de référence du PIRLS 2011, ils devraient écrire leurs commentaires expliquant leur choix. Étant donné que l'expert 3 n'a pas participé au processus d'identification de la liste d'attributs lors de la première étape, nous lui avons demandé d'ajouter d'autres attributs à la liste fournie s'il le juge nécessaire.

La chercheuse a contacté individuellement les experts 1 et 3 pour voir s'ils avaient des questions à poser et pour assurer qu'ils avaient bien suivi les procédures. Lors de cette rencontre, les experts ont dit qu'ils n'avaient aucune difficulté à comprendre la notion « attribut cognitif » ni de suivre des instructions fournies et qu'ils se sont beaucoup inspirés des documents de référence du PIRLS 2011. L'experte 1 disait qu'elle avait de la facilité à réaliser ces tâches étant donné qu'elle a beaucoup travaillé avec la base de données de PISA dans la même perspective de diagnostic. L'expert 3 a, quant à lui, également identifié des habiletés pour la base de données du test TFI, mais pour la compréhension orale, il est donc familier avec le type de travail demandé pour notre projet.

Lors de la remise des résultats, aucun attribut n'a été ajouté par l'expert 3. Toutefois, il a proposé de reformuler la définition de l'attribut 1 en remplaçant l'expression « les informations explicites du texte » par les « détails spécifiques du texte », car l'expression des « informations explicites du texte » lui semble très générale. Les experts ont identifié deux attributs et plus pour un grand nombre d'items (environ 75% d'items).

Nous avons calculé les coefficients de corrélation inter-juges pour mesurer le taux de concordance des experts sur l'identification des attributs pour chaque item. Étant donné que nous avons plus de deux juges et que les échelles de mesure sont de types dichotomiques, donc nominales, nous avons utilisé le coefficient Kappa de Fleiss (1971). Ces coefficients Kappa de Fleiss ont été calculés avec le programme AgreeStat2015 pour chaque item. Le tableau 20 présente les résultats de l'élaboration de la matrice Q selon trois experts.

Tableau 20. Matrice Q2 initiale élaborée par les experts

	Experts			
Items	1	2	3	Attributs proposés[1]
1	1; 2	2	2;4	2;4
2	4;5	3;5	3;5	3;5
3	1	1	1	1
4	3;4	3;4	3;4	3;4
5	1;4;5	4;5	1;4;5	1;4;5
6	1	1	1	1
7	1	1	1	1
8	1	4	4	4
9	4;5	1;3;5	3;5	3;5
10	1;4	3;4	3;4	3;4
11	3;4	1;3;4	3;4	3;4
12	3;4	3;4	1;3;4	3;4
13	2;4	2;3;4	2;4	2;4
14	3;5	2;3;5	2;3;5	2;3;5
15	1;2;3;5	1;2;3;4	1;2;3;4;5	1;2;3;4;5
16	2;5	2;3	2;3;5	2;3;5
17	1	1	1	1
18	1;3	1;3	1;3	1;3
19	1	1	1	1
20	1;3	1;2;3	1;3	1;2;3
21	4	3;4	3;4	3;4
22	1	1	1	1
23	1;4	4	3;4	3;4
24	3;5	3;5	3;5	3;5
25	3;5	3;5	3;5	3;5
26	3;5	3;5	3;5	3;5
27	4	3	3;4	3;4
28	3;5	2;3;5	3;5	3;5
29	1	1	1	1
30	2	2;5	2;5	2;5
31	1;3	1;3	1;3	1;3
32	1;3	3	1;3	1;3
33	1;3	1	1;3	1;3
34	1;3	1;3	1;3	1;3
35	1;3	1	1	1

Les résultats suggèrent que 17, 2% des items ont une légère concordance (Kappa de Fleiss de 0 à 0,4); 11,4% des items ont une concordance moyenne (Kappa de Fleiss de

[1] Attributs retenus lors de la discussion avec les experts

0,41 à 0,6) et 31,4% des items ont une concordance importante avec les coefficients Kappa de Fleiss qui varient entre 0,61 et 0,8 et 40% des items obtiennent une concordance presque parfaite des experts avec des coefficients Kappa de Fleiss qui varient entre 0,81 et 1. Les items qui obtiennent une plus grande concordance des experts sont de manière logique surtout des items à un seul attribut. Pour les items qui ont des taux de concordance faible et moyenne, nous avons examiné en détail des commentaires suggérés par les experts afin de proposer la matrice Q2_finale pour les modélisations.

4.1.3. Matrice Q2_finale proposée pour des modélisations

Avec le cadre de référence de l'élaboration du PIRLS 2011 ainsi que les commentaires fournis par des experts lors de l'élaboration de la matrice Q2_initiale, nous avons vérifié encore une fois les items qui ont un taux de concordance faible et moyenne avant de proposer la matrice Q2_finale pour des modélisations telle que présentée dans le tableau 23. Les ajustements ont été accordés pour les items 1; 2; 8; 9; 10; 11; 14; 15; 16; 20; 21; 27 et 35 en nous basant sur les types de commentaires suivants suggérés par les experts.

Premièrement, en s'appuyant sur la description des habiletés, il nous semble impossible de combiner les deux attributs *Repérer des informations explicites et Faire des inférences* comme c'est le cas des items 10; 11 et 23 à cause de leur caractère plus ou moins contradictoire, en effet l'un nécessite une reconnaissance des informations explicites exprimées dans le texte pour répondre aux questions tandis que l'autre demande une compréhension des informations implicites. L'expert 3 a donc suggéré l'identification de deux attributs *Interpréter* et *Faire des inférences* pour ces items. Cette proposition nous semble logique étant donné que les réponses pour ces questions ne sont pas explicitement exprimées dans le texte. Nous avons donc décidé de retenir les attributs *Interpréter* et *Faire des inférences* pour ces questions.

Certains items nécessitent le remplacement de l'attribut *Repérer des informations explicites* par l'attribut *Faire des inférences*, comme le cas des items 1 et 8. Pour l'item 1 et 8, l'expert 1 a identifié l'attribut *Repérer des informations explicites,* l'expert 2 a proposé l'attribut *Compréhension globale* tandis que l'expert 3 a identifié les attributs *Faire des inférences* et *Compréhension globale.* Nous avons dû examiner le contenu et le

corrigé de cette question et constaté que les informations nécessaires pour répondre correctement à cet item ne sont pas explicitement exprimées dans le texte. L'élève doit donc faire des inférences et avoir une compréhension globale pour déterminer la personne qui raconte l'histoire. Dans ce cas, les attributs *Faire des inférences* et *Compréhension globale* sont ceux qui devraient être retenus pour l'item 1. Ce même problème a été constaté pour l'item 8, car le choix de réponse correct pour cette question n'est pas une information explicite que l'élève peut repérer dans le texte. Cette question demande un faible niveau de l'habileté *Faire des inférences* plutôt que l'habileté *Repérer des informations explicites*.

Les experts 2 et 3 suggèrent l'ajout d'un autre attribut pour les items 14; 16; 15; 20; 21 et 27. Par exemple, pour les items 14; 16 et 20, l'attribut *Compréhension globale* devrait être ajouté dans la liste des attributs retenus pour répondre correctement aux items tandis que les items 21 et 27 nécessitent également l'habileté *Interpréter* autre que l'habileté *Faire des inférences* et *Compréhension globale*. Quant à la question 5 qui contient deux sous-questions, tous les cinq attributs devraient être sollicités pour fournir une réponse correcte à cette question, nous avons dû ajouter l'attribut *Interpréter* pour cette question. Dans un autre ordre d'idée, un attribut a été enlevé pour les items 9 et 35. Ces items ne nécessitent pas que l'élève fasse des inférences pour répondre aux questions, car les informations sont explicitement exprimées dans le texte. Nous avons donc enlevé cet attribut dans la liste des attributs identifiés pour ces items. Finalement, nous avons obtenu la matrice Q2_finale telle que présentée dans le tableau 23. Au total, nous avons 9 items à un attribut; 21 items à deux attributs; 4 items à trois attributs et 1 seul item à cinq attributs.

Nous avons utilisé les matrices Q1 (tableau 18) et Q2_finale (tableau 21) pour modéliser les données de 4762 élèves canadiens ayant pris le livret 13 du PIRLS 2011 dont nous présenterons les résultats obtenus dans les lignes qui suivent.

Tableau 21. Matrice Q2_finale proposée pour des modélisations du PIRLS 2011

Items	**A1** Repérer des informations explicites	**A2** Compréhension globale	**A3** Interprétation	**A4** Faire des inférences	A5 Vocabulaire syntaxe
1	0	1	0	1	0
2	0	0	1	0	0
3	1	0	1	0	0
4	0	0	1	1	0
5	1	0	0	1	1
6	1	0	1	0	0
7	1	0	1	0	0
8	0	0	0	1	0
9	0	0	1	0	1
10	0	0	1	1	0
11	0	0	1	1	0
12	0	0	1	1	0
13	0	1	0	1	0
14	0	1	1	0	1
15	1	1	1	1	1
16	0	1	1	0	1
17	1	0	0	0	0
18	1	0	1	0	0
19	1	0	0	0	0
20	1	1	1	0	0
21	0	0	1	1	0
22	1	0	0	0	0
23	0	0	1	1	0
24	0	0	1	0	1
25	0	0	1	0	1
26	0	0	1	0	1
27	0	0	1	1	0
28	0	0	1	0	0
29	1	0	0	0	0
30	0	1	0	0	1
31	1	0	1	0	0
32	1	0	1	0	0
33	1	0	1	0	0
34	1	0	1	0	0
35	1	0	0	0	0

4.2. Estimation des paramètres d'items et de profils des élèves

Dans cette partie, nous présenterons, dans un premier temps, l'évaluation de l'ajustement relatif et absolu des modèles DINA et G-DINA aux données afin de décider quel modèle s'ajuste le mieux aux données. Dans un deuxième temps, nous présenterons l'estimation des paramètres d'items avec les deux matrices Q1 et Q2_finale. Finalement, nous décrirons les profils de maitrise des habiletés des élèves.

4.2.1. Évaluation des ajustements relatifs et absolus des modèles aux données

Comme nous l'avons présenté dans le chapitre 3, trois statistiques ont été retenues pour l'évaluation de l'ajustement relatifs des modèles aux données: -2LL; AIC et BIC. Le modèle ayant les plus petites valeurs sera choisi comme celui qui s'ajuste le mieux aux données. En regardant le tableau 22, nous constatons que le modèle G-DINA s'ajuste mieux aux données que le modèle DINA avec la matrice Q2_finale tandis qu'avec la matrice Q1, le DINA semble s'ajuster mieux aux données. Plus précisément, avec le DINA et avec la matrice Q1, les indices de -2LL, AIC et BIC sont légèrement inférieurs à ceux avec le G-DINA. Par contre, avec la matrice Q2_finale, le modèle G-DINA s'ajuste mieux que le DINA. Dans les deux modèles, les données s'ajustent mieux avec la matrice Q2_finale qu'avec la matrice Q1. Nous concluons donc que le modèle G-DINA est celui qui s'ajuste le plus aux données avec la matrice Q2_finale étant donné les valeurs plus petites de ces statistiques.

Tableau 22. Ajustements relatifs des modèles aux données avec Q1 et Q2_finale

CDM/Matrice Q		-2LL	AIC	BIC
DINA	Q1	171273,2021	171443,2021	171993,0181
	Q2	170382,2495	170584,2495	171237,5603
G-DINA	Q1	171274,9403	171444,9403	171994,7562
	Q2	163575,4671	163969,4671	165243,7464

Quant à l'évaluation de l'ajustement absolu, les trois statistiques suivantes ont été obtenues à partir des modélisations: (1) le résidu entre la proportion d'items corrects prédite et observée; (2) la corrélation transformée de Fisher prédite de chaque paire d'items et (3) le résidu entre le ratio log-odds observé et prédit de chaque paire d'items (Chen, de la Torre, Zhang, 2013). Les valeurs maximales des scores Z de la proportion

correcte (prop), de la corrélation transformée (Z (Corr) et du ratio log-odds (Log (OR) ont été utilisées pour l'évaluation de l'ajustement absolus des modèles aux données. Ces valeurs doivent être proches de 0 pour tous les items pour montrer que le modèle s'ajuste aux données. Les seuils de rejet de ces scores Z ont été également utilisés pour décider si les modèles s'ajustent adéquatement aux données ou pas. En principe, les valeurs maximales de ces statistiques doivent être inférieures aux valeurs critiques pour montrer que le modèle retenu s'ajuste adéquatement aux données. Au cas contraire, l'ajustement du modèle retenu est rejeté (Ma et Meng, 2014).

Le tableau 23 présente les résultats de l'ajustement absolu des modèles aux données avec les deux matrices Q. Pour la proportion correcte, les valeurs maximales des scores Z sont plus ou moins semblables avec les deux matrices Q1 et Q2_finale dans les deux modèles DINA et G-DINA. Ces valeurs sont plus petites avec le modèle G-DINA que le DINA, et elles sont plus petites avec la matrice Q2 qu'avec la matrice Q1. En comparant avec les valeurs critiques des scores Z avec la correction de Bonferroni, ces valeurs obtenues sont toutes inférieures aux valeurs critiques. Nous concluons que les modèles DINA et G-DINA s'ajustent aux données avec tous les items et avec les deux matrices Q.

Tableau 23. Ajustements absolus des modèles aux données avec la Q1 et Q2_finale

	Matrice	Prop		Z (Corr)		Log (OR)	
		DINA	G-DINA	DINA	G-DINA	DINA	G-DINA
Max	Q1	0,0125	0,0076	1,2399	0,9744	12,2611	11,7888
	Q2	0,0089	0,0062	0,9880	0,3658	11,1028	8,4344

Note. Max=valeur maximale des scores Z; Prop=Proportion correcte; Z (corr)= la corrélation transformée et Log (OR)= ratio log-odds; valeur de score Z critique zc = 3,467; 3,649, 4,044 pour α = 0,1; 0,05; 0,01, respectivement (avec la correction de Bonferroni).

Avec les scores Z de la corrélation transformée de Ficher, les valeurs maximales sont proches pour le modèle DINA et G-DINA avec la matrice Q1. La différence est plus grande entre le DINA et le G-DINA dans le cas de la matrice Q2_finale. En ce qui concerne la proportion correcte (prop), le modèle G-DINA semble mieux s'ajuster aux données que le DINA et il s'ajuste plus avec la matrice Q2_finale qu'avec la matrice Q1.

La comparaison de ces valeurs avec les valeurs critiques des scores Z nous amène à confirmer donc que les modèles de DINA et G-DINA s'ajustent adéquatement aux données pour tous les items et avec les deux matrices Q.

Quant aux scores Z du ratio Log-odds, les valeurs obtenues sont grandement différentes de celles obtenues avec la proportion correcte et la corrélation transformée, même si nous observons les mêmes tendances, c'est-à-dire, les valeurs sont plus petites avec le modèle de G-DINA et qu'avec le DINA et plus petites avec la matrice Q2_finale qu'avec la matrice Q1. Cependant, ces valeurs obtenues sont toutes supérieures aux valeurs critiques des scores Z. Nous concluons donc que les modèles de DINA et G-DINA ne s'ajustent pas adéquatement aux données avec tous les items.

En comparant les trois statistiques pour l'évaluation de l'ajustement absolu des données, nous constatons que les valeurs sont proches entre la proportion correcte et la corrélation transformée dans les deux modèles et avec les deux matrices Q, ce qui aboutit à la même décision de ne pas rejeter l'hypothèse nulle et conclure qu'il y a de l'ajustement absolu des modèles aux données avec tous les items. Cette décision nous suggère que la sensibilité de ces deux statistiques dans l'évaluation de l'ajustement absolu des données est plus ou moins identique. Par contre, les valeurs du ratio de Log-odds sont considérablement différentes des deux statistiques précédentes et nous amènent à rejeter l'hypothèse nulle, donc il n'y a pas d'ajustement absolu des modèles aux données avec tous les items. Cette décision nous questionne sur la fiabilité et la sensibilité de cette statistique dans l'évaluation de l'ajustement absolu des données.

En résumé, les lignes précédentes servent à présenter les résultats de l'évaluation de l'ajustement relatif et absolu des modèles DINA et G-DINA aux données avec les deux matrices Q1 et Q2_finale. Nous avons présenté les résultats de six statistiques, dont trois pour l'évaluation de l'ajustement relatif et trois pour l'évaluation de l'ajustement absolu. Ces résultats suggèrent que le modèle G-DINA s'ajuste mieux que le DINA aux données et plus avec la matrice Q2_finale que la matrice Q1. En ce qui concerne l'ajustement absolu, les statistiques de la proportion correcte et de la corrélation transformée suggèrent que les modèles DINA et G-DINA s'ajustent adéquatement aux données avec toutes les deux matrices Q. Cependant, les résultats du ratio de Log-odds nous suggèrent

que les modèles ne s'ajustement pas adéquatement aux données avec tous les items. Nous nous basons donc sur les résultats de la proportion correcte et de la corrélation transformée étant donné que ces résultats sont plus ou moins semblables, ce qui nous amène à la même décision. Dans la partie suivante, nous présenterons l'estimation des paramètres d'items obtenus avec la matrice Q2_finale et la classification des profils de maitrise des habiletés des élèves obtenus avec la matrice Q2_finale et le modèle G-DINA.

4.2.2. Estimation des paramètres d'items

Les résultats des modélisations nous fournissent également l'estimation de deux paramètres d'items, soit le paramètre de pseudo-chance (g) et d'étourderie (s). L'estimation de ces deux paramètres permet d'évaluer la qualité ainsi que la capacité diagnostique des items. La capacité diagnostique de chaque item est déterminée par 1-g-s. Les seuils pour l'interprétation de ces deux paramètres sont souvent biaisés et varient d'un auteur à l'autre. Par exemple, selon de la Torre (2009) et de la Torre et Douglas (2008) les paramètres de pseudo-chance et d'étourderie des items peuvent être classés en trois catégories comme suit: 0-0,1=Bonne qualité; 0,1-0,2= moyenne qualité et 0,2-0,3=Faible qualité. Par contre, selon Wenchao et al. (2016), les items sont classés comme de bonne qualité si ces deux paramètres se trouvent entre 0 à 0,15; de moyenne qualité s'ils sont entre 0,15 et 0,25 et de faible qualité si ces paramètres sont entre 0,25 à 0,35. Finalement, selon Ravand et Widhiarso (2013), les items sont considérés comme de bonne qualité si les deux paramètres sont inférieurs à 0,5 et de faible qualité s'ils sont supérieurs à 0,5.

Ainsi, en nous basant sur le tableau 25, la moyenne du paramètre de pseudo-chance est de 0,36443, c'est-à-dire, qu'en moyenne, l'élève a 36,43% de chance de répondre correctement aux questions même s'il ne maitrise pas tous les attributs nécessaires pour le test. Selon les critères proposés par de la Torre (2009) et de la Torre et Douglas (2008), il y a six items de bonne qualité, à savoir l'item 20; l'item 24; l'item 26; l'item 30; l'item 31 et l'item 34. Les items 15; 32 et 33 sont classés comme de moyenne qualité tandis que les items 5; 7; 16; 25; 28; 29; 35 sont de faible qualité. Les 19 items qui restent sont considérés comme problématiques en termes de pseudo-chance. Il est à noter

que la plupart de items de bonne et de moyenne qualité en termes de pseudo-chance se trouvent dans la deuxième partie du test. La première partie du test ne contient que 4 items de moyenne et de faible qualité. Si nous nous fions aux critères proposés par Ravand et Widhiarso (2013), il y a 24 items de bonne qualité et 11 items de faible qualité en termes de pseudo-chance.

La moyenne du paramètre d'étourderie est de 0,2198; autrement dit, en moyenne, l'élève a 21,98% de chance de donner des réponses incorrectes même s'il maitrise tous les attributs exigés. En nous basant sur les critères de de la Torre (2009) et de la Torre et Douglas (2008), il y a au total 13 items qui sont classés comme étant de bonne qualité, dont la plupart des items (10 sur 13) se trouvent dans la première partie du test. Les items de bonne qualité en termes de pseudo-chance sont les suivants: 1; 2; 3; 6; 8; 9; 10; 11; 12; 13; 32; 33; 34. Il y a 4 items qui sont classés comme de moyenne qualité: 14; 17; 19; 22. Les sept items suivants sont de faible qualité: 5; 21; 23; 25; 28; 29; 34. Finalement, les 11 items qui restent sont considérés comme problématiques en termes d'étourderie. Cependant, les seuils proposés par Ravand et Widhiarso (2013) nous suggèrent qu'il y a seulement trois items de faible qualité et 32 items de bonne qualité.

Tableau 24. Estimation des paramètres d'items avec la matrice Q2_finale et avec DINA

Item	g	s.e	s	s.e	s+g
1	0,6775	0,0110	0,0466	0,0049	0,7241
2	0,7340	0,0087	0,0612	0,0058	0,7952
3	0,5769	0,0125	0,0865	0,0053	0,66 34
4	0,3088	0,0105	0,3246	0,0100	0,6334
5	0,2295	0,0088	0,2741	0,0114	0,5036
6	0,5901	0,0124	0,0898	0,0053	0,6799
7	0,2338	0,0113	0,3441	0,0088	0,5779
8	0,5538	0,0129	0,0798	0,0058	0,6336
9	0,7147	0,0088	0,0525	0,0053	0,7672
10	0,7160	0,0097	0,0083	0,0022	0,7243
11	0,5624	0,0110	0,0693	0,0058	0,6317
12	0,4194	0,0111	0,0981	0,0068	0,5175
13	0,6917	0,0107	0,0152	0,0031	0,7069
14	0,4405	0,0097	0,1383	0,0088	0,5788
15	0,1958	0,0077	0,3179	0,0121	0,5137
16	0,2708	0,0088	0,3898	0,0121	0,6606
17	0,6136	0,0123	0,1027	0,0057	0,7163
18	0,3947	0,0109	0,3441	0,0092	0,7388
19	0,5280	0,0127	0,1384	0,0065	0,6664
20	0,0862	0,0060	0,5939	0,0109	0,6801
21	0,3806	0,0109	0,2656	0,0095	0,6462
22	0,5345	0,0127	0,1614	0,0069	0,6959
23	0,3923	0,0110	0,2013	0,0088	0,5936
24	0,0000	0,0066	0,5207	0,0117	0,5207
25	0,2618	0,0090	0,2390	0,0105	0,5008
26	0,0265	0,0035	0,4474	0,0118	0,4739
27	0,4428	0,0111	0,3008	0,0098	0,7436
28	0,2642	0,0091	0,2885	0,0109	0,5527
29	0,2764	0,0117	0,2879	0,0085	0,5643
30	0,0425	0,0046	0,7207	0,0107	0,7632
31	0,0117	0,0035	0,0189	0,0032	0,0306
32	0,1202	0,0074	0,0606	0,0048	0,1808
33	0,1602	0,0083	0,0777	0,0053	0,2379
34	0,0339	0,0042	0,1999	0,0078	0,2338
35	0,2643	0,0115	0,3271	0,0088	0,5914
Moyenne	0,3643		0,2198		0,5841

La somme des moyennes des paramètres de pseudo-chance et d'étourderie est de 0,5841, ce qui fait qu'en moyenne, la capacité diagnostique des items est de 0,4259.

Selon les seuils proposés par de la Torre (2009) et de la Torre et Douglas (2008), il y a seulement deux items (31 et 32) qui sont classés de bonne qualité avec g+s entre 0 à 0,2, autrement dit, ces deux items ont à la fois des paramètres de pseudo-chance et d'étourderie de bonne qualité. Il y a également deux items qui sont de moyenne qualité (33 et 34) avec la somme des paramètres de pseudo-chance et d'étourderie qui varie entre 0,2 et 0,4. Douze items sont classés comme de faible qualité avec la somme de g+s qui est entre 0,4 et 0,6. Finalement, il y a 19 items qui sont encore problématiques avec la somme de g+s qui dépasse 0,6. La plupart de ces items se trouvent dans la première partie du test. Cependant, selon les seuils proposés par Ravand et Widhiarso (2013), il y a 16 items de bonne qualité et 19 items de faible qualité. La figure 11 présente l'estimation des paramètres et la capacité diagnostique des items. La ligne verte renvoie à la capacité diagnostique de 35 items du PIRLS 2011. Plus la ligne est élevée, meilleure est la qualité diagnostique des items.

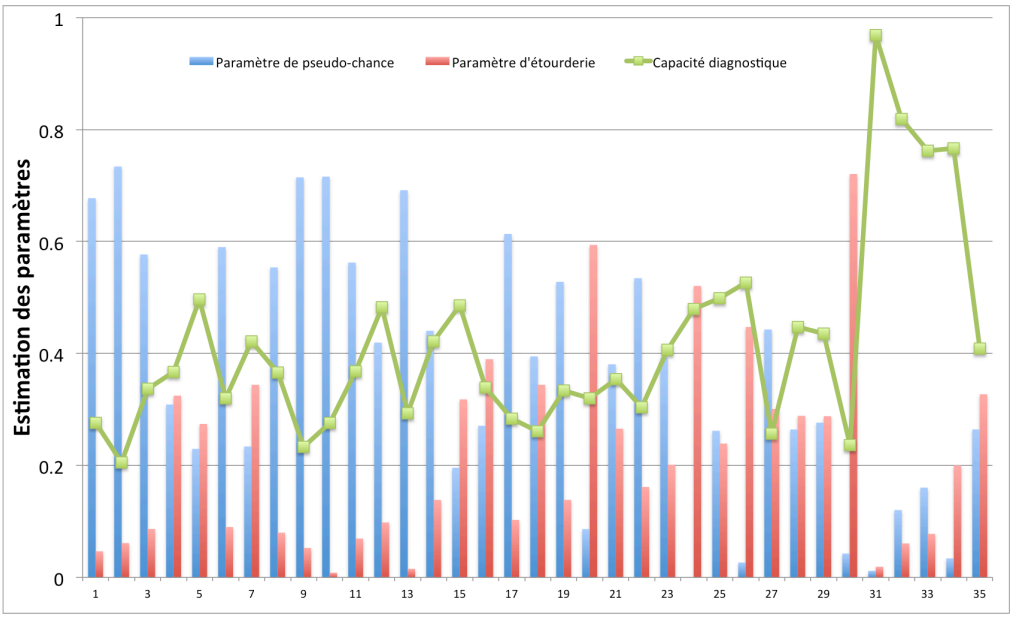


Figure 11. Estimation des paramètres d'items et leur capacité diagnostique

En somme, les résultats issus des modélisations nous suggèrent que la qualité diagnostique des items du PIRLS 2011 est moyenne. La deuxième partie du test semble avoir une meilleure qualité diagnostique des items que la première partie du test. Les explications de ces résultats sont liées aux caractères des passages ainsi qu'aux types de

questions posées, ce dont nous discuterons plus en détail dans le chapitre 5. Une des suggestions pour améliorer la qualité de ces items serait de procéder à une autre étape de raffinement de la matrice Q, ce qui ne sera pas fait dans cette thèse. La prochaine partie est consacrée à présenter les profils de maitrise des habiletés de 4762 élèves canadiens ayant pris le livret 13 du PIRLS 2011.

4.2.3. Profils de maitrise des habiletés des élèves

À titre de rappel, cinq attributs sont identifiés pour le test : *(A1) Repérer des informations explicites; (A2) Compréhension globale; (A3) Interprétation; (A4) Faire des inférences; (A5) Vocabulaire et syntaxe*, nous avons obtenu au total 2^5 profils possibles d'élève, donc 32 profils d'élèves. Le tableau 25 présente les 32 profils avec le pourcentage d'élèves correspondant pour chaque profil. Le profil le plus fréquent des élèves que nous avons analysés est «11110» avec 25,24%, c'est à dire qu'il y a 25,24% de ces élèves qui maitrisent les quatre premiers attributs, mais pas le cinquième. Vient ensuite le profil des élèves qui ne maitrisent aucun des cinq attributs (18,78%). Le profil qui arrive au troisième rang est celui qui maitrise tous les attributs, soit 16,13% des élèves. Le profil «10011» est aussi présent parmi nos participants avec 9,11%. Avec ce profil, l'élève a bien maitrisé l'attribut *Repérer des informations implicites* (A1); l'attribut *Faire des inférences* et *Vocabulaire et syntaxe*, mais pas *Compréhension globale et Interpréter*. Le profil «00011» suit très près avec 8,89%. Aucun participant ne fait partie d'un des 4 profils suivants : «10000»; «11000»; «01001»; «11010». Nous donnerons plus de d'explications sur ces résultats dans le chapitre 5.

Tableau 25. Profil des élèves et des pourcentages des participants

Profil	Pourcentage d'élèves	Profil	Pourcentage d'élèves
00000	18,78	11100	2,06
10000	0	11010	0
01000	0,35	11001	0,01
00100	4,12	10110	1,97
00010	0,40	10101	0,47
00001	2,13	10011	9,11
11000	0	01110	0,55
10100	2,3	01101	0,17
10010	0,15	01011	0,06
10001	0,19	00111	0,03
01100	2,19	11110	25,24
01010	0,12	11101	0,24
01001	0	11011	3,18
00110	00,48	10111	0,32
00101	0,37	01111	0,01
00011	8,89	11111	16,13

Les résultats des modélisations nous fournissent également le profil détaillé de chaque élève sur les probabilités de maitrise pour les habiletés. Le tableau 26 présente l'exemple des profils détaillés de cinq élèves choisis aléatoirement de notre base de données. L'élève 1 ne maitrise aucune des cinq habiletés, car les probabilités sont toutes égales à 0. L'élève 2 maitrise les habiletés A1 et A3 à 100%, l'habileté A3 à 90%, l'habileté A4 à 80% et l'habileté A5 à 0%. L'élève 3, quant à lui, maitrise les habiletés A1 et A3 à 100%, l'habileté A2 à 60%, l'habileté A4 à 80% et l'habileté à 0%. L'élève 4 maitrise les trois habiletés A1; A3 et A4 à 100% tandis que les habiletés A2 et A5 sont maitrisées respectivement à 70% et 0%. Finalement, les habiletés A1; A3; A4 et A5 sont maitrisées à 100% par l'élève 5 alors qu'il maitrise l'habileté A2 à 90%.

Tableau 26. Exemple des profils détaillés de 5 élèves de notre base de données

Élève	A1 Repérer des informations explicites	A2 Compréhension globale	A3 Interprétation	A4 Faire des inférences	A5 Vocabulaire et syntaxe
1	0	0	0	0	0
2	1	0,9	1	0,8	0
3	1	0,6	1	0,8	0
4	1	0,7	1	1	0
5	1	0,9	1	1	1

La figure 12 présente les probabilités de maitrise des cinq habiletés pour l'ensemble des élèves. L'habileté *Faire les inférences* est la mieux maitrisée avec 66,62%, vient ensuite l'habileté *Repérer des informations implicites* qui est maitrisée à 61,31 %. L'habileté *Interpréter* est classée au troisième rang avec 56,64% des probabilités de maitrise. L'habileté *Compréhension globale* est maitrisée par 50,31% et finalement, l'habileté *Vocabulaire et syntaxe* est celle qui est la moins maitrisée avec seulement 41,31%. Ces profils détaillés des élèves ainsi que les probabilités de maitrise des habiletés du groupe ont fait l'objet du processus de l'élaboration des rapports avec le panel d'experts dont nous présentons les résultats dans la partie suivante.

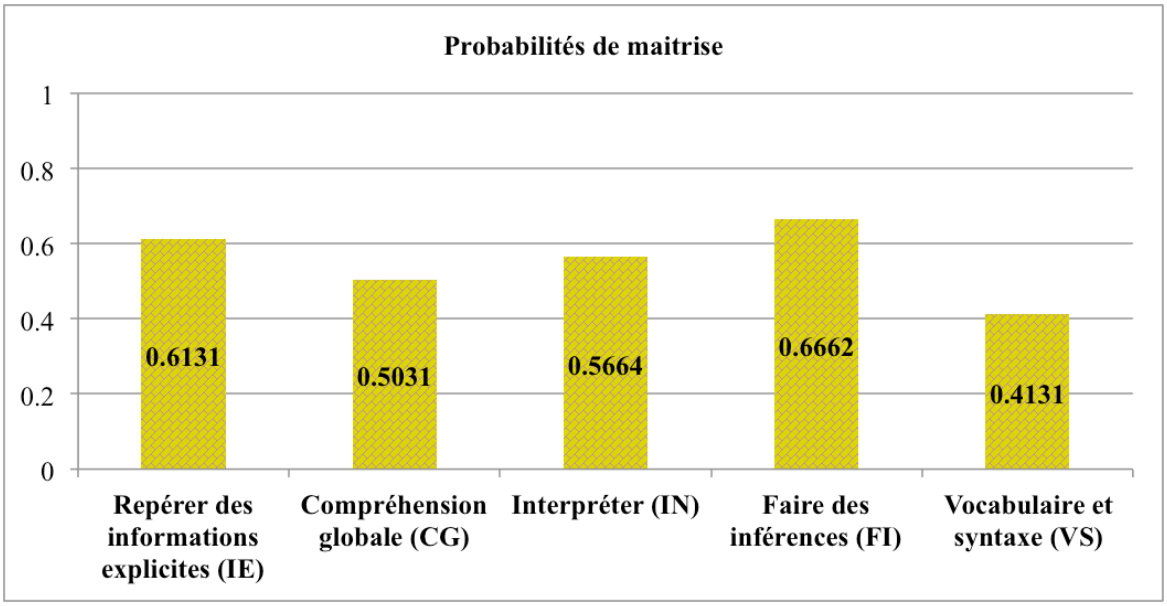


Figure 12. Probabilités de maitrise des habiletés pour l'ensemble des élèves

4.3. Élaboration des rapports diagnostiques avec le panel d'experts

Cette partie est consacrée à la présentation des résultats du processus de l'élaboration des rapports diagnostiques avec le panel d'experts. Nous présentons, dans un premier temps, la reformulation des habiletés en lecture proposées par le panel d'experts. Deuxièmement, nous discutons du choix du profil-type à présenter dans les formats de rapports. Finalement, nous décrivons les trois formats de rapports élaborés à partir du profil choisi avec le panel d'experts comprenant cinq membres, à savoir: (1) une enseignante au primaire, (2) une conseillère pédagogique, (3) la responsable d'une firme de testing, (4) une experte en mesure et évaluation et (5) la chercheuse.

4.3.1. Reformulation des habiletés en lecture

Lors de la première rencontre, nous avons examiné cinq habiletés identifiées pour le test et nous les avons reformulées dans un langage accessible pour les enseignants. En se basant sur « La progression des apprentissages » en lecture proposée par le MELS et le cadre de référence de l'élaboration du PIRS du CMEC (2012), les membres du panel d'experts ont suggéré de remplacer le mot « Localiser » dans la définition de l'habileté *Repérer des informations explicites* par le mot « extraire », car ils trouvent que le terme « Localiser » est très anglophone (localize). La définition de l'habileté *Interpréter* a suscité des discussions entre les membres du panel d'experts, car selon eux, dans le document du MELS, cette habileté est enseignée en partie en 4e année, mais n'a pas à être évaluée. Ainsi, si nous présentons la définition proposée par les experts du panel 1, il y a une forte possibilité que les enseignants ne comprennent pas. Les experts ont constaté également que cette définition est encore très générale et devrait être bien précisée pour ne pas porter à confusion pour les enseignants. Nous avons donc recouru à la définition de l'habileté proposée par le CMEC (2012) dans le cadre de référence de l'élaboration du PIRLS 2011 tout en sachant qu'elle ne change pas la nature de cette habileté. En ce qui concerne l'habileté *Faire des inférences*, les membres considèrent que sa reformulation n'est pas encore claire, car elle a été traduite à partir de la définition en anglais. Nous l'avons donc simplifiée en nous basant sur la définition du CMEC (2012). Quant aux définitions de l'habileté *Compréhension globale* et de *Vocabulaire et syntaxe*, les membres du panel d'experts les ont trouvées claires et correspondant bien à celles

proposées par le MELS. Le tableau 27 présente les définitions de ces cinq habiletés proposées par le panel d'experts lors de l'élaboration de la matrice Q (panel 1) et celles du panel d'experts lors du développement des rapports (panel 2).

Tableau 27. Définitions des habiletés proposées par deux panels d'experts

Attribut	**Définitions proposées par le panel 1**	**Définitions proposées par le panel 2**
(A1) Repérer des informations explicites	Localiser et reconnaitre des informations explicites exprimées dans le texte pour répondre aux questions	**Extraire** et reconnaitre des informations explicites exprimées dans le texte pour répondre aux questions.
(A2) Compréhension globale	Former une compréhension globale d'un paragraphe ou de l'ensemble du texte	Former une compréhension globale d'un paragraphe ou de l'ensemble du texte.
(A3) Interprétation	Clarifier le sens des idées ou des configurations complexes et interpréter des relations	**Acquérir une compréhension plus approfondie du texte en combinant les connaissances antérieures et les informations présentées dans le texte. Les liens à établir ne sont pas seulement implicites; ils peuvent également être ouverts à l'interprétation de la lectrice ou du lecteur (CMEC, 2012).**[2]
(A4) Faire des inférences	Comprendre les informations qui ne sont pas explicitement exprimées en faisant des inférences ou des prédictions	**Déduire des informations justes à partir des indices du texte**.[3]
(A5) Vocabulaire et syntaxe	Exprimer des idées dans une grammaire correcte et compréhensible de l'anglais écrit	Exprimer des idées dans une grammaire correcte et compréhensible de l'anglais écrit.

[2] Les caractères gras présentent la définition adaptée du CMEC (2012) pour l'habileté *Interprétation*

[3] Les caractères gras présentent la définition adaptée du CMEC (2012) pour l'habileté *Faire des inférences*

4.3.2. Choix d'un profil-type

Après avoir reformulé les habiletés en lecture, nous avons décidé de choisir un profil-type parmi les profils de maitrise des élèves et de proposer différents formats de présentation pour ce profil, ce qui rend ces formats plus comparables. Ainsi, afin de choisir un profil-type, le tableau de synthèse des profils des élèves ainsi que les probabilités de maitrise obtenues à partir des modélisations ont été fournis au panel d'experts (voir tableau 29). Dans ce tableau synthèse, nous avons retenu les 10 profils correspondant aux plus grands pourcentages de participants. Les profils 01; 02; 03 sont des exemples des profils plats tandis que les sept profils qui restent (en gras) sont des profils non-plats. Les profils plats sont ceux qui se composent d'une série consécutive de 1 ou des 0 (par exemple, «00000») tandis que dans les profils non-plats, les 1 et les 0 sont combinés d'une manière alternée (par exemple, «10011»).

Les membres du panel d'experts se sont mis d'accord sur le principe qu'il est plus intéressant de présenter un profil non-plat dans un rapport diagnostique, car ce profil peut fournir aux enseignants plus d'informations diagnostiques ainsi que de la variété sur l'aspect visuel. Nous avons donc éliminé les trois profils plats dans la liste des choix de notre profil-type. Pour chacun de sept profils retenus, nous avons choisi aléatoirement un élève dont nous avons produit les graphiques et les tableaux récapitulatifs de réponse avec des habiletés identifiées à chaque question. Ces graphiques ainsi que les tableaux de réponse de sept profils ont été examinés par les experts afin de choisir un profil jugé le plus représentatif d'élèves dans une classe.

Les experts ont classé ces profils en deux catégories: probable et peu probable selon le fait que ce profil ressemble beaucoup ou peu à un cas spécifique de leur classe. L'exemple du profil 4 est un élève qui maitrise bien les habiletés *Repérer des informations explicites; Faires des inférences et Vocabulaire et syntaxe*, mais pas les habiletés *Compréhension globale et Interprétation.* Selon les experts, il s'agit d'un profil probable, parce que la *Compréhension globale* est une des habiletés difficiles en lecture, surtout pour un élève en 4e année. Il peut arriver que des élèves puissent donner des réponses à des questions plus pointues, mais pas aux questions globales.

Par contre, le profil 5 est un profil peu probable, parce qu'il est rare que l'élève puisse bien maitriser l'habileté *Faire des inférences* sans maitriser l'habileté *Repérer des informations explicites*. Il est évident que cela dépend également la manière dont la question a été posée et du type de texte. Dans le cas du PIRLS, la répartition des questions sur l'habileté *Faire des inférences* est assez équitable dans les deux textes, les experts en concluent qu'il est rare de le trouver chez leurs élèves.

Tableau 28. Synthèse de 10 profils les plus populaires avec des exemples

	Profils		Exemple de profils d'élève				
Numéro du profil		%	A1 Repérer des informations explicites	A2 Compréhension globale	A3 Interprétation	A4 Faire des inférences	A5 Vocabulaire et syntaxe
01	11110	25,24	0,7	0,6	1,0	0,5	0,0
02	00000	18,78	0,0	0,0	0,0	0,2	0,1
03	11111	16,13	0,8	1,0	1,0	0,9	1,0
04 (64)[4]	**10011**	**9,11**	**0,8**	**0,3**	**0,0**	**1,0**	**0,8**
05 (220)	**00011**	**8,89**	**0,3**	**0,0**	**0,0**	**1,0**	**0,9**
06 (376)	**11011**	**3,18**	**0,7**	**0,6**	**0,0**	**0,8**	**0,9**
07 (2184)	**10100**	**2,30**	**0,7**	**0,4**	**1,0**	**0,3**	**0,0**
08 (1602)	**01100**	**2,19**	**0,2**	**0,8**	**1,0**	**0,1**	**0,0**
09 (357)	**11100**	**2,06**	**0,6**	**0,5**	**1,0**	**0,2**	**0,0**
10 (1951)	**10110**	**1,97**	**0,9**	**0,3**	**1,0**	**0,8**	**0,0**

[4] Le caractère gras présente les profils non-plats avec les probabilités de maitrise des attributs

Le profil 6 est plus ou moins semblable au profil 4: l'élève a bien maitrisé les habiletés *Repérer des informations implicites; Compréhension globale et Faires des inférences ainsi que Vocabulaire et syntaxe*, mais pas l'habileté *Interprétation.* Ce profil a suscité des discussions intéressantes entre les membres du panel d'experts. Selon la définition du MELS et du CMEC (2012), l'habileté *Faire des inférences* semble plus facile que *l'Interprétation*, car faire des inférences peut amener les élèves à donner les mêmes réponses en se basant sur les éléments du texte. Ce qui n'est pas le cas pour *l'Interprétation*, car chaque élève peut arriver à une réponse différente, l'importance est ses arguments. Ce qui fait que *l'Interprétation* est une habileté plus complexe que *Faire des inférences* selon « La progression des apprentissages » du MELS. Cette idée est aussi cohérente avec les définitions qui se trouvent dans le cadre de référence du PIRLS 2011 et qui stipulent que *l'Interprétation* vise à « acquérir une compréhension plus approfondie du texte en combinant les connaissances antérieures et les informations présentées dans le texte » alors que *Faire des inférences* se définit comme « Combler les lacunes relatives au sens en déduisant des informations à partir du texte ». Il est donc évident que l'élève maitrise bien l'habileté *Faire des inférences,* mais moins l'habileté *Interprétation*. Ce profil est aussi probable selon les experts. En se basant sur les mêmes arguments, les experts trouvent que le profil 7 est peu probable, étant donné qu'il est rare que l'élève puisse maitriser l'*Interprétation* mieux que l'habileté *Faire des inférences*.

Le profil 8 est aussi un profil peu probable, car selon les experts, à la fin de la 4ème année, l'élève devrait bien maitriser les habiletés *Repérer des informations explicites* et *Faire des inférences,* ce qui n'est pas le cas de cet élève. Par contre, l'élève maitrise mieux la *Compréhension globale et l'Interprétation* que le *Vocabulaire et Syntaxe*, ce qui leur semble un peu rare en salle de classe. Pour le profil 9, l'élève maitrise bien *Repérer des informations explicites*, et l'*Interprétation*, mais moins la *Compréhension globale, Faire des inférences et Vocabulaire et syntaxe*. Un des arguments est que peut-être cet élève réussit bien à des questions à choix multiples, mais moins bien à des questions de développement. Il s'agit donc d'un profil probable selon eux.

Finalement, pour le profil 10, l'élève maitrise bien *Repérer des informations explicites; Interprétation et Faire des inférences*, mais moins bien *la Compréhension*

globale et le Vocabulaire et syntaxe. Les experts croient que cela peut arriver à des élèves qui comprennent les idées du texte, mais ne les expriment pas dans une langue correcte. Il est aussi probable que l'élève réussisse à choisir des réponses à choix multiples, mais pas à répondre à des questions de développement. Ce profil est également probable.

Ainsi, parmi les profils probables, nous avons retenu le profil 4, car il est le plus représentatif des élèves. Selon les experts, ce qu'on attend d'un élève en 4ème année, c'est qu'il maitrise très bien *Repérer des informations explicites* et *Faire des Inférences*, mais moins la *Compréhension Globale; l'Interprétation*. Quant à l'habileté *Vocabulaire et syntaxe*, il s'agit d'une habileté assez large à définir, car cela renvoie à la capacité de s'exprimer des idées par l'écrit. La maitrise de cette habileté dépend beaucoup du type des questions posées, et aussi de l'ordre dans lequel ces questions sont posées. Si ces questions de développement sont placées à la fin du test, cela peut influencer des réponses des élèves à cause de la fatigue. Le profil 4 correspond bien aux attentes des experts sur un profil typique de la classe, c'est donc lui qui va être présenté dans nos formats de rapport.

4.4.3. Trois formats de rapports

Lors de la deuxième rencontre, en nous inspirant du protocole de design fourni aux experts, nous avons discuté des éléments à mettre dans les trois formats de rapports à partir de ce même profil. Les membres du panel d'experts se sont mis d'accord sur le fait que ces formats de rapports doivent être simples, clairs et interprétables pour les participants afin de souligner leur sentiment de compétence tout en contenant des informations complexes que nous avons obtenues des modélisations. Nous sommes passés à travers les quatre parties du rapport proposées dans le protocole du design pour discuter des éléments à y présenter.

Dans la partie sur l'identification de l'élève, les experts se sont mis d'accord qu'il est peu important de mettre le nom de l'enseignant, il suffit de mettre le code du groupe et de l'école. Par contre, un des membres du panel suggère qu'il est nécessaire d'ajouter la date de naissance de l'élève, ce qui peut nous renseigner sur son niveau de retard. Le sexe de l'élève est aussi une information nécessaire à considérer dans cette partie. En ce qui concerne les directives pour les rapports diagnostiques, les experts croient que les objectifs du rapport devraient être déplacés à la deuxième page (guide d'interprétation du rapport) en ajoutant une phrase qui réfère à cette page afin d'éviter la surcharge des informations présentées dans la première page.

En ce qui concerne la deuxième partie portant sur le profil de l'élève, les experts ont mis l'accent sur l'importance d'intégrer le profil du groupe en parallèle avec le profil individuel de l'élève afin de fournir aux enseignants un portrait global de leur groupe-classe. Les formats de présentations devraient être variés avec les tableaux et des graphiques de différentes couleurs pour présenter les niveaux de maitrise des habiletés. Afin d'assurer la simplicité et la familiarité ainsi que l'interprétabilité des formats de présentation, les experts ont choisi le diagramme en barres comme type de graphique à mettre dans ces rapports. Les habiletés devraient être classées en ordre de difficulté et présentées en différentes couleurs selon le niveau de maitrise. L'utilisation des étoiles ou des flèches est recommandée pour mieux montrer le positionnement de l'élève par rapport au groupe. Des questions concernant la nécessité de mettre le tableau qui résume les réponses de l'élève au test ont émergé. Cependant, les experts jugent qu'il n'est pas

très pertinent de présenter ce tableau dans le rapport, lorsque les enseignants n'ont pas accès aux épreuves.

Quant au résumé de la performance de l'élève et des pistes d'intervention, les experts suggèrent que cette partie doit permettre de décrire le portrait individuel de chaque élève avec le nombre d'habiletés maitrisées et non-maitrisées, ses points forts et des points à améliorer ainsi que les pistes d'interventions pour réduire des faiblesses. Ces pistes d'interventions doivent être les plus détaillées possibles et doivent être écrites dans un langage accessible aux enseignants. Nous avons donc décidé de nous référer aux documents ministériels tels que *La Progression des apprentissages* et le *Référentiel d'intervention en lecture des élèves de 10 à 15 ans* du MELS pour présenter cette partie.

La quatrième partie du rapport constitue le guide d'interprétation, les membres du panel d'experts pensent qu'il est important de préciser les objectifs du rapport, les descriptions plus détaillées des habiletés identifiées pour le test. Cependant, fournir des exemples des questions pour chaque habileté est jugé peu important, car les enseignants n'ont pas d'accès aux épreuves. Le mode d'emploi pour lire les graphiques doit contenir des exemples concrets pour l'interprétation du profil de l'élève. Les experts proposent également de ne pas présenter très en détail des informations supplémentaires sur les modélisations diagnostiques, car ces informations peuvent rendre les rapports difficiles à comprendre et pour éviter la surcharge d'informations dans le guide d'interprétation. Les lignes suivantes visent à décrire plus en détail ces différentes parties de nos versions finales des rapports.

Chaque format de rapport final contient deux pages. La première page sert à présenter le profil de l'élève tandis que la deuxième page est le guide d'interprétation qui contient des informations supplémentaires pour l'interprétation des résultats. La première page se compose de quatre parties: (1) L'identification de l'élève; (2) les directives pour lire les rapports; (3) la présentation du profil de l'élève et du groupe et (4) la description du profil de l'élève et des pistes d'intervention. Les parties (1), (2) et (4) sont les mêmes dans les trois formats de rapport; ce qui diffère est la partie (3) sur la présentation des graphiques. La première partie contient des informations permettant de personnaliser le rapport de l'élève, par exemple: son nom et son prénom; sa date de naissance; son genre; son niveau; son groupe et son école. La partie (2) contient des informations sur les

directives pour lire les rapports telles que l'objectif du rapport; le nombre d'habiletés identifiées pour le test et le lien avec la deuxième page du rapport.

Dans le format A (voir figure 13), le profil de l'élève et celui du groupe sont présentés dans deux graphiques séparés. Le graphique en barres à gauche présente le profil de l'élève sur les cinq habiletés. Ces barres sont distinguées en trois couleurs selon le niveau de maitrise des habiletés: rouge pour les niveaux de 0 à 0,30; jaune pour les niveaux de 0,31 à 0,70 et vert pour les niveaux de 0,71 et 1,0. Ces trois seuils ont été choisis en nous basant sur les trois catégories proposées par la firme de testing privée de Brisson Legris sur les niveaux de maitrise des élèves présentés dans leurs rapports. La ligne verticale en bleu représente le point de coupure de 0,5 qui est aussi le seuil de coupure proposé par la plupart des recherches avec les MCD pour décider si l'élève maitrise ou non d'une habileté. Autrement dit, si le niveau de maitrise d'une habileté est inférieur à 0,5, cette habileté est considérée comme non-maitrisée. Elle est considérée comme maitrisée si son niveau de maitrise est supérieur ou égal à 0,5. Le graphique en barres à droite présente les pourcentages de maitrise des habiletés du groupe. Chaque barre représente une habileté et se distingue en trois couleurs: rouge pour le groupe ayant les niveaux de 0 à 0,30; jaune pour le groupe avec les niveaux de 0,31 à 0,70 et vert pour le groupe dont les niveaux sont de 0,71 et 1,0. Les chiffres présentés dans les barres sont les pourcentages du groupe. Par exemple, pour l'habileté *Compréhension globale,* 46,4 % des élèves ont un niveau de maitrise de 0 à 0,3 alors que 9,1 % des élèves ont un niveau de maitrise de 0,31 à 0,7 et que 44,5 % des élèves ont un niveau de maitrise de 0,71 à 1.

La partie (4) contient deux sous-parties: (a) la description du profil de l'élève et (b) les pistes d'intervention. Le profil de l'élève a été décrit comme suit: *François*[5] *démontre une compréhension adéquate des faits ou des évènements présentés de manière explicite (Repérer des informations explicites-IE) et implicite (Faire des inférences simples-FI), d'un texte courant et d'un texte littéraire. Il est capable de formuler des réponses aux questions de façon adéquate (Vocabulaire et Syntaxe-VS). François ne démontre pas une compréhension globale (Compréhension globale-CG) et approfondie des textes lui permettant de formuler des interprétations plausibles (Interpréter-IN).* En se basant sur

[5] Il s'agit d'un nom fictif.

les documents ministériels tels que La progression des apprentissages et le référentiel d'intervention en lecture pour les élèves de 10 à 15 ans, les quatre pistes d'intervention suivantes pour l'élève ont été identifiées par les membres du panel d'experts: *Amener l'élève à : (a) cerner l'information importante dans les phrases plus longues et plus complexes (CG); (b) défendre son interprétation personnelle en donnant des raisons (IN);* (c) *surligner les éléments d'information qui appuient son interprétation (IN); (d) vérifier dans le texte s'il n'y pas de contradiction avec l'interprétation retenue (IN).* L'habileté ciblée sous sa forme abrégée explicitée en page 2 pour chaque piste d'intervention est mise entre les parenthèses. Ces deux sous-parties sont identiques dans les trois formats de rapport.

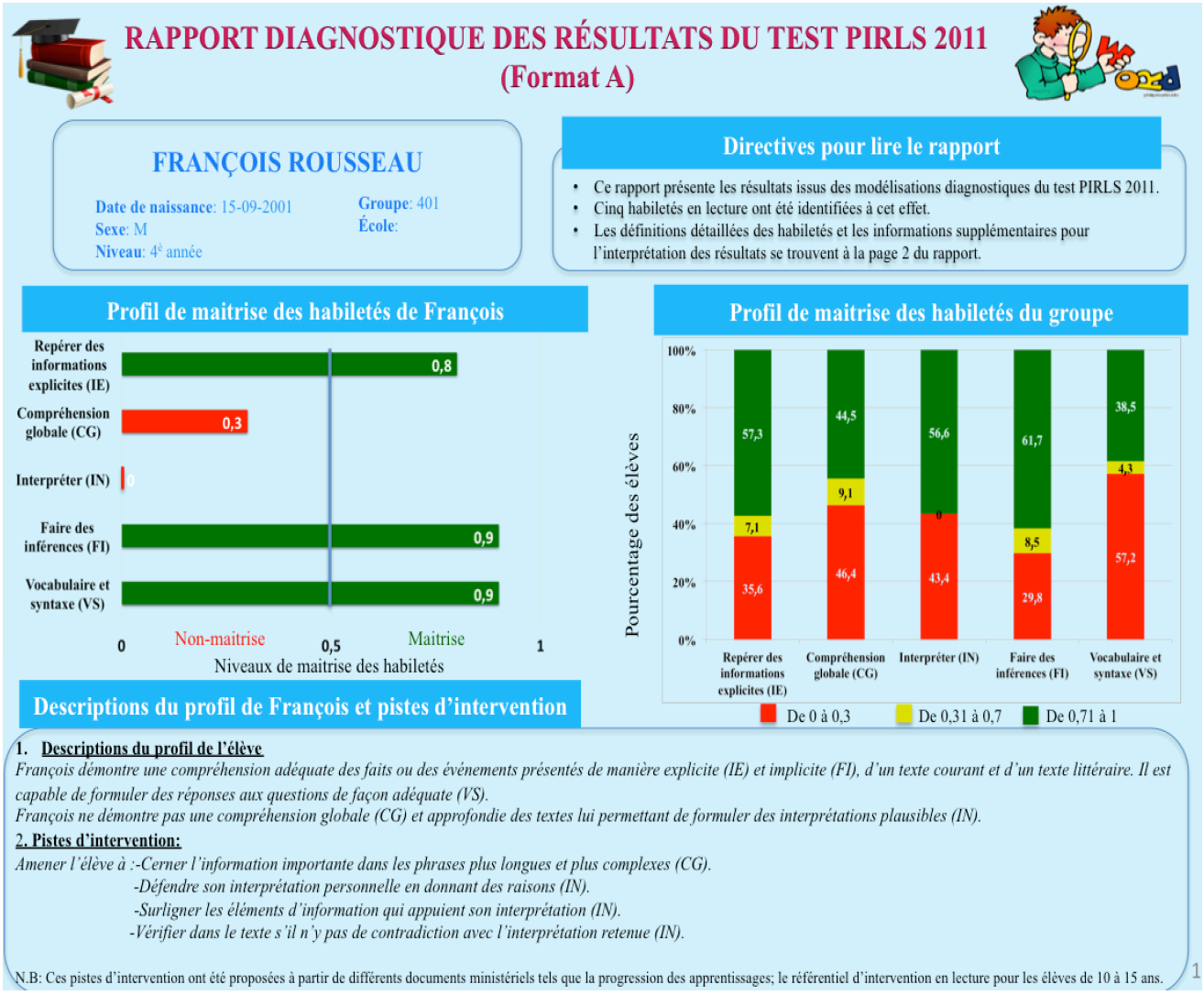


Figure 13. Format A du rapport.

Dans le format B (figure 14), le profil de l'élève et celui du groupe sont présentés dans le même diagramme en barres. Les barres présentent des niveaux de maitrise des habiletés de l'élève et du groupe. La couleur des barres change selon les niveaux de maitrise obtenus: rouge pour les niveaux de 0 à 0,3; jaune pour les niveaux de 0,31 à 0,7; vert pour les niveaux de 0,71 à 1. Les chiffres qui figurent sur les barres correspondent aux niveaux de maitrise des habiletés de François et du groupe. À titre d'exemple, pour l'habileté ***Compréhension globale***, le niveau de maitrise de l'élève est 0,3, donc 30% tandis que celui du groupe est 0,5, ce qui équivaut à 50%. L'élève a donc un niveau de maitrise plus faible que celui du groupe pour cette habileté.

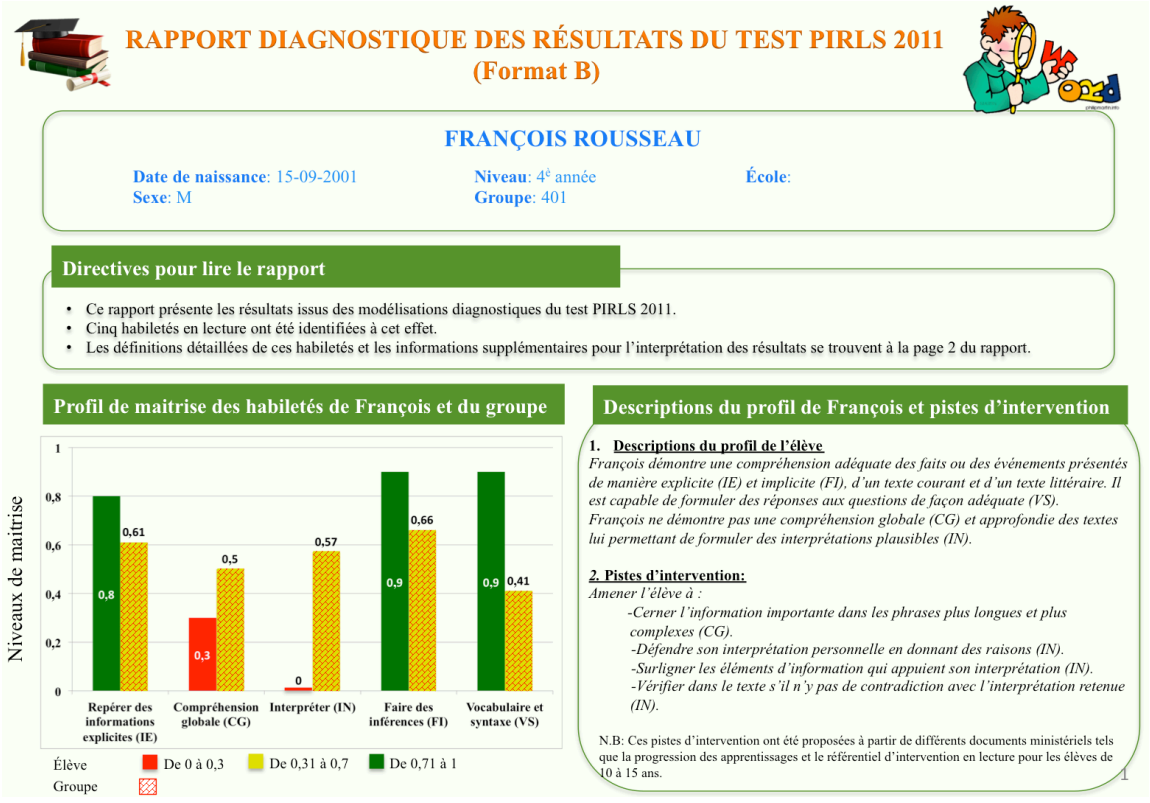


Figure 14. Format B du rapport

Finalement, dans le format C (figure 15), le profil de l'élève et du groupe sont présentés dans le même graphique. Les niveaux de maitrise des habiletés se distinguent en trois couleurs: rouge pour les élèves ayant les niveaux de 0 à 0,3; jaune pour les élèves ayant les niveaux de 0,31 à 0,7; vert pour les élèves ayant les niveaux de 0,71 à 1. Les chiffres qui figurent sur les barres représentent les pourcentages des élèves. Par

exemple, pour l'habileté ***Compréhension globale*** : 46,4 % des élèves ont un niveau de maitrise de 0 à 0,3; 9,1 % des élèves ont un niveau de 0,31 à 0,7 et 44,5 % des élèves ont un niveau de maitrise de 0,71 à 1. Les flèches correspondent au positionnement de l'élève par rapport au groupe. À titre d'exemple, pour l'habileté ***Repérer des informations explicites***, l'élève obtient un niveau de maitrise de 0,8 et fait donc partie du groupe vert.

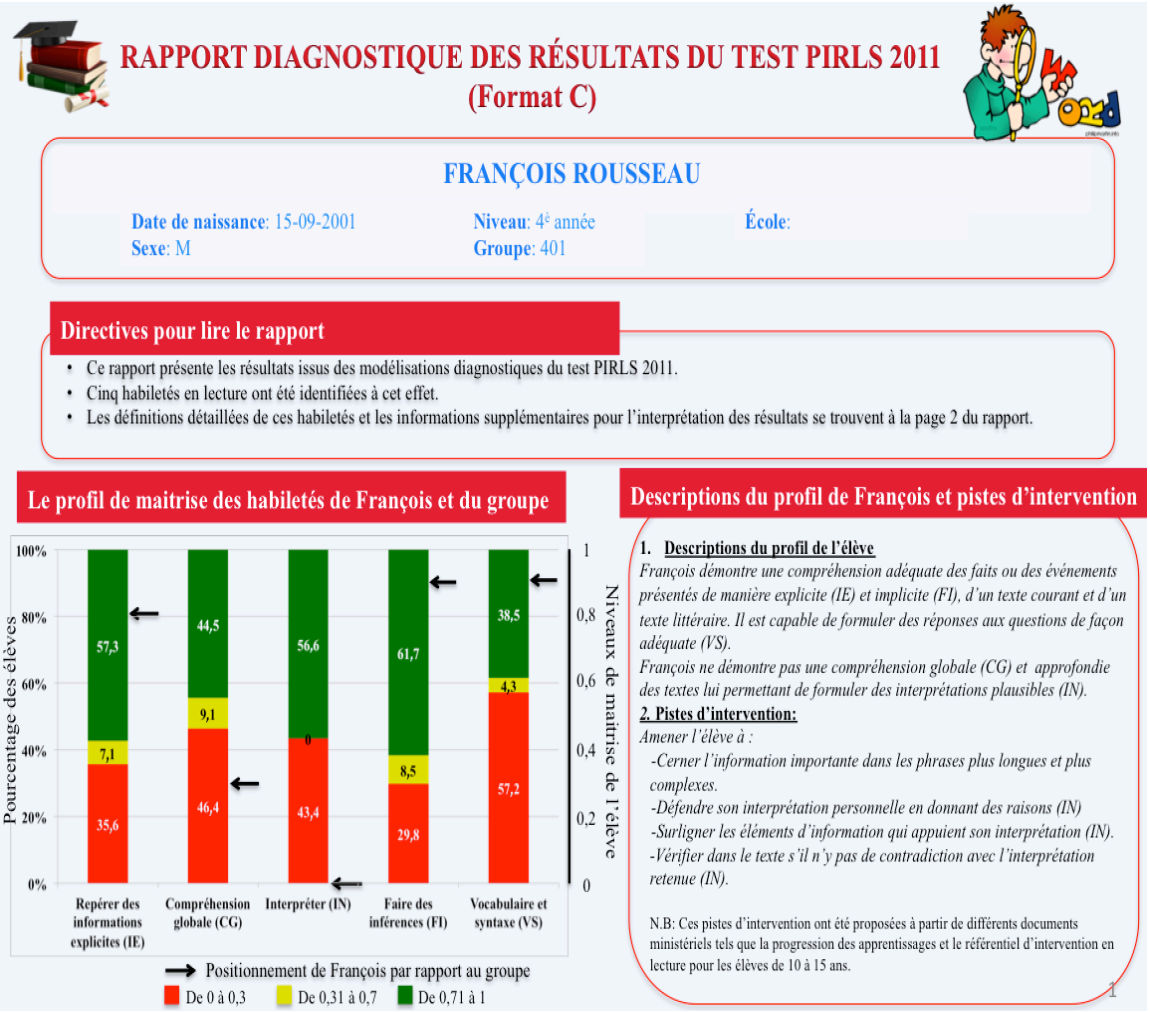


Figure 15. Format C du rapport

La deuxième page des trois formats est le guide d'interprétation qui se compose de trois parties: (a) objectifs du rapport diagnostique; (b) descriptions des habiletés en lecture et (c) mode d'emploi pour lire le graphique. Après avoir discuté avec le panel d'experts, nous concluons que les rapports diagnostiques visent les deux objectifs suivants: (1) dresser un portrait des forces et des points à améliorer de l'élève sur les cinq

habiletés en lecture identifiées pour le test PIRLS 2011 et (2) proposer des pistes d'intervention. Quant à la description des habiletés en lecture, cette partie contient des définitions détaillées sur les cinq habiletés reformulées par le panel d'experts. Le mode d'emploi pour lire le graphique contient des informations sur les modélisations du PIRLS 2011, l'interprétation plus détaillée du terme *niveau de maitrise* avec un exemple ainsi que les directives pour l'interprétation du graphique accompagnée d'un exemple concret. La figure 16 présente le guide d'interprétation du format C. En général, ce guide d'interprétation est le même dans les trois formats, ce qui diffère est le mode d'emploi pour la lecture des graphiques avec des exemples concrets qui sont différents d'un format à l'autre.

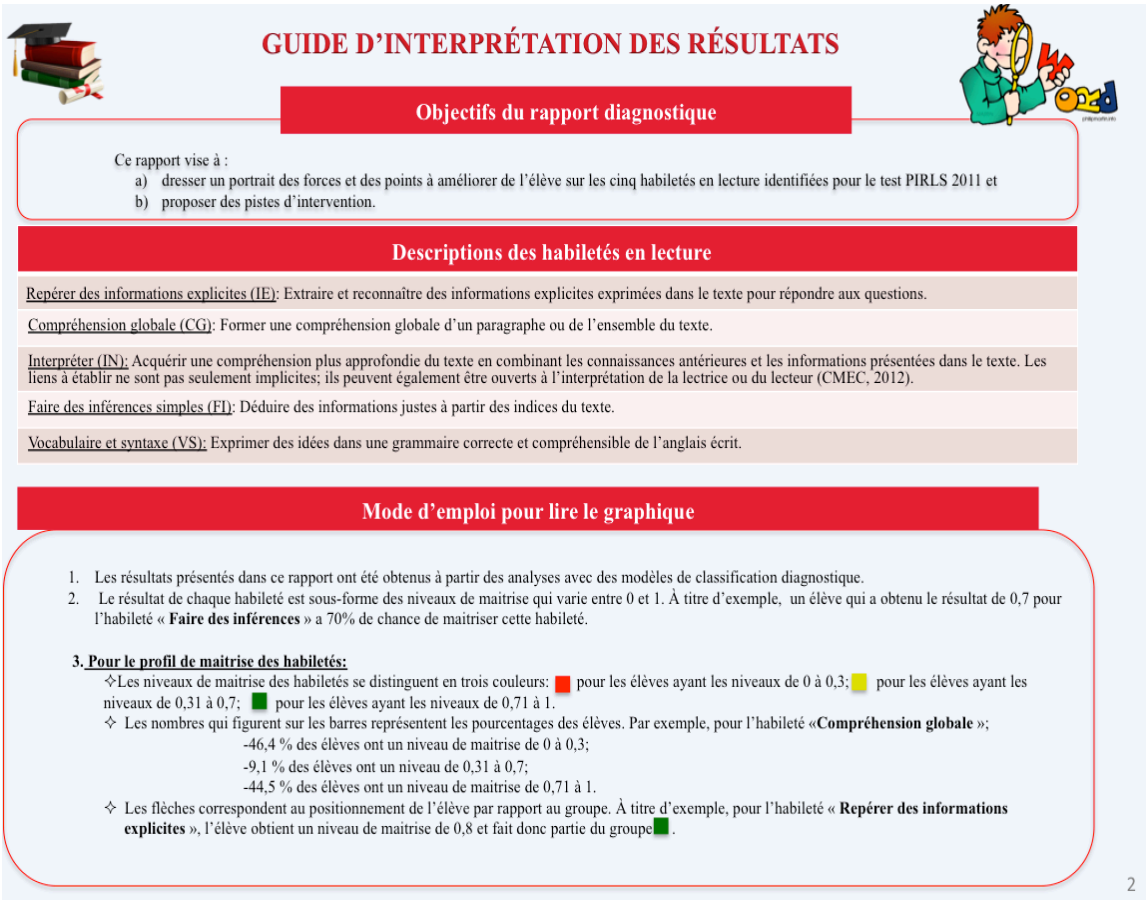


Figure 16. Guide d'interprétation du format C.

Avec ces trois formats de rapports, nous avons conçu un questionnaire (voir annexe 6) pour l'évaluation externe des rapports par des enseignants au primaire, des conseillers pédagogiques et des orthopédagogues. Dans un premier temps, la chercheuse

a conçu la version préliminaire du questionnaire en adaptant les questions utilisées dans la recherche de Zapata-Rivera, Vezzu et VanWinkle (2013) et celles de la recherche de Vezzu, VanWinkle et Zapata (2012) et en inspirant de quatre critères d'un bon rapport suggérés dans la littérature. Cette version préliminaire a été présentée au panel d'experts afin qu'ils puissent y apporter des commentaires. Des suggestions sur la reformulation des questions ont été proposées par les experts. Par exemple, les mots « probabilités de maitrise » doivent être remplacés par « niveaux de maitrise » afin de ne pas créer de la confusion chez les participants. Les experts trouvent également que le mot « point de coupure » est un peu difficile à comprendre dans les questions sur la compréhension et qu'ils doivent donc reformuler autrement. Selon eux, il faudrait également prendre en considération la répartition assez équitable des questions en lien avec les différentes parties du rapport. Avec ces suggestions, la version finale du questionnaire contient 55 questions portant sur quatre aspects: (1) la préférence des rapports; (2) l'évaluation de la qualité des rapports; (3) la compréhension des rapports et (4) les informations sociodémographiques dont les exemples de questions sont présentés dans le tableau 29. La version complète du questionnaire se trouve dans l'annexe 6.

En résumé, la section 4.4.3 sert à présenter les résultats du processus de l'élaboration des rapports avec le panel d'experts. Nous avons présenté, en premier lieu la reformulation de cinq habiletés dans un langage accessible aux enseignants en nous basant sur les documents ministériels. En deuxième lieu, nous avons choisi un profil-type parmi les sept profils les plus populaires présentés au panel d'experts. Le profil choisi est celui qui maitrise bien *Repérer des informations implicites; Faire des inférences et Vocabulaire et syntaxe*; par contre, dans ce profil, l'élève ne maitrise pas *la compréhension globale et l'Interprétation*. Ce profil a été présenté dans les trois formats de rapports selon différentes manières de présenter les résultats du profil de l'élève. Nous avons décrit en détail chacun de ces trois formats ainsi que le guide d'interprétation qui les accompagne. Ces trois formats de rapports sont ensuite évalués par les enseignants au primaire, par les orthopédagogues et par les conseillers pédagogiques à l'aide d'un questionnaire administré en ligne, dont les résultats seront présentés dans la partie suivante .

Tableau 29. Exemples des questions administrées aux enseignants

Dimensions		Description des questions	Exemple des questions	Références
Préférence		-Quatre questions sur le format de représentations préféré (3 choix de réponse pour chaque question)	*-Parmi les trois formats des rapports présentés, lequel préférez-vous?*	Zapata-Rivera, Vezzu et VanWinkle (2013)
Évaluation des rapports	**Accessibilité** **Lisibilité** **Utilité** **Validité**	32 énoncés de type Likert à quatre points (degré d'accord ou désaccord) -1 questions ouvertes	*-Les chiffres présentés dans les barres facilitent le positionnement de l'élève par rapport au groupe.* *-Les niveaux de maitrise des habiletés de l'élève sont faciles à comparer avec ceux du groupe*	Vezzu, VanWinkle et Zapata (2012)
	Compréhension	- 9 questions à choix multiples portent sur certains éléments spécifiques du profil de l'élève	*Dans le graphique sur le profil de maitrise des habiletés du groupe, pour l'habileté* ***Compréhension globale****, quel groupe a le pourcentage le plus élevé?*	
Socio-démographie		-Genre; tranche d'âge, formation, ancienneté, etc.	*-Quel est votre genre?* *-Quelle est votre tranche d'âge?* *-Quel est votre nombre d'années d'expérience?*	

4.4. Perceptions des enseignants au primaire, des orthopédagogues et des conseillers pédagogiques sur les formats de rapports diagnostiques

Cette section vise à décrire les perceptions des enseignants au primaire, des orthopédagogues et des conseillers pédagogiques sur les trois formats de rapports diagnostiques. Dans un premier temps, nous présentons, la description de l'échantillon, ensuite la préférence des participants quant aux trois formats de rapport, le tout en lien avec des variables sociodémographiques. Viennent ensuite l'étude de la dimensionnalité des questions sur l'évaluation de la qualité des rapports et l'analyse des perceptions des participants sur la qualité des rapports en tenant compte des variables sociodémographiques. Finalement, nous présentons l'évaluation que font les participants de leur compréhension de trois formats de rapports. Le lien entre l'évaluation et la compréhension des rapports sera abordé.

4.4.1. Description de l'échantillon

Le sondage a été envoyé par courriel à 27 écoles primaires réparties dans différentes régions comme Montréal, Rive-Nord, Rive-Sud, Outaouais et Québec. Nous l'avons également annoncé sur les réseaux sociaux tels que Linkedln, Twitter et Facebook. Au total, 151 personnes ont accédé au questionnaire. Cependant, 49 participants n'ont rempli que des questions sociodémographiques et 4 participants n'ont pas accordé leur consentement. Nous avons dû éliminer ces participants. Finalement, un total de 98 participants a été retenu pour les analyses, dont 33 participants (33,7%) pour le format A, 32 participants (32,7%) pour le format B et 33 participants (33,7%) pour le format C.

Tableau 30. Répartition des participants selon les formats de rapports

Format	Fréquence	Pourcentage valide	Pourcentage cumulé
Format A	33	33,7	33,7
Format B	32	32,7	66,3
Format C	33	33,7	100,0
Total	98	100,0	

Parmi les 98 participants, 67 participants sont des enseignants au primaire, ce qui représente 68,4%; 15 participants sont des orthopédagogues (15,3%) et 16 participants sont des conseillers pédagogiques (16,3%). La plupart des participants sont des femmes (86,6%) tandis que seulement 13,4% sont des hommes. En ce qui concerne les tranches d'âge, deux tiers des participants se trouvent dans les tranches d'âge entre 30-39 ans (32,7%) et 40-49 ans (33,7%). Le tiers des participants qui restent sont répartis en tranches d'âge de 20-29 ans (16,3%), de 50-59 ans (14,3%) et de 60 ans et plus (3,1%). Vu la répartition inéquitable des participants selon les tranches d'âge, nous les avons regroupés en deux catégories: le groupe de 20 à 39 ans (49%) et celui de 40 ans et plus (51%).

Quant à l'ancienneté, plus d'une moitié des participants ont de 11 à 15 ans d'expérience (26,5%) et de 16 à 20 ans d'expérience (25,5%). Il y a 21,4% des participants ayant de 1 à 5 ans d'expérience et 10,2% des participants avec plus 25 ans d'expérience. Seulement 1% des participants ont moins d'un an d'expérience. Nous avons également recodé cette variable pour avoir trois catégories de réponse: 31,6% des participants ayant moins de 10 ans d'expérience; 52% des participants ayant de 11 à 20 ans d'expérience et 16,3% des participants avec plus de 20 ans d'expérience. En ce qui concerne les régions de provenance, plus de 90% des participants viennent du Québec tandis que 6% des participants proviennent du Nouveau-Brunswick, d'autres provinces du Canada ou même d'autres pays. Ces régions de provenance ont été ensuite regroupées comme Québec (94%) et hors-Québec (6%).

Près d'une moitié des participants (48%) ont obtenu ou sont en voie d'obtention d'un baccalauréat (bac); 23,5% ont suivi des cours en microprogramme ou DESS; 23,5% ont obtenu ou sont en voie d'obtention d'une maitrise et seulement 5,1% ont obtenu un doctorat ou sont en voie de l'obtenir. Nous avons donc regroupé les diplômes en trois catégories: Bac (48%); microprogramme ou DESS (23,5%) et maitrise ou doctorat (28,6%). Quant aux domaines de formation, plus d'une moitié des participants (57%) ont suivi le programme en enseignement au primaire ou préscolaire; 24,5% en adaptation scolaire. Seulement 4,1% des participants ont suivi une formation en psychopédagogie et 14,3% des répondants ont des domaines de formation autres que mentionnés tels que l'enseignement en langue seconde, en didactique du français, etc. Cette variable a été

également recodée en trois groupes: enseignement au primaire ou préscolaire (57,1%); enseignement en adaptation scolaire (24,5%) et psychopédagogie et autres (18,4%). Parmi les personnes interrogées, 77,6 % des participants n'ont pas suivi un cours en méthodes quantitatives pendant leur formation au cours de cinq dernières années tandis que 22,4% des participants en ont suivi un. Le tableau 31 présente la répartition des participants selon les variables sociodémographiques qui ont été utilisées pour des analyses inférentielles ultérieures sur la préférence, sur l'évaluation de la qualité et de la compréhension des rapports.

Tableau 31. Répartition des participants selon les informations sociodémographiques

Informations sociodémographiques		N	Pourcentage
Postes occupés	Enseignant au primaire	67	68,4
	Orthopédagogue	15	15,3
	Conseiller pédagogique	16	16,3
Genre	Féminin	84	86,6
	Masculin	13	13,4
Tranche d'âge	20-29 ans	16	16,3
	30-39 ans	32	32,7
	40-49 ans	33	33,7
	50-59 ans	14	14,3
	60 ans et plus	3	3,1
Années d'expérience	Moins d'un an	1	1,0
	1-5 ans	21	21,4
	6-10 ans	9	9,2
	11-15 ans	26	26,5
	16-20 ans	25	25,5
	21-25 ans	6	6,1
	Plus de 25 ans	10	10,2
Régions de provenance	Nouveau-Brunswick	2	2,0
	Québec	92	93,9
	Autre province du Canada	2	2,0
	Autre pays	2	2,0
Diplôme obtenu	Baccalauréat	47	48,0
	Microprogramme ou DESS	23	23,5
	Maitrise	23	23,5
	Doctorat	5	5,1
Domaine de formation	Enseignement au primaire ou préscolaire	56	57,1
	Enseignement en adaptation scolaire	24	24,5
	Psychopédagogie	4	4,1
	Autre	14	14,3
Cours méthodes quantitatives	Non	76	77,6
	Oui	22	22,4

4.4.2. Préférence des enseignants, des orthopédagogues et des conseillers pédagogiques sur les trois formats de rapports

La deuxième section du questionnaire contient quatre questions sur la préférence des participants selon les trois formats de rapports proposés (voir tableau 32). Pour la première question sur le format de rapport préféré, une moitié de participants a choisi le format B, vient ensuite le format C avec 30,6% des participants et finalement le format A avec 19,4% des participants. En ce qui concerne la deuxième question portant sur le format qui permet de comprendre le mieux les niveaux de maitrise des habiletés de l'élève, c'est également le format B qui est le plus choisi (52%), le format C arrive au deuxième rang avec 29,6% et le format A a été choisi par 18,4% des participants.

Tableau 32. Préférence des participants sur les trois de rapports élaborés

Question	Format A	Format B	Format C
1-Parmi les trois formats de rapports présentés, lequel préférez-vous?	19,4%	**50%**	30,6%
2-Lequel, parmi ces trois formats, vous permet de comprendre le mieux les niveaux de maitrise des habiletés de l'élève?	18,4%	**52,0%**	29,6%
3-Lequel, parmi ces trois formats, vous permet de comprendre le positionnement de l'élève par rapport au groupe?	14,3%	**46,9%**	38,8%
4-Lequel, parmi ces trois formats, pensez-vous que les enseignants vont préférer?	19,4%	**51,0%**	29,6%

La question 3 porte sur le format de rapport permettant de mieux comprendre le positionnement de l'élève par rapport au groupe, le format B est encore une fois le plus choisi par les participants avec 46,9%. Comme dans le cas des deux questions précédentes, le format C arrive au deuxième rang avec 38,8% et le format A est choisi par seulement 14,3% des participants. Selon nous, la raison pour laquelle les formats B et C sont les plus choisis par les participants est que le profil de l'élève et celui du groupe sont présentés sur le même graphique, ce qui n'est pas le cas du format A. Le fait de présenter les deux profils dans le même graphique permet aux participants de mieux

identifier le positionnement de l'élève par rapport au groupe. Nous donnerons plus d'explications à ce propos dans le chapitre de discussion.

La dernière question de cette partie porte sur le format de rapport que, selon les répondants, les enseignants vont préférer de manière générale. C'est là encore le format B qui est le plus choisi, par plus d'une moitié des participants (51%), vient ensuite le format C avec environ 20% de moins, soit 29,6%. Finalement, le format A a été choisi par 19,4%. Si nous les classons selon l'ordre croissant, nous constatons que les participants ont tendance à choisir le format qu'ils préfèrent à titre individuel pour les enseignants, ce qui se justifie par le fait que la répartition des participants selon la question 1 et la question 4 est très semblable.

À la suite de ce constat, nous nous intéressons au lien entre la question 1 et trois autres questions sur la préférence. L'idée est de voir si la réponse des participants à la question 1 peut être reliée à leurs réponses aux trois autres questions. Les tests de khi-deux de Pearson nous suggèrent qu'il y a un lien de dépendance entre les réponses de la question 1 et celles de la question 2 (chi2=136,212, p=0,000); et la question 3 (chi2=81,212; p=0,000) et la question 4 (chi2=180,803, p=0,000). C'est-à-dire que les participants ont tendance à choisir le même format de rapport dans la question 1 que dans les trois autres questions. Par exemple, parmi 19,4% des participants qui ont choisi le format A comme leur préférence dans la question 1, 16,3% ont choisi le même format dans la question 2; 10,4% dans la question 3; 19,4% dans la question 4. Ce même constat est fait quand on observe les résultats des participants qui préfèrent le format B ou C. Le tableau 33 ci-dessous présente le lien entre la préférence des participants pour les questions 1 et les trois autres questions.

Tableau 33. Lien entre la préférence des rapports pour la question 1 et les trois autres questions

Question sur la préférence		**Question 1**					
		A	**B**	**C**	**Total**	**Chi 2**	**P**
		(%)	**(%)**	**(%)**	**(%)**		
2-Lequel, parmi ces trois formats, vous permet de comprendre le mieux les niveaux de maitrise des habiletés de l'élève?	Format A	**16,3**	2,0	1,0	19,4		
	Format B	1,0	**46,9**	2,0	**50,0**	**136,212**	**0,000**
	Format C	1,0	3,1	**26,5**	30,6		
3-Lequel, parmi ces trois formats, vous permet de comprendre le positionnement de l'élève par rapport au groupe?	Format A	**10,2**	3,1	6,1	19,4		
	Format B	4,1	**40,8**	5,1	**50,0**	**81,212**	**0,000**
	Format C	0	3,1	**27,6**	30,6		
4-Lequel, parmi ces trois formats, pensez-vous que les enseignants vont préférer?	Format A	**19,4**	0	0	19,4		
	Format B	0	**49,0**	1,0	**50,0**	**180,803**	**0,000**
	Format C	0	2,0	**28,6**	30,6		

4.4.2.1. Lien entre la préférence des rapports et les postes occupés

Au total, 68,4% des participants sont des enseignants au primaire, 15,3% des participants sont des orthopédagogues tandis que 16,3% des participants sont des conseillers pédagogiques. Vu la répartition inéquitable des participants dans les trois groupes, nous avons décidé de regrouper les conseillers pédagogiques et les orthopédagogues ensemble. L'intérêt ici est de savoir s'il y a un lien entre les postes occupés par des participants et leur préférence quant aux formats de rapports. Chez les enseignants, pour la question 1, il y a 37,3% des participants qui ont choisi le format B, vient ensuite le format C avec 22,4% et finalement le format A avec 12,2%. Pour les orthopédagogues et les conseillers pédagogiques, 16,3 % des participants préfèrent le format B, 8,2 % choisissent le format C et 7,1 % ont une préférence pour le format A. Le test de Khi-deux (Chi2=0,604; p=0,739) montre qu'il n'y a pas de lien de dépendance entre les postes occupés des participants quant à leur préférence des formats de rapports pour la question 1.

Les participants ont tendance à avoir les mêmes préférences pour la question 2 C'est-à-dire que peu importe leurs postes occupés, ils préfèrent d'abord le format B, ensuite le format C et puis le format A. En ce qui concerne la question 3 sur le format de rapport permettant de comprendre mieux les niveaux de maitrise des habiletés de l'élève, les mêmes tendances se trouvent chez les enseignants, c'est-à-dire, ils choisissent en premier lieu le format B (33,7%), puis le format C (25,5%) et enfin le format A (9,2). Cependant, chez les conseillers pédagogiques et les orthopédagogues, ils sont aussi nombreux à choisir le format B (13,3%) et C (13,3%), puis le format A (5,1%). Pour la question 4, les conseillers pédagogiques et les orthopédagogues préfèrent le format B (17,3%), puis les formats A et C (7,1%). Les tendances restent les mêmes chez les enseignants dans cette question,

Les tests de Khi-deux ne sont pas statistiquement significatifs, ni pour la question 2 (Chi2=0,312; p=0,856), ni pour la question 3 (Chi2=0,466; p=0,793), ni pour la question 4 (Chi2=1,121; p=0,571). Il n'y donc pas de lien entre les postes occupés par les participants et leur préférence de formats de rapports pour ces deux questions. Peu

importe le poste occupé, les participants préfèrent toujours le format B. Le tableau 34 résume les tendances dans le choix des formats de rapport préférés et les postes occupés.

Tableau 34. Lien entre la préférence des rapports et les postes occupées

Question sur la préférence		Poste				
		Enseignant (%)	Orthopédagogue et Conseiller pédagogique (%)	Total (%)	Chi 2	p
1-Parmi les trois formats de rapports présentés, lequel préférez-vous?	Format A	12,2	7,1	19,4	170	
	Format B	33,7	16,3	50,0	0,604	0,739
	Format C	22,4	8,2	30,6		
2-Lequel, parmi ces trois formats, vous permet de comprendre le mieux les niveaux de maitrise des habiletés de l'élève?	Format A	12,2	6,1	18,4		
	Format B	34,7	17,3	52,0	0,312	0,856
	Format C	21,4	8,2	29,6		
3-Lequel, parmi ces trois formats, vous permet de comprendre le positionnement de l'élève par rapport au groupe?	Format A	9,2	5,1	14,3		
	Format B	33,7	13,3	46,9	0,466	0,792
	Format C	25,5	13,3	38,8		
4-Lequel, parmi ces trois formats, pensez-vous que les enseignants vont préférer?	Format A	12,2	7,1	19,4		
	Format B	33,3	17,3	51,0	1,121	0,571
	Format C	22,4	7,1	29,6		

4.4.2.2. Lien entre la préférence des rapports et les tranches d'âge

Nous avons recodé la variable des tranches d'âge pour avoir seulement deux catégories: le groupe de 20 à 39 ans (49%) et le groupe de 40 ans et plus (51%). Pour la question 1, les participants sont les plus nombreux à choisir le format B dans les deux groupes d'âge: 27,6% pour le groupe de 20 à 39 ans et 22,4% pour le groupe de 40 ans et plus. Le même pourcentage a été constaté dans le groupe de 20 à 30 ans et celui de 40 ans et plus qui choisissent le format C (15,3%). Finalement, les participants du groupe de 40 ans et plus qui préfèrent le format A est deux fois plus important (13,3%) que celui des 20 à 39 ans (6,1%) qui choisissent le même format. Cependant, le test de Khi-deux (Chi2=3,050; p=0,218) n'est pas significatif, ce qui confirme qu'il n'y a pas de lien entre les tranches d'âge des participants et leur préférence des rapports pour la question 1.

Quant à la question 2, les participants préfèrent davantage le format B, puis le format C et enfin le format A. Pour le groupe qui choisit le format B, les participants de 20 à 39 ans (31,6%) sont plus nombreux que ceux ayant 40 ans et plus (20,4%). Toutefois, pour le groupe qui choisit les formats A (12,2% contre 6,1%) et C (18,4% contre 11,2%), les participants dans la tranche d'âge de 40 ans et plus sont les plus nombreux. Le test de Khi-deux (Chi2=6,024; p=0,049) confirme qu'il y a un lien de dépendance entre les tranches d'âge des participants et le fait qu'ils préfèrent le format de rapport leur permettant de comprendre le mieux les niveaux de maitrise des habiletés de l'élève. Cependant, la probabilité associée au test (0,049) est très proche au seuil de rejet de 0,05.

En ce qui concerne la question 3 sur le format de rapport permettant de mieux comprendre le positionnement de l'élève par rapport au groupe, le format B est celui que les participants du groupe de 40 ans et plus (26,5%) préfèrent, tandis que le groupe de 20 à 39 ans préfère le format C (21,4%). En deuxième lieu, le groupe de 20 à 39 ans choisit le format B (20,4%) tandis que le groupe de 40 ans de plus choisit le format C (17,3%). Toutefois, le test de Khi-deux (Chi2=1,163; p=0,559) confirme qu'il n'y a pas de lien entre les tranches d'âge des participants et leur préférence quant aux rapports permettant de mieux comprendre le positionnement de l'élève par rapport au groupe. Nous observons les mêmes tendances chez les répondants tant pour la question 4 que dans la

question 2. Ce qui diffère est que le test de khi-deux (Chi2=4,13; p=0,127) n'est pas statistiquement significatif, ce qui confirme qu'il n'y a pas de lien de dépendance entre les tranches d'âge et les formats de rapports que les participants pensent que les enseignants vont préférer. Le tableau 35 présente le lien entre les tranches d'âge des participants et leur préférence des rapports.

Tableau 35. Lien entre la préférence des rapports et les tranches d'âge

Question sur la préférence		Age				
		20-39 ans	**40 ans et plus**	**Total**	**Chi 2**	**p**
1-Parmi les trois formats de rapports présentés, lequel préférez-vous?	Format A	6,1	13,3	19,4		
	Format B	27,6	22,4	**50,0**	3,050	0,218
	Format C	15,3	15,3	30,6		
2-Lequel, parmi ces trois formats, vous permet de comprendre le mieux les niveaux de maitrise des habiletés de l'élève?	Format A	6,1	12,2	18,4		
	Format B	31,6	20,4	**52,0**	6,024	**0,049**
	Format C	11,2	18,4	19,6		
3-Lequel, parmi ces trois formats, vous permet de comprendre le positionnement de l'élève par rapport au groupe?	Format A	7,1	7,1	14,3		
	Format B	20,4	26,5	**46,9**	1,163	0,559
	Format C	21,4	17,3	38,8		
4-Lequel, parmi ces trois formats, pensez-vous que les enseignants vont préférer?	Format A	6,1	13,3	19,4		
	Format B	29,6	21,4	**51,0**	4,13	0,127
	Format C	13,3	16,3	29,6		

4.4.2.3. Lien entre la préférence des rapports et l'ancienneté

Les trois groupes préfèrent d'abord le format B, puis le format C et enfin, le format A. Cette tendance est la même pour les quatre questions concernant la préférence. Par exemple, dans la question 1, parmi les 31,6% des répondants ayant moins de 10 ans d'expérience, il y a 18,4% des participants qui choisissent le format B; 9,2% préfèrent le format C ; tandis que le format A a été choisi par seulement 4,1%. Chez les participants ayant de 11 à 20 ans d'expérience, 23,5% des répondants ont choisi le format B; 17,3% ont la préférence pour le format C ; 11,2% des participants préfèrent le format A. Cependant, le test de Khi-deux (Chi2=2,019; p=0,732), qui n'est pas statistiquement significatif, montre qu'il n'y a pas de lien de dépendance entre le fait que les participants préfèrent un rapport plutôt qu'un autre et leur ancienneté en ce qui concerne la question 1.

Nous constatons ces mêmes tendances pour les questions 2; 3 et 4 (voir tableau 37). Pour la question 3, le pourcentage des participants ayant moins de 10 ans d'expérience qui choisissent le format C (12,2%) est plus ou moins égal au pourcentage de ceux qui choisissent le format B (13,3%). Cette tendance est la même pour le groupe ayant de 11 à 20 ans d'expérience: 24,5% optent pour le format B contre 19,4% pour le format C. Tandis que pour le groupe ayant plus de 20 ans d'expérience, le taux des participants qui choisissent le format B (8,2%) est deux fois plus élevé que pour ceux qui ont une préférence pour le format C (4,1%). Les tests de Khi-deux ne sont pas statistiquement significatifs pour la question 2 (Chi2=4,118; p=0,390), pour la question 3 (Chi2=1,591; p=0,810) ou pour la question 4 (Chi2=2,557; p=0,634). Ainsi, peu importe l'ancienneté des participants, leur préférence des rapports est la même.

Tableau 36. Lien entre la préférence des rapports et l'ancienneté

Question sur la préférence		Ancienneté					
		Moins de 10 ans	11-20 ans	20 ans et plus	Total	Chi 2	p
1-Parmi les trois formats de rapports présentés, lequel préférez-vous?	Format A	4,1	11,2	4,1	19,4		
	Format B	18,4	23,5	8,2	**50,0**	2,019	0,732
	Format C	9,2	17,3	4,1	30,6		
2-Lequel, parmi ces trois formats, vous permet de comprendre le mieux les niveaux de maitrise des habiletés de l'élève?	Format A	5,1	9,2	4,1	18,4		
	Format B	20,4	23,5	8,2	**52,0**	4,118	0,390
	Format C	6,1	19,4	4,1	29,6		
3-Lequel, parmi ces trois formats, vous permet de comprendre le positionnement de l'élève par rapport au groupe?	Format A	6,1	6,1	2,0	14,3		
	Format B	13,3	24,5	9,2	**46,9**	1,591	0,810
	Format C	12,2	21,4	5,1	38,8		
4-Lequel, parmi ces trois formats, pensez-vous que les enseignants vont préférer?	Format A	4,1	11,2	4,1	19,4		
	Format B	19,4	23,5	8,2	**51,0**	2,557	0,634
	Format C	8,2	17,3	4,1	29,6		

4.4.2.4. Lien entre la préférence des rapports et le diplôme obtenu

Dans notre échantillon, il y a 48% des participants qui ont obtenu ou qui sont en voie d'obtenir un Bac, 23,5% des participants ont complété ou sont en train de compléter un microprogramme ou un DESS, tandis que 28,6% des participants ont suivi ou suivent actuellement un programme de maitrise ou de doctorat. Pour les questions 1, 2, 4, plus d'une moitié des participants préfèrent le format B dans des pourcentages respectifs de 50,0%, 52,0% et 51,0%. Pour la question 3, un total de 46,9% des répondants préfère le format B. Chez les participants qui ont obtenu ou qui sont en voie d'obtenir un Bac et ceux qui ont une maitrise ou un doctorat, nous constatons qu'un plus grand nombre choisissent le format B, puis le format C. Ce qui est un peu différent pour le groupe de ceux qui ont suivi un microprogramme ou qui ont obtenu un DESS : leur plus grande préférence va toujours vers le format B, mais leur deuxième choix est le format A, non pas le format C comme dans les deux autres groupes. Ces mêmes tendances ont été observées pour les 4 questions.

Toutefois, les tests de Khi-deux réalisés avec la question 1 (Chi2=9,289; p=0,054), la question 2 (Chi2=5,005; p=0,287), la question 3 (Chi2=5,303; p=0,258) et la question 4 (Chi2=8,506; p=0,075) ne sont pas statistiquement significatifs. Nous concluons donc qu'il n'y a pas de lien de dépendance entre le diplôme obtenu et la préférence pour l'un ou l'autre formats de rapports. Les participants avec un Bac et ceux avec une maitrise et un doctorat ont tendance à choisir les formats B et C en préférence, tandis que les répondants avec un microprogramme ou un DESS préfèrent les formats B et A.

Tableau 37. Lien entre la préférence des rapports et le diplôme obtenu

Question sur la préférence		**Diplôme**					
		BAC	**DESS et micro-programme**	**Maitrise et doctorat**	**Total**	**Chi 2**	**p**
1-Parmi les trois formats de rapports présentés, lequel préférez-vous?	Format A	6,1	6,1	7,1	19,4		
	Format B	21,4	15,3	13,3	**50,0**	9,289	0,054
	Format C	20,4	2,0	8,2	30,6		
2-Lequel, parmi ces trois formats, vous permet de comprendre le mieux les niveaux de maitrise des habiletés de l'élève?	Format A	7,1	6,1	5,1	18,4		
	Format B	22,4	14,3	15,3	**52,0**	5,005	0,287
	Format C	18,4	3,1	8,2	29,6		
3-Lequel, parmi ces trois formats, vous permet de comprendre le positionnement de l'élève par rapport au groupe?	Format A	5,1	5,1	4,1	14,3		
	Format B	19,4	13,3	14,3	**46,9**	5,303	0,258
	Format C	23,5	5,1	10,2	38,8		
4-Lequel, parmi ces trois formats, pensez-vous que les enseignants vont préférer?	Format A	6,1	6,1	7,1	19,4		
	Format B	22,4	15,3	13,3	**51,0**	8,506	0,075
	Format C	19,4	2,0	8,2	29,6		

4.4.2.5. Lien entre la préférence des rapports et le domaine de formation

Au total, il y a 57,1% des participants qui ont suivi une formation en enseignement au primaire ou préscolaire, 24,5% des participants ont une formation en adaptation scolaire et 18,4% ont suivi une formation en psychopédagogie ou dans un autre domaine. Les tendances à choisir le format de préférence semblent différentes dans les trois groupes. Dans la question 1, les participants avec une formation en enseignement au primaire ou préscolaire et ceux en adaptation scolaire semblent avoir les mêmes préférences. Ils sont les plus nombreux à choisir le format B (29,6% et 14,3%), puis le format C (20,4% et 7,1%) et enfin le format A (7,1% et 3,1%). Quant au groupe ayant suivi une formation en psychopédagogie et dans un autre domaine, ils ont tendance à choisir le format A (9,2%), puis le format B (6,1%) et enfin le format C (3,1%). Le test de Khi-deux (Chi2=13,663; p=0,008) montre qu'il y a un lien de dépendance statistiquement significative entre la préférence des participants et leur domaine de formation pour la question 1.

Quant à la question 2, les tendances sont les mêmes que pour le groupe en enseignement primaire ou préscolaire et celui en adaptation scolaire. Pour le groupe en psychopédagogie et autres domaines de formation, les participants sont aussi nombreux à choisir le format B que le format A comme préférence (7,1%). Toutefois, le test de Khi-deux (Chi2=7,980; p=0,092) n'est pas significatif : il montre qu'il n'y a pas de lien entre le domaine de formation des participants et leur préférence des rapports.

En ce qui concerne la question 3, les préférences sont les mêmes que pour les questions 1 et 2 en ce qui concerne les participants en enseignement au primaire ou préscolaire et ceux en adaptation scolaire. Ils préfèrent le format B (27,6% et 12,2%), puis le format C (22,4% et 11,2%) et enfin le format A (7,1% et 1,0%). Par contre, le groupe ayant une formation en psychopédagogie et en d'autres domaines sont les plus nombreux à choisir le format B (7,1%%), puis le format A (6,1%) et enfin le format C (5,1%). Le test de Khi-deux (Chi2=7,606; p=0,107), qui n'est pas significatif, confirme qu'il n'y a pas de lien entre la préférence des participants pour l'un ou l'autre format des rapports quant à leur domaine de formation dans cette question.

Finalement, pour la question 4, nous constatons les mêmes tendances à choisir les formats de rapports que pour la question 1. C'est-à-dire que les participants avec une formation en enseignement au primaire ou préscolaire et ceux en adaptation scolaire préfèrent davantage le format B (29,6% et 15,3%), puis le format C (20,4% et 6,1%) et enfin le format A (7,1% et 3,1%). Quant au groupe en psychopédagogie et en autres domaines de formation, ils préfèrent mieux le format A (9,2%), puis le format B (6,1%) et enfin le format C (3,1%). Le test de Khi-deux (Chi2=14,284; p=0,006), qui est significatif, confirme qu'il y a un lien de dépendance entre le domaine de formation et la préférence des rapports dans la question 4. Ces tendances diffèrent entre le groupe des répondants qui ont une formation en enseignement primaire et préscolaire et en adaptation scolaire d'une part et le groupe de répondants qui ont une formation en psychopédagogie ou dans un autre domaine, d'autre part. Le tableau 38 présente ces tendances. Nous donnerons plus d'explications sur ces résultats dans la discussion.

Tableau 38. Lien entre la préférence des rapports et le domaine de formation

Question sur la préférence		Enseigne -ment au primaire	Adaptation scolaire	Formation Psycho-pédagogie et autre	Total	Chi 2	p
1-Parmi les trois formats de rapports présentés, lequel préférez-vous?	Format A	7,1	3,1	9,2	19,4		
	Format B	29,6	14,3	6,1	**50,0**	**13,663**	**0,008**
	Format C	20,4	7,1	3,1	30,6		
2-Lequel, parmi ces trois formats, vous permet de comprendre le mieux les niveaux de maitrise des habiletés de l'élève?	Format A	7,1	4,1	7,1	18,4		
	Format B	29,6	15,3	7,1	**52,0**	7,980	0,092
	Format C	20,4	5,1	4,1	29,6		
3-Lequel, parmi ces trois formats, vous permet de comprendre le positionnement de l'élève par rapport au groupe?	Format A	7,1	1,0	6,1	14,3		
	Format B	27,6	12,2	7,1	**46,9**	7,606	0,107
	Format C	22,4	11,2	5,1	38,8		
4-Lequel, parmi ces trois formats, pensez-vous que les enseignants vont préférer?	Format A	7,1	3,1	9,2	19,4		
	Format B	29,6	15,3	6,1	**51,0**	**14,284**	**0,006**
	Format C	20,4	6,1	3,1	29,6		

4.4.2.6. Lien entre la préférence des rapports et le suivi des cours en méthodes quantitatives

Parmi nos participants, il y a 77,6% des répondants qui n'ont pas suivi des cours en méthodes quantitatives au cours de 5 dernières années, contre seulement 22,4% des participants en ont suivis. Qu'ils aient suivi ou pas les cours en méthodes quantitatives, leur préférence pour l'un ou l'autre des formats de rapports ne semble pas très différente. Dans les deux groupes, ils préfèrent le format B, ensuite le format C et enfin le format A et ceci peut s'observer dans les questions 2 et 4 (voir le tableau 39). Plus d'une moitié des participants manifestent leur préférence pour le format B dans la question 1 (50,0%), la question 2 (52,0%) et la question 4 (51,0%). Près d'une moitié des participants (46,9%) préfèrent le format B dans la question 3.

Cependant, dans le groupe ayant suivi des cours en méthodes quantitatives, la plus grande préférence est toujours le format B (13,3%) et leur deuxième choix est le format A (5,1%), mais pas le format C (4,1%) pour la question 1. Ces mêmes tendances ont été remarquées pour la question 4, même si la différence entre les pourcentages de préférence du format C et celui du format A n'est pas très grande (1%).

Les tests de Khi-deux réalisés avec la question 1 (Chi2=2,064; p=0,356); la question 2 (Chi2=0,723; p=0,697); la question 3 (Chi2=1,677; Chi2=0,432) et la question 4 (Chi2=1,773; p=0,412) ne sont pas statistiquement significatifs ; ils montrent qu'il n'y a pas de lien entre la préférence pour l'un ou l'autre des rapports et le fait d'avoir suivi ou non des cours en méthodes quantitatives. Autrement dit, le fait d'avoir suivi des cours en méthodes quantitatives au cours des cinq dernières années ne joue pas un rôle dans le choix des formats de rapports.

Tableau 39. Lien entre la préférence des rapports et le suivi des cours en méthodes quantitatives

Question sur la préférence		**Cours en méthodes quantitatives**				
		Non	**Oui**	**Total**	**Chi 2**	**p**
1-Parmi les trois formats de rapports présentés, lequel préférez-vous?	Format A	14,3	5,1	19,4		
	Format B	36,7	13,3	**50,0**	2,064	0,356
	Format C	26,5	4,1	30,6		
2-Lequel, parmi ces trois formats, vous permet de comprendre le mieux les niveaux de maitrise des habiletés de l'élève?	Format A	14,3	4,1	18,4		
	Format B	38,8	13,3	**52,0**	0,723	0,697
	Format C	24,5	5,1	29,6		
3-Lequel, parmi ces trois formats, vous permet de comprendre le positionnement de l'élève par rapport au groupe?	Format A	12,2	2,0	14,3		
	Format B	37,8	9,2	**46,9**	1,677	0,432
	Format C	27,6	11,2	38,8		
4-Lequel, parmi ces trois formats, pensez-vous que les enseignants vont préférer?	Format A	14,3	5,1	19,4		
	Format B	37,8	13,3	**51,0**	1,773	0,412
	Format C	25,5	4,1	29,6		

En résumé, la section 4.4.2 est réservée à la description de la préférence des rapports des participants en lien avec des variables sociodémographiques. Dans la plupart des cas, nous constatons que les participants sont les plus nombreux à choisir le format B, ensuite le format C et enfin le format A. Cet ordre de préférence est un peu différent dans le groupe ayant suivi un microprogramme ou un DESS et le groupe avec une formation en psychopédagogie ou dans d'autres domaines : ils ont tendance à choisir le format B, puis le format A et enfin le format C.

En ce qui concerne le lien entre la préférence des rapports et les variables sociodémographiques telles que les postes occupés, les tranches d'âge, l'ancienneté, le diplôme obtenu et le fait d'avoir suivi ou non des cours en méthodes quantitatives, les tests de Khi-deux ne sont pas statistiquement significatifs, il n'y a donc pas de lien entre ces variables sociodémographiques et la préférence pour l'un ou l'autre des rapports. Autrement dit, elles ne jouent aucun rôle dans le choix des formats de rapports préférés par les participants. Quant au lien entre la préférence des rapports et le domaine de formation, les tests sont significatifs pour la question 1 et 4, les participants en enseignement au primaire et en adaptation scolaire ont la tendance de choisir le format B en premier lieu, puis le format C et enfin le format A. Par contre, ceux ayant suivi une formation en psychopédagogie ou en d'autres domaines préfèrent le format A que les formats B et C. Nous reviendrons sur ce point dans la discussion. Reste à savoir comment ces variables peuvent influencer les perceptions des participants sur l'évaluation de la qualité des rapports, ce qui fait l'objet de la partie suivante.

4.4.3. Évaluation de la qualité des rapports

4.4.3.1. Études de la dimensionnalité sur l'évaluation de la qualité des rapports

La troisième section du questionnaire (annexe 6) porte sur l'évaluation de la qualité des rapports répartie en quatre sous-parties: (1) évaluation des directives pour lire les rapports; (2) évaluation de la présentation du profil de l'élève; (3) évaluation de la description du profil et des pistes d'intervention et (4) évaluation du guide d'interprétation. Cette section est constituée de 32 items de type Likert à quatre points de réponses selon le degré d'accord ou désaccord. Au total, il y a 68 participants qui ont répondu à cette section. La moyenne des réponses varie entre 2,46 et 3,35 avec un écart-type qui varie entre 0,575 et 0,880 (voir tableaux 1.1 et 1.2 de l'annexe 1), ce qui suggère que les participants ont une perception positive de la qualité des formats de rapports.

Les données recueillies ont fait l'objet d'analyses factorielles exploratoires pour l'ensemble de 32 items afin d'étudier la dimensionnalité, puis d'analyses séparées pour chaque partie afin d'étudier leur dimension respective. Bien que la revue de littérature nous suggère une taille de l'échantillon de 100 comme seuil minimal pour les analyses factorielles (Field, 2009; Gorsuch, 1978; Kline, 1979; MacCallum, Widaman, Zhang et Hong, 2001), les recherches empiriques montrent qu'il est possible de faire des analyses factorielles avec des échantillons plus petits lorsque le nombre des facteurs est limité à une ou deux dimensions et que ces facteurs sont bien définis (Loye et Barroso Da Costa, 2013). Nous avons effectué des analyses avec SPSS, avec la méthode d'extraction par factorisation en axes principaux, la rotation orthogonale de Varimax et une saturation plus élevée que 0,30.

Pour l'ensemble de 32 items, l'analyse de la fiabilité nous donne un alpha de Cronbach de 0,938, ce qui est considéré comme excellent parce qu'il dépasse le seuil minimal de 0,70 (Nunaly, 1978). Le test de sphéricité de Barlett ($\chi 2$ (496) = 1543,653, p = 0,000) ainsi que la mesure Kaiser-Meyer-Olkin (KMO) ont indiqué un très bon ajustement des items aux facteurs latents (KMO=0,759). La solution à un facteur explique 35,779% de la variance totale, en gardant tous les items. Cependant, les indices de saturation de l'item 1.4 (*J'aimerais que les objectifs du rapport diagnostique soient ajoutés dans les directives du rapport)* et l'item 3.5 *(J'aimerais bien avoir des pistes d'intervention pour le groupe)* sont plus petits que 0,30, ce qui suggère des problèmes de l'ajustement de ces deux items aux facteurs. Le tableau 13 présente l'ensemble des indices de saturation des items avec et sans les items 1.4 et 3.5. Il est à noter que ces deux questions sont de même nature, car les questions ne portent pas sur les éléments précis de la partie de l'évaluation des directives et de l'évaluation des pistes d'intervention. Ces questions ont été reformulées sous forme d'une hypothèse qui peut les rendre plus difficiles à comprendre. Ceci pourrait expliquer une saturation faible de ces deux items.

La suppression des items 1.4 et 3.5 apporte une amélioration de l'indice de saturation pour la plupart des items, soit 20 items sur 30. Aucun changement remarquable de l'indice de saturation n'a été constaté pour les 10 items qui restent. Tous les items ont donc une saturation supérieure à 0,30. Le test de sphéricité de Barlett ($\chi 2$ (435) = 1511,181, p = 0,000) ainsi que la mesure Kaiser-Meyer-Olkin (KMO) ont indiqué un très bon ajustement des items aux facteurs latents (KMO=0,760). La solution à un facteur explique 38,060 % de la variance totale en supprimant l'item 1.4 et 3.5. De plus, l'analyse de la fiabilité pour ces 30 items suggère un alpha de Cronbach de 0,943. Nous avons donc décidé de les enlever dans nos analyses et de les traiter séparément. Ainsi, ces 30 items peuvent former un facteur intitulé *Évaluation de la qualité globale des rapports*.

Tableau 40. Indice de saturation de 32 items de l'évaluation de la qualité des rapports

Items	Saturation avec items 1.4 et 3.5	Saturation sans items 1.4 et 3.5
1.1- Les informations fournies sur l'élève sont suffisantes pour personnaliser le rapport.	0,555	0,556
1.2- Les directives pour lire le rapport sont définies avec clarté	0,491	0,496
1.3- Ces directives sont utiles pour mon interprétation du rapport	0,438	0,439
1.4- J'aimerais que les objectifs du rapport diagnostique soient ajoutés dans les directives du rapport	0,111	
2.1-Les graphiques présentés sont faciles à comprendre.	0,487	0,488
2.2-Dans les graphiques, l'utilisation de différentes couleurs selon les niveaux de maitrise des habiletés est pertinente pour ma compréhension	0,538	0,538
2.3-La taille des graphiques facilite ma lecture.	0,652	0,651
2.4-Les légendes sont claires pour mon interprétation des graphiques	0,544	0,545
2.5-Les légendes sont utiles pour mon interprétation des graphiques	0,644	0,644
2.6-Les chiffres présentés dans les barres facilitent le positionnement de l'élève par rapport au groupe	0,606	0,607
2.7-Le positionnement de l'élève par rapport au groupe permet de mieux identifier les pistes d'intervention	0,673	0,675
2.8-Les niveaux de maitrises des habiletés de l'élève sont faciles à comparer avec ceux du groupe	0,374	0,378
2.9-J'ai confiance aux résultats diagnostiques présentés	0,593	0,596
3.1-La description du profil est utile pour mon interprétation du graphique.	0,546	0,544
3.2-La description du profil correspond à ma compréhension des graphiques	0,640	0,638
3.3-Les pistes d'intervention pour l'élève (individuel) sont utiles pour mon enseignement	0,543	0,539
3.4-Les pistes d'intervention pour l'élève (individuel) sont utiles pour mon évaluation	0,539	0,537
3.5-J'aimerais bien avoir des pistes d'intervention pour le groupe	0,139	
3.6-Dans l'ensemble, le vocabulaire utilisé est facile à comprendre	0,693	0,694
3.7-Les paragraphes sont bien formulés	0,681	0,681
4.1-Les objectifs du rapport diagnostique sont présentés avec clarté.	0,733	0,732
4.2-La description des habiletés facilite la compréhension du rapport	0,677	0,678
4.3-Le mode d'emploi pour lire les graphiques est utile pour ma compréhension du rapport	0,568	0,565
4.4-Les exemples utilisés dans cette partie facilitent mon interprétation des graphique	0,663	0,662
4.5-La quantité des informations fournies est suffisante pour comprendre le profil de l'élève	0,830	0,830
4.6-Dans l'ensemble, le vocabulaire utilisé est facile à comprendre	0,682	0,682
4.7-Les paragraphes sont bien formulés	0,789	0,790
4.8-Dans l'ensemble, le rapport présenté a un aspect visuel qui me plaît.	0,513	0,515
4.9-Ce rapport est plus détaillé que le rapport fait par un professionnel (par exemple, les orthopédagogues, les conseillers pédagogiques, etc.)	0,336	0,336
4.10-J'aimerais avoir accès à ce format de rapport pour mes élèves	0,740	0,740
4.11-J'ai confiance dans mon interprétation du profil de mon élève	0,801	0,799
4.12-Le profil de l'élève présenté dans ce rapport diagnostique correspond à celui de certains de mes élève	0,605	0,603

4.4.3.2. Études de la dimensionnalité sur les sous-parties de l'évaluation de la qualité des rapports

Dans un deuxième temps, nous nous intéressons à connaitre les perceptions des participants sur l'évaluation de la qualité des parties distinctes du rapport. C'est pour cette raison que nous avons décidé de faire des analyses factorielles séparées pour chaque partie, car les suggestions de regroupement des questions dans différents facteurs nous permettent de comprendre plus en profondeur les perceptions des participants sur la qualité de chaque partie du rapport. Au début, l'évaluation des directives contient quatre questions portant sur la première partie des rapports. Le test de sphéricité de Barlett ($\chi 2$ (6) =78,893, p = 0,000) ainsi que la mesure Kaiser-Meyer-Olkin (KMO) ont indiqué un bon ajustement des items aux facteurs latents (KMO=0,697). La solution à un facteur explique 45,465 % de la variance totale en conservant tous les quatre items. Cependant, la saturation de l'item 1.4 est moins de 0,30, ce qui montre un mauvais ajustement de cet item au facteur. L'analyse de fiabilité suggère un alpha de Cronbach faible qui est de 0,512.

En supprimant l'item 1.4, le test de sphéricité de Barlett ($\chi 2$ (3) =73,545, p = 0,000) ainsi que la mesure Kaiser-Meyer-Olkin (KMO) ont indiqué un bon ajustement des items aux facteurs latents (KMO=0,693). La solution à un facteur explique 58,900% de la variance totale avec les trois items, ce qui présente une augmentation de 13,4% de la variance totale. Tous ces trois items ont un indice de saturation supérieure à 0,30. Le tableau 42 présente la saturation des items avec et sans item 1.4. L'analyse de fiabilité suggère une augmentation considérable de l'alpha de Cronbach, soit de 0,805, qui est excellent. Nous avons donc décidé d'enlever cet item dans nos analyses inférentielles ultérieures et de le traiter à part.

Tableau 41. Indices de saturation pour les items de l'évaluation des directives

Items	Saturation avec item 1.4	Saturation sans item 1.4
1.1- Les informations fournies sur l'élève sont suffisantes pour personnaliser le rapport.	0,696	0,699
1.2- Les directives pour lire le rapport sont définies avec clarté	0,754	0,718
1.3- Ces directives sont utiles pour mon interprétation du rapport	0,839	0,874
1.4- J'aimerais que les objectifs du rapport diagnostique soient ajoutés dans les directives du rapport	0,249	

L'évaluation du profil de l'élève se compose de neuf questions portant sur les critères de la présentation visuelle des graphiques et du contenu tels que la clarté des graphiques, l'utilité des chiffres et des légendes, etc. Le test de sphéricité de Barlett ($\chi 2$ (36) =340,791, p = 0,000) ainsi que la mesure Kaiser-Meyer-Olkin (KMO) ont indiqué un très bon ajustement des items aux facteurs latents (KMO=0,880). L'analyse factorielle exploratoire avec la méthode d'extraction par factorisation en axes principaux avec la rotation Varimax nous suggère deux facteurs avec une explication de 58,787 % de la variance totale. Les neuf items ont un indice de saturation supérieur à 0,30. Le tableau 43 présente les indices de saturation de tous les items.

En nous basant sur les indices de saturation, le premier facteur se compose des cinq questions de 2.1 à 2.5 tandis que le deuxième facteur contient des questions de 2.6 à 2.9. Il est à noter que les items 2.3; 2.4; 2.5; 2.6; 2.7; 2.8; 2.9 saturent à la fois sur deux facteurs avec les indices de saturation supérieurs à 0,30. Nous avons donc choisi le facteur avec un indice de saturation plus grand pour décider du regroupement des items. Ainsi, le facteur 1 qui contient des questions 2.1; 2.2; 2.3; 2.4 et 2.5 portent sur l'évaluation de la présentation du graphique tandis que le facteur 2 avec les questions 2.6; 2.7; 2.8; 2.9 portent sur l'évaluation du contenu des graphiques. L'analyse de la fiabilité suggère un alpha de Cronbach de 0,886 pour le facteur 1 et de 0,797 pour le facteur 2, ce

qui est excellent. Ces deux regroupements vont être retenus pour nos analyses descriptives et inférentielles ultérieures.

Tableau 42. Indice de saturation des items de l'évaluation du profil de l'élève.

	Facteur Présentation du graphique	Contenu du graphique
2.1-Les graphiques présentés sont faciles à comprendre.	0,789	0,229
2.2- Dans les graphiques, l'utilisation de différentes couleurs selon les niveaux de maitrise des habiletés est pertinente pour ma compréhension	0,769	0,162
2.3-La taille des graphiques facilite ma lecture.	0,691	0,354
2.4-Les légendes sont claires pour mon interprétation des graphiques	0,715	0,339
2.5-Les légendes sont utiles pour mon interprétation des graphiques	0,637	0,492
2.6-Les chiffres présentés dans les barres facilitent le positionnement de l'élève par rapport au groupe	0,458	0,574
2.7-Le positionnement de l'élève par rapport au groupe permet de mieux identifier les pistes d'intervention	0,327	0,703
2.8-Les niveaux de maitrises des habiletés de l'élève sont faciles à comparer avec ceux du groupe	0,084	0,620
2.9-J'ai confiance aux résultats diagnostiques présentés	0,377	0,669

La troisième partie de cette section du questionnaire porte sur l'évaluation de la description du profil de l'élève et des pistes d'intervention. Cette partie se compose de sept items sur le contenu de la description du profil de l'élève et la proposition des pistes d'intervention (questions de 3.1 à 3.5) ainsi que la qualité de la reformulation (questions 3.6 et 3.7). Le test de sphéricité de Barlett ($\chi 2$ (21) =217,423, p = 0,000) ainsi que la mesure Kaiser-Meyer-Olkin (KMO) ont indiqué un ajustement moyen des items aux facteurs latents (KMO=0,674). La solution à un facteur explique de 45,748% de la variance totale. Les indices de saturation des items sont supérieurs à 0,30, sauf pour l'item 3.5. L'analyse de fiabilité avec tous les items suggère un alpha de Cronbach de 0,772.

Avec la suppression de l'item 3.5, l'indice de KMO=0,670 est plus ou moins semblable, cependant la solution à un facteur explique 52,700% de la variance totale, ce qui témoigne une augmentation de 12%. L'indice de saturation de tous les six items est

supérieur à 0,30. Une augmentation de l'alpha de Cronbach a été également constatée, soit de 0,819. Le tableau 43 présente l'indice de saturation de tous les items avec et sans item 3.5. Ainsi, nous avons fait le regroupement de ce facteur sur l'évaluation de la description du profil et des pistes d'intervention avec six items. L'item 3.5 a été enlevé pour le traiter séparément.

Tableau 43. Indice de saturation des items de la description du profil et des pistes d'intervention

	Saturation avec item 3.5	Saturation sans item 3.5
3.1-La description du profil est utile pour mon interprétation du graphique.	0,389	0,381
3.2-La description du profil correspond à ma compréhension des graphiques	0,335	0,331
3.3-Les pistes d'intervention pour l'élève (individuel) sont utiles pour mon enseignement	0,388	0,386
3.4-Les pistes d'intervention pour l'élève (individuel) sont utiles pour mon évaluation	0,469	0,468
3.5-J'aimerais bien avoir des pistes d'intervention pour le groupe	0,047	
3.6-Dans l'ensemble, le vocabulaire utilisé est facile à comprendre	0,849	0,849
3.7-Les paragraphes sont bien formulés	0,840	0,840

La dernière partie de cette section du questionnaire est consacrée à l'évaluation du guide d'interprétation et de l'ensemble du rapport. Cette partie se compose de douze questions de type Likert à quatre points selon le degré d'accord ou désaccord, dont sept items sur l'évaluation du guide d'interprétation (items de 4.1 à 4.7) et cinq items sur l'évaluation de l'ensemble du rapport (items de 4.8 à 4.12). Le test de sphéricité de Barlett ($\chi 2$ (66) =518,600, p = 0,000) ainsi que la mesure Kaiser-Meyer-Olkin (KMO) ont indiqué un très bon ajustement des items aux facteurs latents (KMO=0,864). L'analyse factorielle par factorisation en axes principaux avec la rotation « Varimax » nous suggère deux facteurs, ce qui explique 56,281% de la variance totale. Les indices de saturation sont présentés dans le tableau 44.

Tableau 44. Indices de saturation des items de l'évaluation du guide d'interprétation et de l'ensemble du rapport.

Items	Facteur Guide d'interprétation	Ensemble du rapport
4.1-Les objectifs du rapport diagnostique sont présentés avec clarté.	0,857	0,168
4.2-La description des habiletés facilite la compréhension du rapport	0,806	0,159
4.3-Le mode d'emploi pour lire les graphiques est utile pour ma compréhension du rapport	0,655	0,211
4.4-Les exemples utilisés dans cette partie facilitent mon interprétation des graphiques	0,735	0,246
4.5-La quantité des informations fournies est suffisante pour comprendre le profil de l'élève	0,691	0,505
4.6-Dans l'ensemble, le vocabulaire utilisé est facile à comprendre	0,637	0,396
4.7-Les paragraphes sont bien formulés	0,602	0,527
4.8-Dans l'ensemble, le rapport présenté a un aspect visuel qui me plaît	0,312	0,430
4.9-Ce rapport est plus détaillé que le rapport fait par un professionnel (par exemple, les orthopédagogues, les conseillers pédagogiques, etc.)	0,006	0,587
4.10-J'aimerais avoir accès à ce format de rapport pour mes élèves	0,498	0,629
4.11-J'ai confiance dans mon interprétation du profil de mon élève	0,655	0,444
4.12-Le profil de l'élève présenté dans ce rapport diagnostique correspond à celui de certains de mes élèves	0,412	0,491

En nous basant sur les indices de saturation, les items 4.5; 4.6; 4.7; 4.8; 4.9; 4.10; 4.11; 4.12 saturent à la fois à deux facteurs avec les indices de saturation supérieurs à 0.30. Nous avons donc choisi le facteur avec un indice de saturation plus grand pour décider le regroupement des items. Ainsi, le facteur 1 qui porte sur l'évaluation du guide

d'interprétation se compose huit items 4.1; 4.2; 4.3; 4.4; 4.5; 4.6; 4.7 et 4.11. Le facteur 2 qui sert à évaluer l'ensemble du rapport contient quatre items 4.8; 4.9; 4.10 et 4.12. Il est à noter que l'item 4.11 *J'ai confiance dans mon interprétation du profil de mon élève, qui* est supposé évaluer le facteur 1 fait désormais partie du facteur 2, ce qui nous semble également logique. En effet, cet item 4.11 représente un lien avec le guide d'interprétation qui fournit des modes d'emploi pour lire les graphiques, ce qui pourrait influencer la confiance dans l'interprétation du profil de l'élève. L'analyse de fiabilité suggère un alpha de Cronbach de 0,924 pour le facteur 1 et de 0,691 pour le facteur 2.

En résumé, la partie précédente présente les résultats des analyses factorielles exploratoires avec la méthode d'extraction par factorisation en axes principaux des questions de l'évaluation de la qualité des rapports. Nous avons effectué des analyses factorielles pour l'ensemble de 32 items, que nous avons ensuite séparé en quatre sous-parties conformes à la structure de notre questionnaire. L'idée de regrouper les questions en un seul facteur nous permet de comprendre l'évaluation de la qualité globale des rapports tandis que le regroupement des questions en quatre sous-parties nous renseigne sur l'évaluation de la qualité des rapports d'une manière plus spécifique. Les analyses factorielles pour les sous-sections nous permettent également de retrouver la structure du questionnaire. Les résultats suggèrent que les items ont un bon ajustement aux facteurs proposés pour l'ensemble des questions et pour chaque partie séparée. Ces suggestions de regroupement nous amènent à créer des variables composites qui servent faire des analyses descriptives et inférentielles dans les parties suivantes.

4.4.3.3. Analyses descriptives de l'évaluation de la qualité des rapports

Afin de faire des analyses descriptives et inférentielles sur l'évaluation de la qualité des rapports, nous avons calculé des variables composites selon les suggestions de regroupements obtenus à partir des analyses factorielles. Plus précisément, nous avons créé la variable *évaluation de la qualité globale des rapports* en regroupant les 30 items. La variable *évaluation des directives pour lire les rapports* se compose de trois items 1.1; 1.2 et 1.3. La variable *évaluation de la présentation des graphiques* regroupe les items 2.1; 2.2; 2.3; 2.4 et 2.5 tandis que la variable *évaluation du contenu des graphiques* inclut les variables 2.6; 2.7; 2.8 et 2.9. La variable sur *l'évaluation de la description du profil et des pistes d'intervention* contient des items 3.1; 3.2; 3.3; 3.4; 3.6 et 3.7. La

variable sur *l'évaluation du guide d'interprétation* se compose des items 4.1; 4.2; 4.3; 4.4; 4.5; 4.6 ; 4.7 et 4.11 tandis que la variable sur l'évaluation de l'ensemble du rapport contient des items 4.8; 4.9; 4.10 et 4.12. Les variables composites ont été calculées par les moyennes de tous les items de chaque facteur. Les items 1.4 et 3.5 ont été retirés de la liste des items et traités séparément. Le tableau 45 présente tous les regroupements.

Tableau 45. Facteurs proposés par des analyses factorielles**.**

Variables	**Items**
Évaluation de la qualité globale des rapports	30 items (en supprimant les items 1.4 et 3.5)
Évaluation des directives pour lire les rapports	1.1; 1.2 et 1.3
Évaluation de la présentation visuelle des graphiques	2.1; 2.2; 2.3; 2.4 et 2.5
Évaluation du contenu des graphiques	2.6; 2.7; 2.8 et 2.9
Évaluation de la description du profil et des pistes d'intervention	3.1; 3.2; 3.3; 3.4; 3.6 et 3.7
Évaluation du guide d'interprétation	4.1; 4.2; 4.3; 4.4; 4.5; 4.6 ; 4.7 et 4.11
Évaluation de l'ensemble du rapport	4.8; 4.9; 4.10 et 4.12

Dans l'ensemble, les participants ont des perceptions très positives de l'évaluation de la qualité globale des rapports (m=3,1098; s[6]=0,43701) (tableau 1.3 de l'annexe 1). Les moyennes des variables sont supérieures à 3,0, sauf pour la variable de l'évaluation de l'ensemble du rapport (m=2,8676; s=0,52453). La moyenne la plus grande est pour l'évaluation de la description du profil et des pistes d'intervention (m=3,2384; s=0,45864). La moyenne de l'évaluation du guide d'interprétation (m=3,2298; s=0,52795) suit de très près celles de l'évaluation du profil de l'élève et des pistes d'intervention. La moyenne de l'évaluation des directives (m=3,0395; s=0,63003), celle de l'évaluation de la présentation des graphiques (m=3,0932; s=0,67006) et celle de l'évaluation des graphiques (m=3,0000; s=0,63191) sont très semblables les unes des

[6] Écart-type

autres. L'écart-type de l'évaluation de la présentation des graphiques est le plus grand (s=0,67006), ce qui signifie la plus grande dispersion des réponses des participants autour de la moyenne. Cependant, les écarts-type de l'évaluation des directives (s=0,63003) et de l'évaluation du contenu des graphiques (s=0,63191) sont également très proches. L'écart-type le plus petit est celui de l'évaluation de la qualité des rapports (s=0,43701), ce qui montre que dans l'ensemble, les réponses des participants sont moins dispersées autour de la moyenne que celles de chaque facteur séparé. La figure 17 résume l'évaluation de la qualité de différentes parties du rapport.

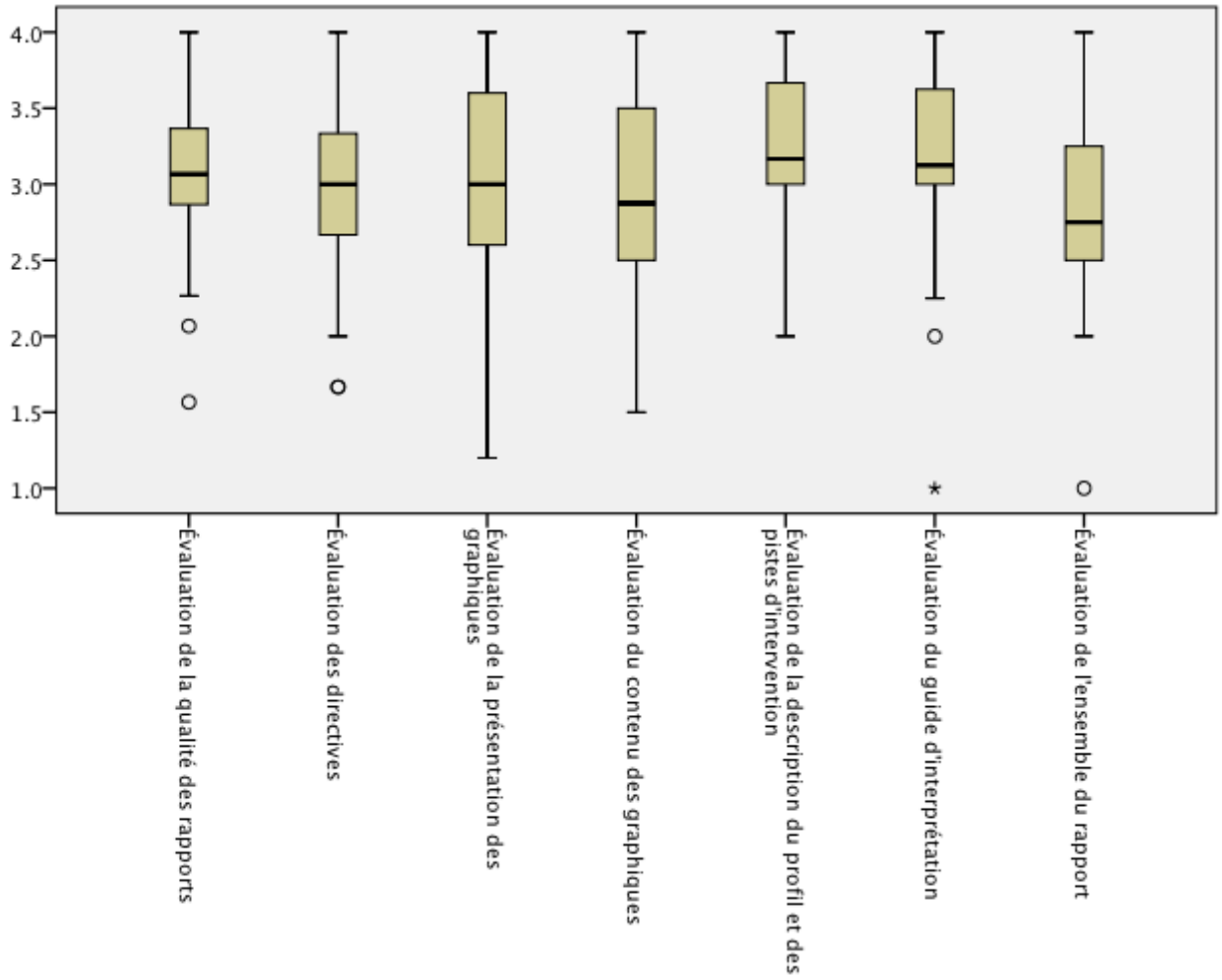


Figure 17. Évaluation de la qualité de différentes parties des rapports

4.4.3.4. Évaluation de la qualité des rapports selon le format de rapport évalué

À titre de rappel, le format évalué a été attribué aléatoirement aux répondants afin d'assurer une exposition équitable de trois formats de rapports. Selon les résultats

descriptifs, l'évaluation de la qualité selon le format de rapport évalué est très positive, car les moyennes sont supérieures à 3,0 (voir tableau 1.4 de l'annexe 1). Parmi les trois formats, l'évaluation de la qualité globale du format B est légèrement plus positive (m=3,1333; s=0.57879), vient ensuite celle du format C (m=3,1143; s=0,36089) et finalement le format A (m=3,0772; s=0,3772). Cet ordre croissant de l'évaluation de la qualité des rapports correspond bien aux tendances observées chez les participants quant aux formats de rapports que chacun a privilégiés. La figure 18 met en évidence de manière claire cet ordre.

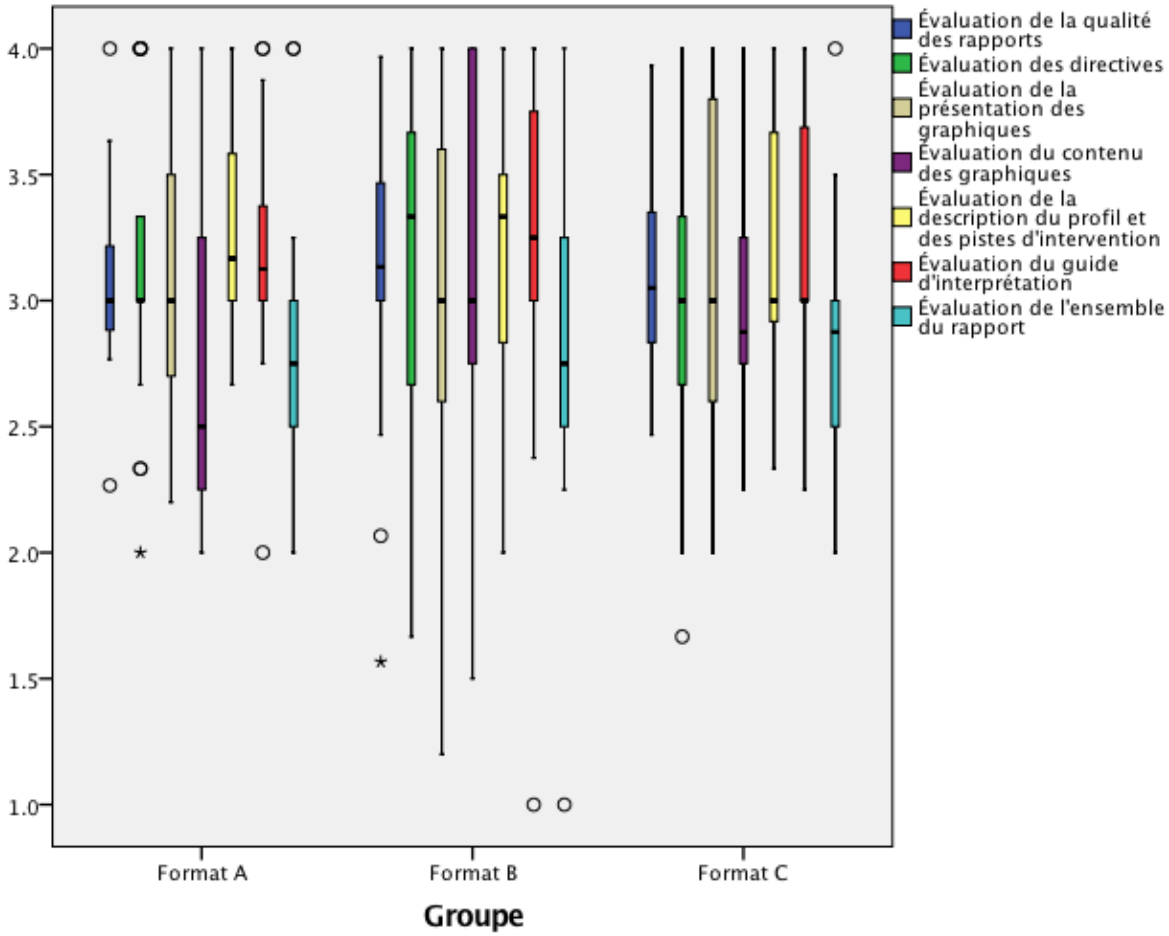


Figure 18. Évaluation de la qualité des rapports selon le format évalué

Si nous examinons l'évaluation de chaque facteur séparé, nous constatons ces mêmes tendances pour l'évaluation du contenu des graphiques et l'évaluation du guide d'interprétation. Cependant, les perceptions des participants sont un peu différentes pour

d'autres parties du rapport. Pour l'évaluation des directives, l'évaluation est la plus positive pour le format B (m=3,1733; s=0,60949); suivi de très près du format A (m=3,1111; s=0,55109) et finalement du format C (m=2,8778; s=0,68079). Quant à l'évaluation de la présentation des graphiques et celle de la description du profil et des pistes d'interprétation, l'évaluation est la plus positive pour le format A (m=3,1300; s=0,58138 et m=3,2833; s=0,37502), vient ensuite le format C (m=3,1000; s=0,78443 et m=3,2278; s=0,48440), et enfin le format B (m=3,0522; s=0,78443 et m=3,2121; s=0,50705). Quant à l'évaluation du guide d'interprétation, les perceptions des répondants sont aussi positives pour le format C (m=2,8750; s=0,45896), puis le format A (m=2,8684; s=0,51619) et enfin le format B (2,8571; s=0,63033). Par ailleurs, l'évaluation de ce facteur s'avère moins positive que celle des autres facteurs, car les moyennes des trois formats sont toutes inférieures à 3,0. Avec ces différentes tendances, nous nous intéressons à savoir si les formats de rapports évalués influencent les perceptions des participants sur l'évaluation de la qualité des rapports.

Nous avons réalisé des tests Anova à un facteur de classification pour vérifier si l'évaluation de la qualité des rapports et de chaque facteur présente une différence statistiquement significative selon le format de rapport évalué. Pour ce faire, nous avons fait des analyses préliminaires pour vérifier certains postulats comme le test de Kolmogorov-Smirnov pour un échantillon (tableau 2.1 de l'annexe 2) pour la normalité, et le test de Levene pour l'égalité des variances dont les résultats se trouvent dans l'annexe 2.

Les résultats de l'Anova à un facteur de classification (tableau 3.1 de l'annexe 3) montrent que les perceptions des répondants sur l'évaluation de la qualité des rapports ne varient pas selon le format de rapport évalué (F(2)=0,021; p=0,921). Si nous regardons les facteurs séparés, nous remarquons les mêmes résultats pour l'évaluation des directives (F(2)=1,720; p=0,186); pour l'évaluation de la présentation des graphiques (F(2)=0,073; p=0.930); pour l'évaluation du contenu des graphiques (F(2)=2,787; p=0,068) et pour l'évaluation de la description du profil et des pistes d'interprétation (F(2)=0,137; p=0,872); pour l'évaluation du guide d'interprétation (F(2)=0,125; p=0,882) ainsi que l'évaluation de l'ensemble du rapport (F(2)=0,007; p=0,993). Les participants ont donc les mêmes perceptions sur la qualité des formats de rapports évalués même s'il y a certaines tendances différentes observées au niveau des résultats descriptifs.

4.4.3.5. Évaluation de la qualité des rapports selon les postes

Dans l'ensemble, les participants ont un regard positif sur l'évaluation de la qualité globale des rapports, peu importe les postes occupés (tableau 1.5 de l'annexe 1). Les conseillers pédagogiques sont ceux qui sont les plus positifs (m=3,2472; s=0,43099), viennent ensuite les orthopédagogues (m=3,0867; s=0,29740) et finalement les enseignants au primaire (m=3,0790; s=0,46382). Nous observons les mêmes tendances pour l'évaluation de la présentation des graphiques; l'évaluation du contenu des graphiques; l'évaluation de la description du profil et des pistes d'intervention et l'évaluation du guide d'interprétation. Quant à l'évaluation des directives et l'évaluation de l'ensemble du rapport, ce sont toujours les conseillers pédagogiques qui sont les plus positifs (m=3,1667; s=0,58104 et m=3,1042; s=0,39107); les enseignants au primaire arrivent au deuxième rang (m=3,0553; s=0,63303 et m=2,8370; s=0,56573) et enfin les orthopédagogues (m=2,8333; s=0,67420 et m=2,7250; s=0,39878). Les participants ont le regard le moins positif pour la partie sur l'évaluation de l'ensemble du rapport.

Les résultats des tests de Levene montrent qu'il y a de l'égalité des variances pour toutes les variables (tableau 3.2 de l'annexe 3). Les résultats de l'Anova à un facteur suggèrent qu'il n'y a pas de différence de perceptions des participants sur les rapports évalués que ce soit sur l'évaluation de la qualité (F(2)=0,716; p=0,493); l'évaluation des directives (F(2)=0,938; p=0,396; l'évaluation de la présentation des graphiques (F(2)=0,989; p=0,377); l'évaluation du contenu des graphiques (F(2)=1,248; p=0,293); l'évaluation de la description et des pistes d'intervention (F(2)=0,658; p=0,521) et sur l'évaluation du guide d'interprétation (F(2)=0,183; p=0,883) ainsi que l'évaluation de l'ensemble du rapport (F(2)=1,704; p=0,190). Par conséquent, les postes occupés n'influencent pas l'évaluation des participants sur la qualité des rapports.

4.4.3.6. Évaluation de la qualité des rapports selon les régions de provenance

Malgré un décalage dans le nombre des participants en provenance du Québec et ceux hors-Québec, nous constatons des tendances différentes dans les perceptions des participants sur les formats de rapports évalués (tableau 1.6 de l'annexe 1). Sur l'évaluation de la qualité globale, les participants non-québécois semblent avoir un regard plus positif (m=3,1278; s=0,52974) que les participants provenant du Québec (m= 3,1081; s=0,43212). Ces tendances sont les mêmes pour l'évaluation de la description du profil et des pistes d'intervention; pour l'évaluation du guide d'interprétation et pour l'évaluation de l'ensemble du rapport.

Quant à l'évaluation des directives (m=3,0476; s=0,62499 et m=2,9444; s=0.74287), l'évaluation de la présentation des graphiques (m=3,1104; s=0,67356 et m= 2,9000; s=0,65422) et l'évaluation du contenu des graphiques (m=3,0075; s=0,63659 et m=2,9167; s=0,62583), les participants québécois sont plus positifs que leurs collègues non québécois. Les participants québécois semblent les moins positifs dans l'évaluation de l'ensemble du rapport (m=2,8427; s=0,51461), tandis que leurs collègues non québécois sont les moins positifs quant à l'évaluation de la présentation des graphiques (m=2,9444; s=0,74287). Les deux groupes sont les plus positifs pour l'évaluation de la description du profil de l'élève et des pistes d'intervention (m=3,3333; s=0,58689 et m=3,2298; s=0,44985). Avec ces différentes tendances, nous avons fait des tests de Student (test T) pour voir si les différences entre les deux sous-groupes sont statistiquement significatives.

Les résultats du test T (t(66)=0,105; p=0,917) pour deux échantillons indépendants montrent que l'évaluation de la qualité globale des rapports n'est pas différente selon les régions de provenance (tableau 4.1 de l'annexe 4). Ces mêmes résultats ont été également remarqués pour l'évaluation des facteurs séparés telle que l'évaluation des directives (t(74)= -0,383 p=0.703); l'évaluation de la présentation des graphiques (t(71)= -0,735; p=0,465); l'évaluation du contenu des graphiques (t(71)=-0,335; p=0,739); l'évaluation de la description du profil et des pistes d'intervention (t(70)= 0,527; p=0,600); l'évaluation du guide d'interprétation (t(66)= 0,299; p=0,766) et l'évaluation de l'ensemble du rapport (t(66)= 1,264; p=0,211). Autrement dit, les régions de provenance

n'influencent pas les perceptions des participants sur l'évaluation de la qualité des rapports.

4.4.3.7. Évaluation de la qualité des rapports selon l'âge

Dans l'évaluation de la qualité des rapports, la répartition des participants selon les deux tranches d'âges 20-39 ans (N=30) et le groupe de 40 ans et plus (N=38) est assez équitable. Les participants du groupe de 20 à 39 ans semblent plus positifs dans l'évaluation de la qualité globale des rapports (m=3,1422; s=0,34662) que leurs collègues plus âgés (m=3,0842; s=0,5013). Les tendances sont les mêmes pour l'évaluation de la présentation des graphiques (m=3,1212 et s=0,63234 et m=3,0700 et s=,70682); pour l'évaluation de la description du profil et des pistes d'intervention (m=3,2879; s=0,45122 et m=3,0875; s=0,67594); pour l'évaluation du guide d'interprétation (m=3,2792; s=0,44125 et m=3,1908; s=0,59042) et pour l'évaluation de l'ensemble du rapport (m=2,8750 et s=0,42926 et m=2,8618 et s=0,59473).

Quant à l'évaluation des directives, les participants de 40 ans et plus (m=3,0488; s=0’62165 sont plus positifs que leurs collègues moins âgés (m=3,0286; s=0,64864). Nous avons constaté ces mêmes tendances avec l'évaluation du contenu des graphiques (m=3,0875; s=,67594 et m=2,8939; s=0,56607). Comme dans les cas précédents, les participants de deux groupes sont les moins positifs pour l'évaluation de l'ensemble du rapport (m=2,8750 et s=0,42926 et m=2,8618 et s=0,59473). Les participants de deux tranches d'âge sont les plus positifs pour l'évaluation de la description du profil et des pistes d'intervention (m=3,2879; s=0,45122 et m=3,0875; s=0,67594). Le tableau 1.7 de l'annexe 1 présente les résultats descriptifs de l'évaluation des rapports selon les tranches d'âges.

Malgré les différentes tendances observées dans les résultats descriptifs, les résultats de test T nous indiquent qu'il n'y a pas de différence statistiquement significative quant à l'évaluation de la qualité globale des rapports (t(66)=0,541; p=0,591); l'évaluation des directives rapports (t(74)=-0,138; p=0,890); l'évaluation de la présentation des graphiques (t(71)=0,323; p=0,748); l'évaluation du contenu des graphiques (t(71)= -1,309; p=0,195); l'évaluation de la description du profil et des pistes d'intervention (t(70)=0,840; p=0,404); l'évaluation du guide d'interprétation (t(66)=0,683; p=0,497) et l'évaluation de l'ensemble du rapport (t(66)=0,102; p=0,919)

selon les tranches d'âge. Ainsi, les tranches d'âges n'influencent pas les perceptions des participants sur l'évaluation de la qualité des rapports (tableau 4.2 de l'annexe 4).

4.4.3.8. Évaluation de la qualité des rapports selon l'ancienneté

Dans l'ensemble, les participants les moins expérimentés avec moins de 10 ans sont les plus positifs dans l'évaluation de la qualité globale des rapports (m=3,1500; s=0,35467) en comparaison avec ceux qui sont les plus expérimentés (m=3,1154 et s=0,48926) et les participants qui ont de 11 à 20 ans d'expérience (m=3,0848; s=0,46900) (tableau 1.8 de l'annexe 1). Si nous regardons l'évaluation de chaque facteur séparé, nous ne constatons pas les mêmes tendances. Par exemple, pour l'évaluation des directives et l'évaluation de la présentation des graphiques, plus les participants sont expérimentés, plus ils sont positifs face à l'évaluation des rapports. Quant à l'évaluation du contenu des graphiques, les participants ayant 20 ans d'expérience et plus sont les plus positifs face à des rapports évalués (m=3,2667; s=0,69093), viennent ensuite les participants avec moins de 10 ans d'expérience (m=2,9565; s=0,60138) et finalement ceux qui ont de 11 à 20 ans d'expérience (m=2,9143; s=0,61220).

En ce qui concerne l'évaluation de la description du profil et des pistes d'intervention et l'évaluation du guide d'interprétation, les participants les moins expérimentés sont ceux qui sont les plus positifs (m=3,2899; s=0,43886 et m=3,3375; s=0,43886). Les participants de 20 et plus d'expérience et plus arrivent au deuxième rang dans le cas de l'évaluation de la description du profil et des pistes d'intervention (m=3,2262; s=0,43871) tandis que ce sont les participants ayant de 11 à 20 ans d'expérience qui arrivent au deuxième rang dans l'évaluation du guide d'interprétation (m=3,2071; s=0,57598). Les participants de trois groupes sont les moins positifs dans l'évaluation de l'ensemble du rapport. Avec ces différentes tendances des résultats descriptifs, nous nous intéressons à savoir s'il y a des différences dans les perceptions des participants sur l'évaluation des rapports selon leur ancienneté.

Les résultats de l'Anova à un facteur (tableau 3.2 de l'annexe 3) montrent qu'il n'y a pas de différence statistiquement significative dans les perceptions des participants quant à l'évaluation de la qualité globale des rapports ((F(2)=0,130; p=0,870) et l'évaluation des directives (F(2)=0,386; p=0,681); l'évaluation de la présentation des graphiques (F(2)=0,761; p=0,471); l'évaluation du contenu des graphiques (F(2)= 1,748; p=0,182); l'évaluation de la description du profil et des pistes d'intervention (F(2)= 0,214; p=0,808) l'évaluation du guide d'interprétation (F(2)= 0,698; p=0,501) ainsi que l'évaluation de l'ensemble du rapport (F(2)= 0,186; p=0,831) selon l'ancienneté. Nous concluons donc que l'ancienneté des participants n'influence pas leurs perceptions sur l'évaluation des rapports.

4.4.3.9. Évaluation de la qualité des rapports selon le diplôme obtenu

Pour l'évaluation de la qualité globale des rapports, les participants ayant une maitrise ou un doctorat ont les perceptions les plus positives face à des rapports évalués (m=3,1318; s=0,49564), viennent ensuite les participants avec un baccalauréat (m=3,1144; s=0,44787) et enfin ceux qui ont suivi un microprogramme ou un DESS (m=3,0708; s=0,34488) (voir le tableau 1.9 de l'annexe 1). Ces mêmes tendances ont été constatées pour l'évaluation de la description du profil et des pistes d'intervention. Toutefois, dans ce cas, les participants avec un baccalauréat (m=3,1970 et s=0,45349) et ceux qui ont suivi un microprogramme ou un DESS (m=3,1979; s=0,38112) ont presque les mêmes perceptions. En ce qui concerne l'évaluation du contenu des graphiques et l'évaluation du guide d'interprétation, ce sont les participants avec un baccalauréat qui sont les plus positifs (m=3,0379 et s=0,60635; m=3,2583; s=0,58991). Les participants ayant suivi un microprogramme ou un DESS sont les plus positifs en ce qui concerne l'évaluation de la présentation des graphiques (m=3,1059; s=0,71105), viennent ensuite le groupe avec un baccalauréat (m=3,0970; s=0,64638) et enfin ceux qui détiennent une maitrise ou un doctorat (m=3,0783; s=0,70256)

Quant à l'évaluation du contenu des graphiques et de l'évaluation du guide d'interprétation, nous constations que plus le niveau de scolarité est élevé, l'évaluation semble moins positive. Autrement dit, les participants avec un baccalauréat sont ceux qui sont les plus positifs, viennent ensuite les participants qui ont suivi un microprogramme ou qui ont un DESS et enfin ceux qui ont une maitrise ou un doctorat. Ce qui est le

contraire dans le cas de l'évaluation de l'ensemble du rapport : plus le niveau de scolarité des participants est élevé, plus leurs perceptions sont positives. Avec ces tendances observées, nous avons fait des tests Anova à un facteur afin de vérifier ces différences

Les résultats de l'Anova à un facteur (tableau 3.3 de l'annexe 3) suggèrent qu'il n'y a pas de différence des perceptions des participants sur l'évaluation de la qualité ($F(2)=0,091$; $p=0,913$); sur l'évaluation des directives ($F(2)=0,399$; $p=0,672$; sur l'évaluation de la présentation des graphiques ($F(2)=0,009$; $p=0,991$); sur l'évaluation du contenu des graphiques ($F(2)=0,112$; $p=0,895$); sur l'évaluation de la description et des pistes d'intervention ($F(2)=0,611$; $p=0,546$) et sur l'évaluation du guide d'interprétation ($F(2)=0,137$; $p=0,872$) ainsi que l'évaluation de l'ensemble du rapport ($F(2)=0,260$; $p=0,772$). Ainsi, peu importe le diplôme obtenu, les participants ont les mêmes perceptions sur l'évaluation des rapports.

4.4.3.10. Évaluation de la qualité des rapports selon le domaine de formation

Les participants ayant une formation en psychopédagogie et dans d'autres domaines semblent ceux qui sont les plus positifs quant à l'évaluation de la qualité globale des rapports ($m=3,2357$; $s=0,45469$); viennent ensuite les participants ayant suivi une formation en enseignement au primaire ($m=3,1186$ et $s=0,49937$) et enfin sont ceux avec une formation en adaptation scolaire ($m=3,0067$; $s=0,27543$). Nous constatons les mêmes tendances pour l'évaluation des directives, l'évaluation du contenu des graphiques et l'évaluation de l'ensemble du rapport (voir tableau 1.10 de l'annexe 1).

Quant à l'évaluation de la présentation des graphiques, ce sont les répondants du groupe avec une formation en enseignement au primaire qui sont les plus positifs ($m=3,1676$; $s=0,72650$), les participants avec une formation en psychopédagogie et dans d'autres domaines arrivent au deuxième rang ($m=3,1429$; $s=0,65365$) et enfin ce sont ceux qui ont une formation en adaptation scolaire ($m=2,9364$; $s=0,57698$) qui arrivent en dernier.

En ce qui concerne l'évaluation de la description du profil et des pistes d'intervention et l'évaluation du guide d'interprétation, les participants avec une formation en psychopédagogie et dans d'autres domaines sont toujours les plus positifs ($m=3,3333$ et $s=0,47592$ et $m=0,3393$ et $s=0,35651$), ceux avec une formation en adaptation arrivent au deuxième rang ($m=3,2197$ et $s=0,43152$ et $m=3,2313$; $s=0,35651$)

et enfin, au dernier rang, ceux qui ont une formation en enseignement au primaire (m=3,2130; s=0,47578 et m=3,1838; s=0,64052),

L'analyse d'Anova à un facteur de classification (tableau 3.4 de l'annexe 3) nous indique qu'il n'y a pas de différence statistiquement significative selon les domaines de formation quant à l'évaluation de la qualité globale des rapports (F(2)=1,150; p=0,323) et l'évaluation des directives (F(2)=2,761; p=0,07); l'évaluation de la présentation des graphiques (F(2)=0,866; p=0,425); l'évaluation du contenu des graphiques (F(2)=1,035; p=0,361); l'évaluation de la description du profil et des pistes d'intervention (F(2)=0,367; p=0,694); l'évaluation du guide d'intervention (F(2)=0,423; p=0,637) ainsi que l'évaluation de l'ensemble des rapports (F(2)=2,364; p=0,102). Nous pouvons conclure que peu importe le domaine de formation des participants, leurs perceptions sur l'évaluation des rapports ne changent pas.

4.4.3.11. Évaluation de la qualité des rapports selon le suivi des cours en méthodes quantitatives

Malgré le nombre limité des participants ayant suivi le cours en méthodes quantitatives pendant les cinq dernière années, ils sont les plus positifs en ce qui concerne l'évaluation de la qualité globale des rapports (m=3,3154; s=0,38094) et l'évaluation de différents facteurs en comparaison avec leurs collègues qui ne l'ont pas fait (m=3,0612; s=0,43883). Parmi les facteurs étudiés, les membres du groupe ayant répondu « oui » sont les plus positifs en ce qui concerne l'évaluation de la description du profil et des pistes d'intervention, tandis que leurs collègues qui ont répondu « non » sont les plus positifs pour l'évaluation du guide d'interprétation (m=3,1841 et s=0,5343). Les deux groupes sont les moins positifs pour l'évaluation de l'ensemble des rapports (m=2,9808; s=0,51500 et m=2,8409; s=0,52784). Le tableau 1.11 de l'annexe 1 présente ces tendances

Les résultats des tests T (tableau 4.3 de l'annexe 4) suggèrent qu'il y a une différence statistiquement significative dans l'évaluation de la description du profil et des pistes d'intervention (t(70)=2,700; p=0,009). Les personnes ayant suivi le cours sont plus positives que celles qui ne l'ont pas suivi.

Quant à l'évaluation de la qualité globale des rapports (t(66)=1,924; p=0,059); l'évaluation des directives (t(74)=1,059; p=0,293); l'évaluation de la présentation des

graphiques (t(71)=0,952; p=0,345); l'évaluation du contenu des graphiques (t(71)=1,865; p=0,066); l'évaluation du guide d'interprétation (t(66)=1,481; p=0,143) et l'évaluation de l'ensemble des rapports (t(66)=0,863; p=0,391), les résultats montrent que les perceptions des participants ne sont pas différentes selon qu'ils ont suivi ou pas des cours en méthodes quantitatives.

4.4.3.12. Lien entre l'évaluation des rapports et la préférence

Dans cette partie, nous nous intéressons à savoir s'il y a un lien entre l'évaluation des rapports des participants et leur préférence. L'idée est de savoir si l'évaluation de la qualité globale des rapports et des facteurs séparés dépend du fait que les participants montrent une préférence pour les rapports qu'ils ont évalués. Pour ce faire, pour chaque question de préférence, nous avons réparti les répondants en deux groupes. Le groupe 1 comprend les participants qui préfèrent le format de rapport qui leur a été attribué aux fins d'évaluation. Le groupe 2 comprend ceux dont le rapport à évaluer n'est pas celui qu'ils préfèrent. Étant donné que le choix de format de préférence est différent dans les questions, les pourcentages de deux groupes varient donc pour les quatre questions.

En nous basant sur le tableau 46, nous constatons que les participants ont plus la tendance à préférer le format qu'ils évaluent et ce, pour les quatre questions, avec les taux respectifs sont de 58,2% (Q1); 58,2% (Q2); 60,2% (Q3) et 57,1% (Q4). La plus grande dispersion entre les deux groupes se trouve à la question 3 (60,2% contre 39,8%).

Tableau 46. Répartition des participants sur les questions de la préférence

Questions	Q1	Q2	Q3	Q4
Groupe 1	58,2%	58,2%	60,2%	57,1%
Groupe 2	41,8%	41,8%	39,8%	42,9%

Les résultats descriptifs (tableau 1.12 de l'annexe 1) suggèrent que les participants du groupe 1 ont une évaluation plus positive que ceux du groupe 2. Ces tendances ont été constatées non seulement avec l'évaluation de la qualité globale des rapports, mais aussi dans tous les facteurs séparés et dans les quatre questions de préférence. Pour l'évaluation de la qualité globale, les participants ont les perceptions les plus positives

avec la question 3 (m=3,3108; s=0,38426 et m=2,9414; s=0,41047), vient ensuite la question 4 (m=3,2758; s=0,38408 et m=2,9533; s=0,43078). L'évaluation de la qualité de la question 1 arrive au troisième rang (m=3,2688; s=0,38808 et m=2,9685; s=0,43424). Les participants du groupe 1 ont toujours une évaluation plus positive de la qualité globale des rapports que ceux du groupe 2. Nous observons ces mêmes tendances avec l'évaluation de la description du profil et des pistes d'intervention; l'évaluation du guide d'interprétation et l'évaluation de l'ensemble du rapport.

En ce qui concerne les directives, l'évaluation est la plus positive avec la question 4 (m=3,2072; s=0,56847 et m=2,8803; s=0,65108), la question 3 suit de très près (m=3,2059; s=0,56862 et m=2,9048; s=0,65139), viennent ensuite les questions 2 (m=3,2000; s=0,60607 et m=2,9024; s=0,62470) et 1 (m=3,1944; s=0,57113 et m=2,9000; s=0,65459). Quant à l'évaluation de la présentation des graphiques, l'évaluation est toujours la positive avec la question 3 (m=3,5188; s=0,49477 et m=2,7610; s=0,60037), vient ensuite l'évaluation de la question 1 (m=3,5000; s=0,53541 et m=2,7385; s=0,56968), l'évaluation arrive au troisième rang dans la question 4 (m=3,4743; s=0,54898 et m=2,7421; s=0,57687) et enfin est l'évaluation dans la question 2 (m=3,4364; s=0,57327 et m=2,8100; s=0,61427). Ces mêmes tendances ont été constatées avec l'évaluation du contenu des graphiques. Dans tous les cas, l'évaluation est plus positive avec le groupe 1 que le groupe 2. Nous avons donc fait des tests T de Student pour vérifier si ces différences entre les deux groupes sont statistiquement significatives

Les résultats des tests T de Student (tableau 4.4 de l'annexe 4) avec l'évaluation de la qualité globale des rapports sont significatifs pour la question 1 (t(66)=2,991; p=0,004); la question 4 (t(66)=2,296; p=0,025); la question 3 (t(66)=3,804; p=0,000) et la question 4 (t(66)=3,250; p=0,002). Il y a donc une différence dans l'évaluation de la qualité des rapports entre les deux sous-groupes selon leur préférence des formats de rapports.

Si nous regardons l'évaluation des facteurs séparés, les résultats des tests T de Student sont significatifs avec la question 1 (t(74)=2,079; p=0,041); la question 2 (t(74)=2,098; p=0,039); la question 3 (t(74)=2,119; p=0,037) et la question 4 (t(74)=2,326; p=0,023). Nous concluons donc que l'évaluation des directives est statistiquement différente entre les deux groupes.

En ce qui concerne l'évaluation de la présentation des graphiques, les résultats des tests T de Student sont significatifs avec la question 1 (t(71)=5,828; p=0,000); la question 2 (t(71)=4,468; p=0,000); la question 3 (t(71)=5,770; p=0,000) et la question 4 (t(71)=5,544; p=0,000). Il y a donc de la différence dans l'évaluation de la présentation des graphiques entre les deux sous-groupes selon leur préférence des formats de rapports.

Avec l'évaluation du contenu des graphiques, les résultats des tests T de Student sont significatifs avec la question 1 (t(71)=2,711; p=0,008); la question 2 (t(71)=2,196; p=0,031); la question 3 (t(71)=3,282; p=0,002) et la question 4 (t(71)=2,601; p=0,011). Nous concluons donc que l'évaluation des directives est statistiquement différente entre les deux groupes selon le fait qu'ils évaluent le format qu'ils préfèrent ou pas.

Quant à l'évaluation de la description du profil et des pistes d'intervention, les résultats du test T sont significatifs seulement avec la question 3 (t(70)=2,094; p=0,040). Les résultats ne sont pas significatifs avec la question 1 (t(70)=1,366; p=0,176) ni la question 2 (t(70)=1,054; p=0,295) ou la question 4 (t(70)=1,773; p=0,081).

Nous concluons donc qu'il n'y a pas de différence dans l'évaluation de la description du profil et des pistes d'intervention entre les deux groupes pour ces trois questions. Quant à l'évaluation du guide d'interprétation, les tests ne sont significatifs dans aucune des quatre questions. Les résultats nous semblent cohérents, car la partie de l'évaluation de la description du profil et des pistes d'intervention est identique dans les trois formats de rapports, ce qui fait que l'évaluation de cet aspect ne varie pas selon leur préférence des formats de rapports.

Finalement, pour l'évaluation de l'ensemble du rapport, les résultats du test T nous suggèrent qu'il y a une différence statistiquement significative pour la question 3 (t(66)=2,456; p=0,017) et la question 4 (t(66)=2,196; p=0,032), ce qui signifie que l'évaluation de l'ensemble du rapport varie donc selon la préférence des formats de rapports dans ces deux questions. Cependant, les tests ne sont pas significatifs pour les questions 1 (t(66)=1,881; p=0,064) et 2 (t(66)=1,452; p=0,151). Nous reviendrons sur l'interprétation de ces résultats dans notre discussion.

4.4.3.12. Synthèse de la section 4.4.3

En résumé, la section 4.4.3 sert à analyser l'évaluation des rapports des participants en tenant compte des variables sociodémographiques telles que les régions de provenance, les postes occupés, les tranches d'âge, l'ancienneté, le diplôme obtenu, les domaines de formation et le fait d'avoir suivi des cours en méthodes quantitatives. Les résultats descriptifs suggèrent que les participants ont des perceptions très positives des rapports qu'ils ont évalués. Les participants semblent avoir une évaluation plus positive de la qualité pour le format B, puis le format C et enfin le format A, même si ceci peut varier selon les facteurs évalués. En ce qui concerne les postes occupés, les conseillers pédagogiques sont ceux qui sont les plus positifs, viennent ensuite les orthopédagogues, et finalement les enseignants au primaire. Cependant, les résultats des Anova à un facteur confirment que les participants ont globalement les mêmes perceptions sur la qualité des formats peu importe le format évalué et les postes occupés.

Quant aux régions de provenance, malgré le décalage dans le nombre des participants entre les deux sous-groupes, les participants non québécois ont un regard plus positif que les participants provenant du Québec. En ce qui concerne les tranches d'âge, les participants du groupe de 20 à 39 ans sont plus positifs dans l'évaluation de la qualité des rapports que leurs collègues plus âgés. Quant à l'ancienneté, les participants les moins expérimentés (avec moins de 10 ans) sont les plus positifs quant à l'évaluation de la qualité des rapports en comparaison avec ceux qui sont les plus expérimentés. Toutefois, les résultats du test T de Student nous suggèrent que les régions de provenance, les tranches d'âge ainsi que l'ancienneté n'influencent pas les perceptions des participants sur l'évaluation de la qualité des rapports.

Les participants avec une maitrise ou un doctorat ont des perceptions plus positives des rapports évalués, viennent ensuite les participants avec un baccalauréat et enfin ceux qui ont suivi un microprogramme ou un DESS. Si nous considérons les domaines de formation, nous constatons que les participants ayant une formation en psychopédagogie et dans d'autres domaines semblent plus positifs dans l'évaluation de la qualité des rapports que leurs collègues ayant suivi une formation en enseignement au primaire et en adaptation scolaire. Malgré les différences observées dans les analyses descriptives, rien de significatif n'est mis en évidence. Les perceptions des participants quant à l'évaluation

de la qualité des rapports en général et sur les facteurs séparés en particulier ne varient donc pas selon ces variables sociographiques.

Les participants ayant suivi des cours en méthodes quantitatives ont les perceptions plus positives quant à l'évaluation de la qualité des rapports que leurs collègues qui ne les ont pas suivis. Toutefois, le fait d'avoir suivi ou non des cours en méthodes quantitatives n'influence pas les perceptions des participants sur l'évaluation de la qualité des rapports.

En ce qui concerne le lien entre l'évaluation de la qualité des rapports et la préférence des participants, ceux qui évaluent le format de rapport qu'ils préfèrent ont des perceptions plus positives sur l'évaluation des rapports. Les résultats du test T montrent qu'il y a des différences statistiquement significatives entre la préférence des participants et l'évaluation de la qualité des rapports. Ainsi, le fait que les participants évaluent le même format qu'ils préfèrent influence positivement leurs perceptions sur la qualité des rapports.

En conclusion, malgré les différentes tendances observées dans notre échantillon, les perceptions des participants ne changent pas au niveau de la population. Nous concluons donc que les perceptions des participants sur l'évaluation des rapports sont les mêmes peu importe le format de rapport évalué, les régions de provenance, les postes occupés, les tranches d'âge, l'ancienneté, le diplôme obtenu, les domaines de formation et le fait d'avoir suivi ou non des cours en méthodes quantitatives. Par contre, il y a un lien entre l'évaluation de la qualité des rapports et la préférence. Les participants qui évaluent le format de rapport qu'ils préfèrent ont des perceptions plus positives de l'évaluation des rapports Nous donnerons plus d'explications sur ces résultats dans notre chapitre de discussion.

4.4.4. Évaluation de la compréhension des rapports

La section de l'évaluation de la compréhension des rapports comprend neuf questions à choix multiples à deux, trois ou quatre choix de réponse. Ces questions portent sur la capacité d'interpréter des éléments spécifiques du profil de l'élève. Parmi les choix de réponses, un seul choix correspond à une réponse correcte. Nous avons donc codé dichotomiquement ces choix de réponses sous forme de 1 pour des réponses correctes et de 0 pour des réponses incorrectes. Nous avons ensuite créé une variable qui correspond au score total de ces neuf réponses et qui varie donc entre 0 et 9.

Dans cette partie, nous avons fait des analyses descriptives de l'évaluation de la compréhension des rapports avec le score total dans un premier temps, et ensuite fait des analyses inférentielles entre chacune des questions séparées et l'évaluation de la qualité des rapports. Le but est de trouver le lien s'il existe entre la compréhension des rapports des participants et leurs perceptions sur la qualité des rapports.

4.4.4.1. Analyses descriptives de l'évaluation de la compréhension des rapports

En général, les participants ont une très bonne compréhension des rapports évalués, car sur une échelle qui varie entre 0 et 9, la moyenne des participants est supérieure à 7 pour chacun des formats. Malgré la préférence des participants pour le format B, les participants montrent une compréhension légèrement meilleure pour le format A (m=8,4706; s=0,7998) que pour les formats B (m=8,100; s=1,0208) et C (m=7,9603; s=1,5059). Nous pouvons constater que les participants comprennent moins bien le format C qui présente la plus grande dispersion des données. Le score total le plus faible (3) se trouve également dans ce groupe.

En ce qui concerne les postes occupés, les conseillers pédagogiques semblent ceux qui comprennent le mieux les rapports et ce de la manière plus homogène (m=8,4167; s=0,7929), viennent ensuite les orthopédagogues (m=8,3000; s=0,8232) et enfin les enseignants au primaire (m=8,0238; s=1,3702) dont les scores sont les plus dispersés autour de la moyenne. Comme les questions sur l'évaluation de la qualité sont formulées positivement, nous pouvons dire que plus les participants sont d’accord avec ces énoncés, plus ils sont positifs sur l'évaluation de la qualité des rapports. Il nous

semble donc y avoir un lien entre l'évaluation de la qualité et la compréhension des rapports selon les postes occupés. En effet, les conseillers pédagogiques sont les plus positifs sur l'évaluation de la qualité des rapports tandis que les enseignants au primaire sont les moins positifs. Si nous considérons les tranches d'âge, les participants de 40 ans et plus ont une meilleure compréhension (m= 8,2059; m=1,1222) que leurs collègues plus jeunes (m=8,0667; s=1,3113). En ce qui concerne l'ancienneté, plus les participants sont expérimentés, meilleure est leur compréhension des rapports. Il y a également une plus grande dispersion des réponses autour de la moyenne pour le groupe le moins expérimenté (m=7,8500; s=1,5312).

Les répondants en provenance du Québec manifestent une meilleure compréhension des formats de rapports (m=8,2373; s=1,1498) que ceux qui ne viennent pas du Québec (m=7,000; s=1,4142). La différence des moyennes de la compréhension de deux groupes est assez grande (1,2373). Si nous tenons compte du diplôme obtenu, ce sont les participants avec un microprogramme ou un DESS qui ont une meilleure compréhension des rapports avec la plus petite dispersion des réponses autour de la moyenne (m=8,4667; s=0,6399), viennent ensuite ceux qui détiennent une maitrise ou un doctorat (m=8,1570; s=1,0678) et enfin sont ceux avec un baccalauréat (m=7,9667; s=1,4735). Une plus grande dispersion des données autour de la moyenne a été également constatée pour ce groupe.

À la différence des tendances observées dans l'évaluation des rapports, ce sont les participants en enseignement au primaire qui ont la meilleure compréhension des rapports (m=8,3548; s=0,9503); les collègues en adaptation scolaire arrivent au deuxième rang (m=7,9500; s=1,5381) suivis par ceux avec une formation en psychopédagogie et dans d'autres domaines (m=7,9231; s=1,1875). Finalement, les participants qui n'ont pas suivi les cours en méthodes quantitatives au cours de cinq dernières années comprennent mieux les rapports (m=8,1923; s=1,2213) que ceux qui les ont suivis (m=7,9167; s=1,1645). Le tableau 47 résume les résultats descriptifs de la compréhension des participants selon les variables sociodémographiques.

Nous avons fait des analyses factorielles exploratoires pour l'ensemble de neuf questions de cette partie dans l'intention de trouver une suggestion de regroupement des items en un seul facteur qui représente la compréhension des rapports. Cependant, les résultats obtenus des analyses factorielles ne nous permettent pas de faire le

regroupement approprié des items. Nous nous arrêtons donc à l'analyse des résultats descriptifs sur la compréhension des rapports.

Tableau 47. Résultats descriptifs de l'évaluation de la compréhension des rapports

Informations sociodémographiques		N	Moyenne	Écart-type	Minimum	Maximum
Groupe	Format A	17	8,4706	0,7998	7,00	9,00
	Format B	20	8,1000	1,0208	5,00	9,00
	Format C	27	7,9603	1,5059	3,00	9,00
Postes occupées	Enseignant au primaire	42	8,0238	1,3702	3,00	9,00
	Orthopédagogue	10	8,3000	0,8232	7,00	9,00
	Conseiller pédagogique	12	8,4167	0,7929	7,00	9,00
Genre	Féminin	53	8,1698	1,2047	3,00	9,00
	Masculin	10	8,3000	0,8232	7,00	9,00
Tranche d'âge	20-39 ans	30	8,0667	1,3113	3,00	9,00
	40 et plus	34	8,2059	1,1222	5,00	9,00
Années d'expérience	Moins de 10 ans	20	7,8500	1,5312	3,00	9,00
	11-20 ans	33	8,2121	0,9639	5,00	9,00
	20 ans et plus	11	8,4545	1,2135	5,00	9,00
Régions de provenance	Hors-Québec	5	7,0000	1,4142	5,00	9,00
	Québec	59	8,2373	1,1498	3,00	9,00
Diplôme obtenu	Baccalauréat	30	7,9667	1,4735	3,00	9,00
	Microprogramme ou DESS	15	8,4667	0,6399	7,00	9,00
	Maitrise et doctorat	19	8,1570	1,0678	3,00	9,00
Domaine de formation	Enseignement au primaire ou préscolaire	31	8,3548	0,9503	5,00	9,00
	Enseignement en adaptation scolaire	20	7,9500	1,5381	3,00	9,00
	Psychopédagogie et autre	13	7,9231	1,1875	5,00	9,00
Cours méthodes quantitatives	Non	52	8,1923	1,2213	3,00	9,00
	Oui	12	7,9167	1,1645	5,00	9,00

4.4.4.2. Lien entre l'évaluation de la qualité et la compréhension des rapports

Dans cette partie, nous nous intéressons à savoir s'il y a un lien entre l'évaluation des rapports des participants et leur compréhension. L'idée est de savoir si ceux qui ont une évaluation des rapports plus positive ont plus tendance à donner des réponses correctes ou pas. Malgré le décalage dans le nombre des participants qui donnent des réponses correctes et incorrectes, les résultats descriptifs nous suggèrent que l'évaluation de la qualité des rapports est plus positive lorsque les participants donnent des réponses correctes et vice versa, et ceci se trouve dans toutes les questions, sauf pour la question 2 (tableau 1.13 de l'annexe 1). Dans cette question, l'évaluation de la qualité globale des rapports est légèrement plus élevée chez les participants ayant donné des réponses incorrectes (m=3,1667; s=0,1414 et m=3,1226; s=0,4403).

Si nous tenons compte de l'évaluation des facteurs séparés, nous constatons que l'évaluation est plus positive chez les participants ayant des réponses correctes que chez ceux qui donnent les réponses incorrectes dans tous les facteurs pour les questions 3, 6 et 7. Quant aux questions 1 et 5, l'évaluation est plus positive pour le groupe des réponses correctes dans presque tous les facteurs, sauf pour l'évaluation du contenu des graphiques (m=2,9912; s=0,6338 et m=3,1667; s=0,5400) et l'évaluation de l'ensemble du rapport (m=2,8684; s=0,5156 et m=3,000; s=0,6123). Cependant, pour la question 2, les participants ayant donné les réponses incorrectes ont une évaluation plus positive dans la plupart des facteurs comme pour les directives (m=3,1667; s=1,17850; le contenu des graphiques (m=3,1250; s=0,1767), le guide d'interprétation (m=3,3750; s=0,5303) et l'ensemble du rapport (m=3,1250; s=0,1767). Pour la question 4, l'évaluation est plus positive pour le groupe ayant donné des réponses correctes dans presque tous les facteurs, sauf l'évaluation du guide d'interprétation et de l'ensemble du rapport. Ces mêmes tendances ont été constatées pour la question 8.

Quant à la question 9, les participants qui donnent des réponses correctes ont une évaluation plus positive sur tous les facteurs, sauf pour l'évaluation du contenu des graphiques (m=2,9948; s=0,5980 et m=3,0156; s=0,7098). Le tableau 1.13 de l'annexe 1 permet de résumer ces résultats descriptifs.

Avec ces différentes tendances observées dans notre échantillon, nous avons fait des tests non-paramétriques de Mann-Whitney pour voir si ces différences sont statistiquement significatives dans cette population. Nous avons recouru à des tests non-paramétriques alternatifs du test T de Student parce que certains postulats pour le test T n'ont pas été respectés tels que la normalité des variables testées et la taille inégale des sous-échantillons (groupe des réponses correctes et incorrectes). Pour les tests non-paramétriques, aucun postulat n'est demandé, sauf l'exigence que les variables dépendantes soient numériques et ordonnées, ce qui est notre cas.

Les résultats des tests de Mann-Whitney (voir tableau 48) suggèrent qu'il n'existe pas de différence statistiquement significative dans l'évaluation de la qualité globale des rapports pour les questions 1 (U=151,50; p=0,648); 2 (U=54,00; p=0,757); 3 (U=22,00; p=0,122); 4 (U=85,00; p=0,836); 5 (U=53,50; p=0,743); 8 (U=295,50; p=0,548); et 9 (U=370,00; p=0,828) et ce, avec tous les facteurs étudiés. Il n'y a donc pas de lien entre leur compréhension des ces questions et l'évaluation des rapports. Cependant, pour la question 6, les tests de Mann-Whitney nous montrent que l'évaluation est statistiquement différente dans les deux groupes pour les facteurs tels que les directives (U=12,50; p=0,01); la présentation des graphiques (U=17,50; p=0,018) et la description du profil et des pistes d'intervention (U=23,50; p=0,029). Pour la question 7, le test est significatif pour l'évaluation de la qualité globale des rapports (U=69,00; p=0,005); l'évaluation des directives (U=110,00; p=0,049); l'évaluation de la description du profil (U=71,50; p=0,005) et l'évaluation du guide d'interprétation (U=98,50; p=0,028). Nous concluons donc que l'évaluation des participants de ces facteurs varie selon le fait qu'ils donnent des réponses correctes ou incorrectes.

Tableau 48. Lien entre l'évaluation de la qualité et la compréhension des rapports

Question sur la compréhension		Évaluation						
	Statistiques du test	**Qualité globale**	**Directives**	**Présentation des graphiques**	**Contenu des graphiques**	**Descrip-tion du profil**	**Guide inter-prétation**	**Ensemble du rapport**
1-Combien d'habiletés l'élève doit-il maitriser pour le test?	U de Mann-Whitney	151,50	143,00	124,50	141,50	123,50	7,50	153,50
	p	0,648	0,504	0,272	0,485	0,259	0,577	0,677
2-L'élève a-t-il bien maitrisé toutes les habiletés attendues?	U de Mann-Whitney	54,00	59,00	58,50	49,00	42,00	57,50	35,50
	p	0,757	0,906	0,892	0,612	0,434	0,860	0,300
3-L'élève a-t-il bien maitrisé l'habileté *Repérer des informations explicites*"?	U de Mann-Whitney	22,00	28,00	22,50	47,00	23,00	34,00	43,00
	p	0,122	0,180	0,124	0,559	0,127	0,273	0,457
4- L'élève a-t-il bien maitrisé l'habileté *Compréhension globale*?	U de Mann-Whitney	85,00	61,50	71,00	86,50	83,50	75,50	85,50
	p	0,836	0,330	0,512	0,873	0,797	0,606	0,847
5-Le niveau de maitrise de l'élève est-il identique pour les cinq habiletés?	U de Mann-Whitney	53,50	57,00	41,50	34,00	58,00	20,50	35,50
	p	0,743	0,844	0,425	0,275	0,876	0,104	0,300
6-Dans le graphique, quel est le niveau de maitrise de l'élève pour l'habileté *Interpréter*?	U de Mann-Whitney	32,50	12,50	17,50	79,00	23,50	71,00	42,00
	p	0,061	**0,01**	**0,018**	0,688	**0,029**	0,509	0,111
7-Si on considère que le seuil de réussite est 0,5, lequel de ces énoncés est vrai?	U de Mann-Whiney	69,00	110,00	122,00	131,50	71,50	98,50	116,50
	p	**0,005**	**0,049**	0,093	0,140	**0,005**	**0,028**	0,070
8. Dans le graphique sur le profil de maitrise des habiletés du groupe, pour l'habileté *Compréhension globale*, quel groupe a le pourcentage le plus élevé?	U de Mann-Whitney	295,50	252,50	244,50	258,00	329,00	322,50	322,50
	p	0,548	0,178	0,143	0,215	0,966	0,879	0,879
9. Pour l'habileté *Faire des inférences"*, le pourcentage des élèves ayant les niveaux de maitrise de 0,71 à 1 est-il le plus élevé?	U de Mann-Whitney	370,00	695,00	315,00	364,50	369,50	368,00	383,50
	p	0,828	0,818	0,281	0,760	0,820	0,801	0.994

4.5. Synthèse des résultats

Le chapitre 4 se compose de quatre parties principales qui visent à répondre à nos deux objectifs de recherche. Le premier objectif consiste à modéliser les données de 4762 élèves du Canada ayant participé au PIRLS 2011 à visée diagnostique. Ainsi, dans la partie 4.1, nous avons présenté les résultats du processus de l'élaboration des matrices Q1 et Q2 réalisées avec le panel d'experts. La matrice Q1 a été élaborée à partir de quatre processus en lecture identifiés pour le PIRLS 2011 tandis que la Q2 a été élaborée par un panel d'experts comprenant 3 membres. Les coefficients Kappa de Fleiss qui sert à déterminer le degré d'accord d'inter-juges ont été calculés avec le programme AgreeStat2015 pour chaque item. Au total, nous avons 9 items à un attribut; 21 items à deux attributs; 4 items à trois attributs et 1 seul item à cinq attributs

Avec ces deux matrices Q1 et Q2_finale, nous avons modélisé les données avec les modèles DINA et G-DINA et le logiciel Ox. Ces résultats ont été présentés dans la partie 4.2. Pour l'évaluation de l'ajustement des modèles aux données, le modèle G-DINA s'ajuste mieux que le DINA aux données et plus avec la Q2_finale que la Q1. Les modèles DINA et G-DINA s'ajustent adéquatement aux données avec toutes les deux matrices Q, avec la proportion correcte et la corrélation transformée. Cependant, selon les résultats du ratio de Log-odds, les modèles ne s'ajustent pas adéquatement aux données avec tous les items. Nous nous fions donc aux résultats de la proportion correcte et à la corrélation transformée étant donné que ces résultats sont plus ou moins semblables, ce qui aboutit à la même décision. Nous avons décidé de présenter l'estimation des paramètres et des profils de maitrise des habiletés des élèves avec le modèle G-DINA et la matrice Q2_finale. La moyenne du paramètre de pseudo-chance est de 0,36443, tandis que la moyenne du paramètre d'étourderie est de 0,2198 : autrement dit, en moyenne, l'élève a 21,98% de risque de donner des réponses incorrectes, même s'il maitrise tous les attributs exigés pour le test. La capacité diagnostique des items est de 0,4259. Selon les seuils proposés par Ravand et Widhiarso (2013), il y a 16 items de bonne qualité et 19 items de faible qualité. Ces résultats suggèrent donc que des travaux de raffinement de la matrice Q devraient être faits afin d'améliorer la qualité des items dans la perspective de les utiliser pour réaliser un diagnostic.

Quant aux profils de maitrise des habiletés de l'élève, le profil le plus fréquent est «11110» avec 25, 24% de nos élèves qui maitrisent les quatre premiers attributs, mais pas le cinquième attribut. Vient ensuite le profil des élèves qui ne maitrisent aucun des cinq attributs (18,78%). Le profil qui arrive au troisième rang est celui qui maitrise tous les attributs, soit 16,13% des élèves. Aucun participant ne fait partie d'un des quatre profils suivants: «10000»; «11000»; «01001»; «11010». L'habileté *Faire les inférences* est la plus maitrisée avec 66,62%, vient ensuite l'habileté *Repérer des informations implicites* (61,31 %). L'habileté *Interpréter* est classée au troisième rang (56,64%). L'habileté *Compréhension globale* est maitrisée par 50,31% et finalement, l'habileté *Vocabulaire et syntaxe* est la moins maitrisée avec seulement 41,31%.

Les parties 4.3 et 4.4 visent à répondre au deuxième objectif de recherche, soit élaborer et évaluer les rapports diagnostiques à partir des résultats des modélisations. Ainsi, dans la partie 4.3, nous avons commencé le travail avec le panel d'experts en reformulant des habiletés en lecture dans un langage familier aux enseignants, ce qui correspond bien aux documents ministériels du MELS et du CMEC. Nous avons donc reformulé légèrement les habiletés *Interprétation* et *Faire des inférences* pour qu'elles conviennent bien au contexte de notre étude. Après avoir examiné les sept profils-type, nous avons choisi le profil d'un élève qui maitrise bien les habiletés R*epérer des informations implicites; Faire des inférences; Vocabulaire et syntaxe* mais pas les habiletés *Compréhension globale* et *Interprétation* comme profil-type, car ce dernier ressemble le plus à un élève de la classe selon les experts, et qu'il correspond bien aux attentes pour un élève en 4ème année.

Avec ce profil, nous avons élaboré trois formats de rapports avec les différentes manières de présenter les graphiques. Le profil de l'élève et celui de groupe sont présentés dans deux graphiques séparés dans le format A tandis qu'ils sont dans le même graphique pour les formats B et C. Chaque rapport est accompagné d'un guide d'interprétation qui contient des informations plus détaillées sur les objectifs du rapport, du PIRLS 2011 et les directives pour lire des graphiques avec des exemples concrets. Nous avons décrit en détails ces trois formats ainsi que le guide d'interprétation.

La section 4.4 sert à décrire les perceptions des enseignants au primaire, des orthopédagogues et des conseillers pédagogiques sur les trois formats de rapports élaborés à l'aide d'un questionnaire soumis en ligne. L'instrument se compose de 55

questions à choix de réponse sur une échelle de type Likert portant sur 4 aspects: (a) la préférence des enseignants sur les formats de représentation des résultats diagnostiques; (b) l'évaluation des participants sur les trois formats de rapports élaborés et (c) leur compréhension sur les trois formats de rapports et (d) des informations sociodémographiques. Au total, les 98 participants sont répartis équitablement dans les trois formats de rapport. Parmi ces participants, 68,4% sont des enseignants au primaire, 15,3% sont des orthopédagogues et 16,3% sont des conseillers pédagogiques. Les participants sont regroupés en deux tranches d'âge: de 20 à 39 ans (49%) et 40 ans et plus (51%). En ce qui concerne l'ancienneté, 31,6% des participants ont moins de 10 ans d'expérience; 52% des participants ont de 11 à 20 ans d'expérience et 16,3% des participants ont plus de 20 ans d'expérience.

Les analyses descriptives sur la préférence suggèrent que le format B est le préféré pour les quatre questions. Les tests de Khi-deux montrent qu'il n'y a pas de lien statistiquement significatif entre la préférence des rapports et les variables sociodémographiques telles que les postes occupés, les régions de provenance, les tranches d'âge, l'ancienneté, le diplôme obtenu, les domaines de formation, le genre et le suivi ou non des cours en méthodes quantitatives.

Les analyses factorielles exploratoires avec la méthode d'extraction par factorisation en axes principaux et la rotation orthogonale de Varimax ont été réalisées avec 32 questions sur l'évaluation de la qualité des rapports. Nous avons d'abord effectué des analyses factorielles pour l'ensemble des 32 items pour retrouver la structure globale du questionnaire. Par la suite, nous avons réalisé des analyses factorielles dans chaque section séparée dans le but de trouver un moyen pour regrouper les items selon les parties à évaluer du rapport. Les résultats suggèrent que les items ont un bon ajustement aux facteurs proposés pour l'ensemble des questions et pour chaque partie séparée, avec un très bon coefficient d'Alpha de Cronbach. Ces résultats suggèrent sept regroupements des questions, dont un pour l'ensemble de trente questions (évaluation globale de la qualité des rapports) et six pour les sous-parties (évaluation des directives, évaluation de la présentation des graphiques, évaluation du contenu des graphiques, évaluation de la description du profil et des pistes d'intervention, évaluation du guide d'interprétation et évaluation de l'ensemble des rapports). Nous avons donc éliminé deux questions avec des indices de saturation moins de 0,30.

L'évaluation des rapports a été réalisée avec ces sept facteurs en tenant compte des variables sociodémographiques ainsi que la préférence des rapports des participants. Les participants ont des perceptions très positives sur les rapports évalués. Les participants semblent avoir une évaluation de la qualité plus positive pour le format B, puis le format C et enfin le format A. Cependant, les tests d'Anova à un facteur de classification et les tests T indiquent qu'il n'y a pas de différence statistiquement significative dans les perceptions des participants sur l'évaluation de la qualité globale des rapports en général et sur les facteurs séparés, en particulier selon les variables sociographiques, malgré les différentes tendances observées dans notre échantillon. Par contre, les résultats du test T nous suggèrent qu'il y a une différence statistiquement significative entre la préférence et l'évaluation de la qualité des rapports. Les participants qui évaluent le format de rapport qu'ils préfèrent ont des perceptions plus positives sur l'évaluation des rapports.

L'évaluation de la compréhension suggère que les participants ont une très bonne compréhension de trois formats de rapports. Les participants ont une meilleure compréhension du format A que des formats B et C. Quant au lien entre l'évaluation de la qualité globale des rapports et l'évaluation des facteurs séparés et la compréhension des questions, les résultats descriptifs suggèrent que l'évaluation des rapports est plus positive pour le groupe qui donne des réponses correctes que pour le groupe de ceux avec les réponses incorrectes. Les résultats des tests non-paramétriques de Mann-Whitney sont significatifs avec l'évaluation des directives, de la présentation des graphiques et de la description du profil et des pistes d'intervention dans la question 6. Pour la question 7 portant sur le seuil de réussite, le test est significatif pour l'évaluation de la qualité des rapports, l'évaluation des directives, l'évaluation de la description du profil et l'évaluation du guide d'interprétation. Nous concluons donc que l'évaluation des participants de ces facteurs varie selon le fait qu'ils comprennent ou pas.

Les résultats présentés dans ces quatre parties nous permettent de répondre aux deux objectifs ainsi qu'à quatre questions de recherche. Les discussions entre le lien des résultats obtenus avec ces objectifs et ces questions de recherche ainsi qu'avec d'autres recherches empiriques sur le sujet vont être présentées dans le chapitre suivant. Les apports scientifiques et pratiques de notre étude et ses limites seront également abordés.

CHAPITRE 5. INTERPRÉTATION, DISCUSSION ET CONCLUSION

Ce chapitre est consacré à la discussion des résultats obtenus dans le chapitre 4. Premièrement, nous discutons des résultats à la lumière des intérêts de recherche et en lien avec quatre questions spécifiques de notre recherche. Par la suite, nous présentons les apports scientifiques et pratiques de notre étude. Finalement, nous abordons les limites ainsi que les recommandations pour les recherches futures sur le sujet.

5.1. Interprétation et discussion

À titre de rappel, cette étude vise deux objectifs de recherche généraux, soit (a) modéliser des données du PIRLS 2011 à visée diagnostique et (b) élaborer et évaluer des rapports à visée diagnostique destinés aux enseignants à partir des résultats des modélisations. Ainsi, ces deux objectifs généraux se transforment en quatre objectifs spécifiques qui consistent à: (a) vérifier la capacité diagnostique du PIRLS 2011; (b) fournir des profils de maitrise des élèves du Canada ayant fait le PIRLS 2011; (c) élaborer des rapports à visée diagnostique destinés aux enseignants et (d) analyser les perceptions des enseignants au primaire, des conseillers pédagogiques et des orthopédagogues sur les formats de rapports élaborés. Effectivement, les analyses effectuées et les résultats obtenus nous ont permis de répondre à quatre questions de recherche suivantes:

(1)- ***Quelle est la capacité diagnostique des items du test PIRLS 2011?***

Les résultats des modélisations obtenus avec les modèles DINA et G-DINA nous ont permis d'évaluer l'ajustement relatif et absolu de ces deux modèles aux données ainsi que les paramètres d'étourderie et de pseudo-chance, ce qui nous a amené à estimer la qualité diagnostique de 35 items du test PIRLS 2011.

(2)-Quels sont les profils de maitrise des habiletés des élèves canadiens ayant participé au PIRLS 2011?

Les résultats des modélisations avec les modèles DINA et G-DINA nous ont également fourni les probabilités de maitrise pour chacune des cinq habiletés en lecture identifiées pour le test PIRLS. L'habileté *Faire des inférences* est la mieux maitrisée (66,62%) tandis que l'habileté *Vocabulaire et syntaxe* est la moins maitrisée avec une probabilité de 41,31%. De plus, ces résultats nous ont renseigné en détail sur le profil de maitrise sur les cinq habiletés de chaque élève, ce qui nous a permis de sortir les profils les plus représentatifs des élèves canadiens ayant fait la section du livret 13 du test PIRLS 2011.

(3)-*Comment élaborer des rapports à visée diagnostique destinés aux enseignants à partir des résultats obtenus des modélisations?*

L'élaboration des rapports a été réalisée en adaptant les étapes du cadre de référence de Zapata-Rivera, VanWinkle et Zwick (2012), puis en consultant un panel d'experts composé de cinq membres provenant de milieux différents. Ce processus d'élaboration des rapports a été effectué à travers quatre rencontres en équipe dont les objectifs visent à: (1) reformuler les cinq habiletés en lecture identifiées pour le test PIRLS 2011 dans un langage accessible pour les enseignants en salle de classe; (2) choisir un profil-type qui fait l'objet de l'élaboration des rapports diagnostiques parmi les profils de maitrise des habiletés obtenus à partir des modélisations; (3) proposer des éléments du contenu et de la forme pour présenter dans des rapports diagnostiques à partir du protocole du design et (4) pré-valider des formats de rapports élaborés ainsi que le questionnaire conçu sur l'évaluation des rapports. À la fin de ce processus, nous avons réussi à élaborer trois formats de rapports à partir du profil-type choisi et un questionnaire validé et administré en ligne destiné aux enseignants au primaire, aux conseillers pédagogiques et aux orthopédagogues nous permettant de répondre à la 4ème question de recherche.

(4)-*Quelles sont les perceptions des enseignants au primaire, des conseillers pédagogiques et des orthopédagogues sur les formats de rapports diagnostiques?*

La collecte de données a été réalisée auprès de 98 enseignants au primaire, d'orthopédagogues et de conseillers pédagogiques à l'aide d'un questionnaire créé sur la plate-forme Survey Monkey. Les analyses descriptives et inférentielles réalisées avec SPSS nous ont aidé à comprendre les perceptions des participants des trois formats de rapports élaborés en ce qui concerne leur préférence; leur évaluation de la qualité et leur

compréhension des rapports selon des informations sociodémographiques. Les lignes suivantes sont consacrées à la discussion plus en détail des questions de recherche.

5.1.1. Éléments qui influencent l'ajustement des modèles aux données et des paramètres d'items

Afin d'évaluer l'ajustement des modèles de DINA et G-DINA aux données, nous avons recouru à six statistiques, dont trois statistiques, -2LL, AIC et BIC pour l'évaluation de l'ajustement relatif, et trois statistiques, les scores Z de la proportion correcte (prop), de la corrélation transformée (Z (Corr) et du ratio log-odds (Log (OR)), pour l'évaluation absolue des modèles aux données. L'évaluation de l'ajustement relatif des modèles permet de déterminer le modèle le plus approprié qui s'ajuste le mieux aux données parmi plusieurs modèles concurrents avec la valeur la plus petite de ces trois statistiques tandis que l'évaluation de l'ajustement absolu nous renseigne à quel niveau le modèle choisi peut s'ajuster aux données. Plus les trois statistiques s'approchent de 0, plus l'ajustement absolu du modèle aux données est élevé. Nous avons évalué l'ajustement relatif et absolu des modèles de DINA et G-DINA aux données avec la matrice Q1 qui a été élaborée à partir du cadre de référence du PIRLS 2011 et la matrice Q2, élaborée avec le panel d'experts.

Ainsi, pour ce qui est de l'évaluation de l'ajustement relatif, les résultats obtenus suggèrent que le modèle G-DINA s'ajuste mieux que le DINA aux données et davantage avec la matrice Q2_finale qu'avec la matrice Q1. Ces résultats corroborent très bien ceux des recherches précédentes réalisées avec ces deux modèles en maths (Basokcu, 2014; Ma, Lacognagelo et de la Torre, 2016; Chen, de la Torre et Zhang, 2013). Dans ces recherches, les modèles généralisés et saturés comme le cas de G-DINA et GDM s'ajustent toujours mieux aux données que d'autres modèles plus spécifiques comme DINA et DINO, etc. La raison qui l'explique est que les modèles généralisés ne tiennent pas compte de la relation restreinte, c'est-à-dire conjonctive ou disjonctive des attributs afin de répondre correctement à un item (Ravand, Barati et Widhiarso, 2012). Autrement dit, le modèle G-DINA assouplit l'hypothèse de la probabilité égale de réponses correctes lorsque le sujet ne maitrise pas tous les attributs demandés pour répondre correctement à

cet item. Ainsi, même si les sujets ne maitrisent pas tous les attributs qu'il faut pour un item, la probabilité d'y répondre correctement peut varier d'un sujet à l'autre vu le nombre et la nature des habiletés qui ne sont pas maitrisées (Loye, 2010). Une autre raison est que le modèle G-DINA est souvent moins influencé que d'autres modèles spécifiques lorsqu'il y a des changements dans les matrices Q (entre Q1 et Q2_finale) (Basokcu, 2014), ce qui témoigne de son meilleur ajustement relatif aux données par rapport au modèle DINA. Cependant, bien que ces modèles généralisés s'ajustement mieux aux données, les modèles MCD spécifiques, lorsqu'ils sont utilisés correctement, donnent la possibilité d'obtenir des interprétations plus simples, plus stables et de fournir des profils de maitrise des habiletés plus exacts (Ma, Lacognagelo et de la Torre, 2016). Il est donc intéressant de vérifier si nous pouvons utiliser des modèles MCD spécifiques pour fournir des profils de maitrise des habiletés aux apprenants tout en gardant un ajustement adéquat de ces modèles aux données. Une des suggestions est d'utiliser le test de Wald pour chaque item afin de déterminer si un MCD généralisé peut être remplacé par un MCD spécifique sans perdre sa qualité de l'ajustement aux données (Ma, Lacognagelo et de la Torre, 2016).

Quant à l'évaluation de l'ajustement absolu des modèles DINA et G-DINA aux données avec les matrices Q1 et Q2_finale, les statistiques de la proportion correcte et de la corrélation transformée montrent que ces deux modèles s'ajustent adéquatement aux données. Avec ces deux statistiques, le modèle G-DINA semble s'ajuster mieux que le DINA et l'ajustement est meilleur avec la matrice Q2_finale qu'avec la matrice Q1. En ce qui concerne le ratio de Log-odds, les résultats suggèrent que les modèles DINA et G-DINA ne s'ajustent pas adéquatement aux données avec tous les items. Dans le cadre de notre étude, nous nous fions aux résultats de la proportion correcte et de la corrélation transformée étant donné que ces résultats sont plus ou moins semblables, ce qui nous amène à la même décision sur l'ajustement adéquat des modèles aux données. Cependant, le fait que ces trois statistiques puissent fournir des résultats permettant de prendre des décisions différentes sur le rejet ou non de l'hypothèse nulle sur l'ajustement adéquat des modèles aux données nous amène à nous questionner sur la fiabilité et la sensibilité de ces trois statistiques. Ce constat a été souligné dans la recherche de Chen, de la Torre et Zhang (2013) lorsque ces trois statistiques n'aboutissent pas à la même décision sur l'ajustement adéquat des modèles aux données. Cette recherche suggère également qu'il y

a probablement des problèmes d'inexactitude dans la matrice Q qu'il est donc intéressant de détecter avec des techniques plus sophistiquées telles que le test de Wald. De plus, il est important de vérifier quelle statistique pourrait être fiable dans l'évaluation de l'ajustement relatif et absolu selon le type de MCD utilisé et la nature de réponses (dichotomiques ou polychotomiques) ainsi que le type et le nombre d'attributs identifiés.

Les résultats sur l'évaluation de l'ajustement absolu des modèles aux données montrent également que l'ajustement est meilleur avec la matrice Q2_finale qu'avec la matrice Q1, ce qui souligne bien le caractère multidimensionnel de la lecture. En effet, pour la matrice Q1 qui est élaborée en se basant sur les quatre processus en lecture proposés dans le cadre de référence du PIRLS 2011, comme chaque item correspond à un seul processus de lecture, les 35 items sont des items à un seul attribut. Avec la matrice Q2_finale qui contient cinq attributs, nous avons au total neuf items à un attribut; vingt et un items à deux attributs; quatre items à trois attributs et un seul item à cinq attributs. Ainsi, selon le panel d'experts, à part les neuf items à un seul attribut, les vingt-six items qui restent demandent au moins un attribut de plus pour pouvoir donner la réponse correcte. Par exemple, pour l'item 3, à part l'attribut *Repérer des informations explicites* tel qu'identifié dans la matrice Q1, l'élève a besoin de l'attribut *Interprétation* afin d'y répondre correctement. En ce qui concerne l'item 14, deux attributs *Compréhension globale* et *Vocabulaire et Syntaxe* ont été ajoutés en plus de l'attribut *Interprétation* qui est initialement identifié dans le cadre de référence du PIRLS 2011 (Q1). Ainsi, ce problème souligne l'importance du raffinement des habiletés dans la matrice Q, car plus ces attributs sont détaillés, plus fines et enrichies sont des informations diagnostiques obtenues (Lee et Sawaki, 2009; Li, 2011). Cette idée est très appropriée dans le cas du test PIRLS 2011 avec la matrice Q1, car certains processus de lecture proposés sont encore très généraux et nécessitent d'être séparés en deux ou trois habiletés comme le processus *Examiner et évaluer le contenu, le langage et les éléments textuels* ou le processus *Interpréter et combiner des idées et des informations*. Cependant, un nombre élevé d'attributs peut engendrer des problèmes sur la capacité des modélisations, un facteur important à considérer autre que la pertinence des profils de maitrise des habiletés lorsque nous faisons des modélisations avec des MCD.

Ainsi, un défi des concepteurs des tests est de maintenir l'équilibre entre le nombre d'attributs identifiés et la longueur du test, c'est-à-dire qu'il faudrait ajouter plus d'items lorsque le nombre des habiletés sous-jacentes identifiées pour le test est grand (Li, 2011). Parfois, pour un même test, le nombre d'attributs identifiés peut être différent d'un chercheur à l'autre. Par exemple, pour le test de MELAB de 20 items, Gao (2006) suggère dix attributs tandis que Li (2011) identifie seulement cinq attributs. C'est le même constat observé dans la recherche de Jang (2009) et Sawaki, Kim et Gentile (2009) avec le test TOEFL iBT lorsque Jang (2009) a identifié neuf habiletés tandis que Sawaki et ses collaborateurs (2009) en ont identifié seulement quatre. Ceci renvoie à la granularité des habiletés et des sous-habiletés en lecture, car les définitions détaillées des habiletés et des sous-habiletés du test dépendent des balises théoriques (par exemple la représentativité du construit) et des techniques (accessibilité aux items du test) ainsi que des conditions pratiques (objectifs et contexte d'utilisation des rétroactions diagnostiques) (Jang, 2009; Li, 2012). Ceci dit, avec les 35 items du test PIRLS 2011, il est possible d'augmenter le nombre d'attributs en précisant encore plus en détail les habiletés identifiées pour voir si la nouvelle matrice Q serait meilleure que les matrices Q1 et Q2_finale. Parfois, lorsque les conditions sur l'élaboration de la matrice Q ont été respectées (au moins trois items par attribut, par exemple), la décision sur le choix de la matrice Q renvoie aux résultats fournis par les modélisations, autrement dit, nous devons laisser les MCD décider quelle matrice Q s'ajuste le mieux aux données, c'est ce que nous avons fait avec les deux matrices Q1 et Q2_finale du PIRLS 2011.

En ce qui concerne les paramètres de pseudo-chance et d'étourderie, ce qui renvoie à la qualité diagnostique des items du PIRLS 2011, les résultats des modélisations suggèrent qu'en moyenne la qualité diagnostique des items est de 0,4259. Dans l'ensemble, ces deux paramètres de pseudo-chance et d'étourderie sont encore élevés, ce qui signifie que la capacité diagnostique des items est moyenne, probablement parce que le test n'a pas été initialement conçu à visée diagnostique. Cette idée a été clairement confirmée dans les recherches avec les tests à grande échelle en lecture de Li (2011), Jang (2008) et Ravand et Widhiarso (2013), Huang et coll. (2014). Si nous regardons plus en détail les items, il y a deux items de bonne qualité; deux items de moyenne qualité; douze items de faible qualité et dix-neuf items qui sont encore problématiques selon les balises

proposées par de la Torre (2009) et de la Torre et Douglas (2008). Cependant, selon les seuils proposés par Ravand et Widhiarso (2013), il y a seize items de bonne qualité et dix-neuf items de faible qualité. Plus d'une moitié des items (11 sur 19) se trouvent dans la partie 1 du test dont le but de lecture est de *lire pour l'expérience littéraire*, tandis que la partie 2 qui vise à *lire pour acquérir et utiliser des informations* contient huit items problématiques. Ceci peut nous amener à nous questionner sur le lien entre l'objectif de la lecture, le type de texte ainsi que ses caractéristiques et la qualité psychométrique des items. En effet, cette observation a été mise en évidence dans plusieurs recherches sur l'influence des éléments textuels et la compréhension en lecture (Urquhart et Weir, 1998; Jang 2009). Par exemple, Freedle et Kostin (1993) ont rapporté qu'au moins 33% de la variance de la difficulté de l'item du test TOEFL RC a été expliqué par des variables associées au contenu et aux structures des passages. Alderson, Percsich et Szabo (2000) ont soutenu que la compétence en lecture entrainait la capacité de reconnaitre comment les idées sont présentées dans le texte et de comprendre les intentions sous-jacentes de l'auteur dans une séquence d'idées. Jang (2009) a confirmé que les textes avec différentes structures organisationnelles rhétoriques (par exemple, expositoires, chronologiques, comparatives) ont déterminé différents processus cognitifs chez les élèves. Dans cette étude, le passage de la partie 2 du test est de nature informative avec une structure organisationnelle plus claire et cohérente que celui de la partie 1, qui est de nature fictive et vaguement structuré avec de nombreuses phrases de conversation entre les personnages. Ceci peut être un des éléments qui explique la meilleure capacité diagnostique des items de la partie 2 par rapport à ceux de la partie 1 du test.

Un autre élément à considérer est le type de question utilisé dans le test. Les QCM semblent présenter un paramètre de pseudo-chance plus élevé que les questions à développement. Ceci a été justifié par le fait que parmi les dix-neuf items problématiques, dont douze QCM et sept questions à développement, il y a onze QCM ayant le paramètre de pseudo-chance plus élevé que le paramètre d'étourderie. L'item QCM a des paramètres de pseudo-chance et d'étourderie plus ou moins semblables (0,30 et 0,32). Par contre, pour les sept questions à développement, quatre questions ont un paramètre d'étourderie plus élevé que celui de pseudo-chance. Cependant, selon Huang et coll., (2014), le paramètre de pseudo-chance renvoie non seulement aux caractéristiques

des items, mais aussi à l'habileté de l'élève, car la pseudo-chance est l'interaction entre la tendance d'un item qui suscite les devinettes et la capacité de deviner de l'élève. Ainsi, pour les QCM, les élèves ont plus tendance « à procéder par déduction » qu'à répondre de manière aléatoire. De plus, les élèves avec un niveau de maitrise plus élevé peuvent avoir une plus grande probabilité de deviner la réponse correctement que ceux qui sont moins compétents. Par contre, les élèves moins compétents sont susceptibles d'être plus influencés par les facteurs de distraction que ceux qui sont plus compétents (Huang et coll., 2014). Ces résultats pourraient être intéressants dans l'évaluation diagnostique cognitive, notamment la spécification des habiletés sous-jacentes des épreuves, car les différentes variables textuelles peuvent susciter différentes habiletés cognitives; le choix des types des questions, quant à lui, pourrait influencer la qualité diagnostique de ces items (Jang, 2009). Nous devons donc tenir compte de ces facteurs lors de la conception des épreuves à visée diagnostique ainsi que l'élaboration de la matrice Q. Nous discutons sur le profil de maitrise des habiletés des élèves dans les lignes qui suivent.

5.1.2. Reformulation des habiletés en lecture; équivalence entre le programme de formation en anglais et le document du MELS et ses influences sur la matrice Q et les profils de la maitrise des habiletés des élèves

Étant donné que les habiletés identifiées pour le test en anglais ont été initialement définies en anglais et ensuite traduites en français pour être conformes à l'étude, nous avons reformulé ces habiletés dans un langage familier aux enseignants en salle de classe afin d'éviter les confusions que les traductions peuvent apporter. Cette étape a suscité des discussions très intéressantes chez les membres du panel d'experts 2 compte tenu des différences dans le programme de formation du français langue d'enseignement au Québec et celui de l'anglais langue d'enseignement dans d'autres provinces. Les discussions tournent autour de la différenciation de deux habiletés *Interprétation* et *Faire des inférences* ainsi que du niveau de difficulté de ces deux habiletés. Selon CMEC (2012), le processus *Faire des inférences* permet aux lecteurs de combler des lacunes relatives au sens en déduisant des informations à partir du texte, ce qui exige peu d'efforts chez les lecteurs, autrement, ce sont ceux qui sont habiles qui peuvent généralement le faire de façon automatique. Les tâches de ce processus consistent à déduire qu'un

évènement en a entrainé un autre, à tirer des conclusions quant à l'idée principale dans une série d'arguments, etc. Ainsi, cette habileté correspond bien au traitement des informations implicites telle que définies dans les documents du MELS ou l'habileté *Making Inferences/Interpreting texts* proposé par le ministère de l'Éducation de Ontario. Par contre, selon le CMEC (2012), l'interprétation demande aux lecteurs d'avoir une compréhension plus approfondie du texte en combinant les informations présentées dans les textes et les connaissances antérieures. Cette description de l'habileté *Interprétation* correspond plus ou moins à la définition du MELS et à celle de l'habileté *Extended understanding* proposée par le ministère de l'Éducation de l'Ontario. Ainsi, cette habileté représente un degré de difficulté plus élevé que l'habileté *Faire des inférences,* car les liens à établir ne sont pas seulement implicites, mais peuvent être ouverts à l'interprétation des élèves, ce qui peut varier selon le lecteur ou selon ses expériences personnelles (CMEC, 2012). Le degré de difficulté et la complexité des habiletés sont des éléments explicatifs des différents niveaux de maitrise des habiletés chez les élèves.

En effet, parmi les cinq habiletés identifiées par le PIRLS 2011, les habiletés *Faire des inférences* (66,62%) et *Repérer des informations explicites* (61,31%) sont les habiletés les plus maitrisées. Autrement dit, les élèves canadiens ayant fait le livret 13 du PIRLS ont 66,62 % de chance de maitriser l'habileté *Faire des inférences et 61,31% de chance de maitriser l'habileté Repérer des informations explicites*. Viennent ensuite les habiletés *Interprétation* (56,64%) et *Compréhension globale* (50,31%). L'habileté *Vocabulaire et syntaxe* est la moins maitrisée avec 41,31%. Les résultats obtenus corroborent en grande partie les théories en lecture et le degré de difficulté des habiletés, car les habiletés *Repérer des informations implicites* et *Faire des inférences* qui ne demandent qu'une compréhension locale du texte, sont considérées comme plus faciles que les habiletés *Compréhension globale* et *Interprétation* qui exigent un niveau de compréhension plus générale du texte. De plus, le fait que l'habileté *Vocabulaire et syntaxe* est la moins maitrisée semble cohérent avec les résultats de la recherche de Li (2011) avec le test de METLAB lorsque les attributs *Vocabulaire* et *Syntaxe* sont les moins maitrisés (par 25,6% et 28,7%) contre une probabilité de maitrise de 40,1% pour l'attribut *Repérer des informations explicites* et de 32,3% pour l'attribut *Comprendre des informations implicites*. L'habileté *Vocabulaire et syntaxe* semble la plus difficile à

maitriser, ce qui a été renforcé par le constat que le manque de vocabulaire est l'obstacle majeur à la compréhension de la lecture (Garcia, 1991). Une règle de base est que les lecteurs doivent connaitre 95% des mots d'un texte pour lire le texte avec succès (Grabe, 2009; Li, 2012). De plus, conformément à la littérature sur la lecture ainsi que la discussion du panel d'experts, la présente étude a révélé que l'habileté *Interprétation* est plus difficile que *Faire des inférences* et *Repérer des informations explicites*, elle nécessite donc plus de processus cognitifs, ce qui explique le niveau de maitrise moins élevé des élèves pour ces deux habiletés. Cependant, bien que l'habileté *Repérer des informations explicites* soit jugée plus facile que *Faire des inférences*, elle est moins maitrisée par les élèves avec 5 % de différence. La raison est que parmi les seize items dont l'habileté *Repérer des informations explicites* a été identifiée comme nécessaire, une moitié des questions (8 items) sont des items à un seul attribut. Par contre, il y a seulement un item à un attribut parmi les douze items pour lequel l'habileté *Faire des inférences* a été identifiée comme nécessaire pour obtenir une réponse correcte. Les onze items qui restent sont des items à deux, trois, quatre et cinq attributs. À notre avis, c'est la complémentarité de ces attributs au sein d'une même question qui contribue à augmenter la probabilité de répondre correctement aux items et au niveau de maitrise des habiletés. Ceci peut être un des éléments qui explique que l'habileté *Faire des inférences* est plus maitrisée que *Repérer des informations explicites*, bien qu'elle soit plus difficile.

Le profil le plus représentatif des élèves (25,24%) est celui qui regroupe ceux qui maitrisent les quatre habiletés *Repérer des informations explicites*, *Compréhension globale*, *Interprétation, Faire des inférences*, mais pas *Vocabulaire et syntaxe*. Vient ensuite le profil des élèves qui ne maitrisent aucun des cinq attributs (18,78%). Le profil qui arrive au troisième rang est celui qui détermine les élèves qui maitrisent l'ensemble des attributs, soit 16,13% des élèves. Ces résultats correspondent bien à nos attentes, étant donné que l'habileté *Vocabulaire et Syntaxe* est l'habileté qui représente le plus grand défi, donc elle s'avère la moins maitrisée chez les élèves. Les résultats des modélisations nous suggèrent quatre profils peu probables chez les élèves.

Selon nous, c'est la nature compensatoire des habiletés qui explique les résultats de quatre profils peu probables chez les élèves. Le profil «10000» correspondant à l'élève qui ne maitrise que l'habileté *Repérer des informations explicites*; il est le moins présenté

dans notre étude parce qu'une moitié des questions dans lesquelles cette habileté a été identifiée comme nécessaire est constituée d'items à deux attributs et plus, c'est-à-dire que l'élève a besoin d'au moins une autre habileté pour répondre correctement aux items. Le profil «11000» correspond aux élèves qui ne maitrisent que les deux habiletés *Repérer des informations explicites* et *Compréhension globale*; cette combinaison est assez rare, car l'habileté *Compréhension globale* fait partie de la compréhension des informations implicites et dans notre matrice Q, elle est souvent en lien avec *Faire des inférences* ou *Interprétation*. Il est donc peu probable que l'élève maitrise la compréhension globale, mais pas *l'interprétation* et qu'il ne peut pas, non plus, faire des inférences. Quant au profil «01001» qui revoie à l'élève qui maitrise seulement les deux habiletés *Compréhension globale* et *Vocabulaire et syntaxe*, ceci est un profil peu présenté chez les élèves du PIRLS, parce l'habileté *Vocabulaire et syntaxe* est bien souvent nécessaire pour répondre aux questions sur *l'interprétation* et *faire des inférences*, mais elle est moins sollicitée pour la *Compréhension globale*. Finalement, le profil «11010» est celui qui maitrise les habiletés *Repérer des informations explicites*; *Compréhension globale* et *Faire des inférences*, mais pas *Interprétation* et *Vocabulaire et syntaxe*. Ceci peut être expliqué par le fait que dans notre étude, l'habileté *Interprétation* est toujours liée à une des quatre habiletés, ce qui fait qu'il est rare de maitriser les trois autres habiletés sans maitriser l'habileté *Interprétation*.

Si nous faisons le lien avec les modèles théoriques en lecture, nous constatons que les modèles interactifs sur lesquels le cadre de référence de PIRLS s'appuie mettent en lumière cette nature de complémentarité des habiletés en lecture. Par exemple**,** le modèle contemporain en lecture (Irwin, 1986; Giasson, 1991) suppose que la compréhension en lecture se réalise dans l'interaction de trois composantes: le texte, le lecteur et le contexte ou le lecteur joue le rôle actif. Les modèles interactifs en lecture (Rumelhart, 1977), la théorie du schéma (Alderson et Pearson, 1984, 1988) et le modèle de construction et intégration en lecture (Kintsch, 1988) considèrent que le lecteur doit mobiliser un ensemble des connaissances linguistiques et culturelles, des stratégies cognitives et métacognitives pour construire du sens. Et le modèle interactif compensatoire de Stanovich (1980) affirme qu'un déficit dans une stratégie de lecture particulière du lecteur

entrainera un plus grand recours à d'autres sources de connaissances quel que soit son niveau de traitement hiérarchique.

De plus, le constat sur la complémentarité ou non des attributs en lecture a suscité beaucoup de discussions sur la question de choisir un MCD compensatoire ou non-compensatoire lors des modélisations des tests en lecture. En réalité, il n'existe pas encore de réponse claire à ce propos, car les MCD permettent à la fois des relations compensatoires et non-compensatoires des habiletés. Par exemple, selon Jang (2005), les habiletés en lecture du test TOEFL iBT doivent être un mélange des interactions compensatoires et non-compensatoires. En réalité, les modèles non-compensatoires ont plus tendance à être choisis parce qu'ils peuvent générer des informations diagnostiques plus fines (Li et Suen, 2012; Li, 2011). Cependant, Lee et Sawaki (2009) ont montré que peu importe le type de MCD utilisé (compensatoire ou non), la classification des profils de maitrise des élèves doit être similaire. La preuve est qu'ils ont utilisé trois différents modèles (compensatoires et non-compensatoires) pour le test TOEFL iBT en lecture et en écoute, et que les résultats obtenus sur les profils des élèves sont très semblables d'un modèle à l'autre. En outre, ces interactions dépendent également du degré de difficulté des habiletés jugées nécessaires pour répondre correctement à l'item, ce qui peut varier d'un item à l'autre (Li, 2011; Li et Suen, 2012). Ainsi, dans leur recherche, le fait de choisir le modèle compensatoire ou non-compensatoire n'a aucune conséquence sur les résultats de la classification des profils des élèves.

5.1.3. Retour aux cadres de référence de l'élaboration et de l'évaluation des rapports et aux critères d'un bon rapport

Le processus de l'élaboration et de l'évaluation des rapports se base principalement sur les cadres de référence de Zapata-Rivera, VanWinkle et Zwick (2012) et de Hambleton et Zenisky (2010) portant sur quatre étapes : (a) identifier le public visé et analyser ses besoins; (b) documenter les recherches et les formats de rapports qui existent; (c) développer des rapports et (d) évaluer des rapports. Dans le cadre de notre étude, nous n'avons pas réalisé d'une manière empirique l'étape de l'identification du public visé et de l'analyse de ses besoins. Cependant, nous avons fait un état de lieux sur les formats de rapports conçus à visée diagnostique et non diagnostique qui existent

actuellement au Québec. Nous nous sommes renseignée également auprès des membres du panel d'experts provenant de différents milieux éducatifs sur les besoins de recevoir des rapports diagnostiques. Le constat auquel nous arrivons est qu'il n'existe pas encore de formats de rapports conçus à visée diagnostique, même si le volet de soutien à l'apprentissage constitue un souci primordial au MELS. De plus, bien que l'administration du test PIRLS 2011 prenne beaucoup de temps et d'argent, les résultats ne servent qu'à des fins de recherche et à l'élaboration des politiques éducatives. Autrement dit, le CMEC ne communique pas de résultats à titre personnel aux élèves ni aux écoles qui ont participé au test ni aux ministres de l'Éducation. Par contre, chaque province peut choisir sa propre manière de communiquer des résultats (Labrecque et al., 2012). Cette situation justifie donc la nécessité d'avoir une moyenne de présentation des résultats à titre individuel des épreuves à grande échelle comme le PIRLS 2011.

L'étape de la documentation des recherches et des formats de rapports qui existent a été effectuée par la recension des écrits sur les cadres de référence et des recherches empiriques de l'élaboration et de l'évaluation des rapports à la fois à visées diagnostique et non-diagnostique. Cette étape nous a permis de sélectionner le cadre de référence le plus approprié pour le développement des rapports de notre recherche ainsi que de développer ce qui nous apparait comme les quatre critères d'un bon rapport, à savoir l'accessibilité, la lisibilité, la validité et l'utilité. Les recherches empiriques nous ont également proposé des principes de présentation du contenu et de la forme qu'il faut retenir dans les rapports. Ces principes portent tant sur des éléments textuels que non textuels tels que proposés dans la recherche de Robert et Gierl (2010). Des informations textuelles englobent la structure interne qui renvoie aux techniques d'organisation et de séquence tandis que la structure externe porte sur l'accessibilité à la structure, la typographie et la disposition du texte. Les éléments non textuels comprennent des tableaux et des graphiques qui servent à présenter des résultats. Le résultat final de cette étape de documentation est que nous sommes arrivée à concevoir un protocole du design sur l'élaboration des rapports diagnostiques qui précisent des éléments textuels et non textuels que nous avons utilisés pour travailler avec le panel d'experts lors de l'étape suivante.

Le développement des rapports a été réalisé avec un panel d'experts comprenant cinq membres provenant de différents milieux éducatifs à travers quatre rencontres en équipe. À la différence des recherches réalisées dans la même perspective diagnostique que celles de Jang (2005) et de Robert et Gierl (2010), nous avons commencé le travail avec le panel d'experts par la reformulation des habiletés en lecture et le choix d'un profil type à présenter dans les rapports. Cette étape permet d'assurer la cohérence entre les résultats des modélisations obtenus et le portrait réel des élèves en salle de classe, étant donné que nous avons travaillé avec deux panels d'experts différents dans les deux étapes. Les discussions sur les informations du fond et de la forme pour présenter dans les rapports ont été menées au cours de trois autres rencontres en équipe. À la fin de ces rencontres, nous avons réussi à élaborer trois formats différents de rapports diagnostiques à partir d'un profil type choisi par le panel d'experts et un questionnaire utilisé pour l'évaluation externe de ces rapports.

En comparaison avec les rapports de Roberts et Gierl (2010) et Jang (2005), conçus dans la même perspective diagnostique que notre étude, nos rapports présentent des différences et des ressemblances. Ils se distinguent notamment par une amélioration sur l'aspect visuel. Au niveau du contenu, nos rapports contiennent les cinq types d'informations proposées par Roberts et Gierl (2010), à savoir : a) les objectifs du rapport; (b) une description des habiletés évaluées; (c) la performance pour chaque habileté; (d) un résumé de la maitrise des habiletés; (e) un guide d'interprétation des résultats. Dans les deux cas, sans compter le guide d'interprétation des résultats, les informations ont été organisées en trois parties. Dans le rapport proposé par Roberts et Gierl (2010), la première partie contient des informations sur les scores totaux obtenus ainsi que les directives pour lire le rapport. Selon eux, ces informations normatives peuvent être accompagnées des scores totaux permettant aux utilisateurs d'obtenir des informations plus claires de leurs résultats. Cette idée est un peu différente dans notre cas, car dans cette première partie, nous n'avons mis que des informations personnelles de l'élève (nom, prénom, école, groupe et âge) qu'il convient préciser au regard du caractère à grande échelle du PIRLS 2011. Cette idée permet de renforcer le critère de l'accessibilité des rapports en mettant plus d'accent sur leur personnalisation. Les directives pour lire les rapports ont été présentées d'une manière très brève pour éviter la

surcharge d'informations et pour laisser plus d'espace à la présentation des résultats dans la première page. Quant à la recherche de Jang (2005), les directives pour lire des rapports ont été jumelées avec le mode d'emploi pour l'interprétation des graphiques dans la deuxième partie. À notre avis, le fait de présenter les directives dans la première partie constitue une amélioration de notre recherche, ce qui permet de mieux orienter des enseignants dès le début lorsqu'ils reçoivent des rapports diagnostiques. Nous avons rédigé ces directives dans un langage simple et compréhensible, pour ne pas nuire à l'intérêt immédiat des enseignants lors de l'interaction avec ces rapports.

La deuxième partie du rapport porte sur la présentation des résultats. Nous avons utilisé trois types de graphiques différents pour présenter le profil de l'élève avec les trois couleurs rouge, jaune et vert, pour distinguer les différents niveaux de maitrise des habiletés. L'utilisation de différentes couleurs selon les niveaux de maitrise est également un des points forts de nos rapports, car elle corrobore très bien les principes de présentation proposés par Cleveland et McGill (1985) et Koslyn (1994). En effet, d'après eux, il est important: (a) d'utiliser le contraste pour exprimer les informations importantes; (b) d'utiliser les redondances des indices visuels pour accentuer les informations présentées (couleurs différentes); (c) d'utiliser la proximité pour regrouper les éléments similaires ensemble; (d) d'utiliser un alignement commun pour marquer la structure visuelle des informations.

Nous avons décidé de choisir de présenter un graphique au lieu d'un tableau pour présenter les résultats, étant donné que cela permet de mieux communiquer les changements et les tendances des données comme l'ont affirmé Shah, Mayer et Hegarty (1999). De plus, l'utilisation d'un tableau est recommandée lorsqu'on présente une petite quantité de données et qu'il faut fournir des chiffres exacts pour les comparaisons (Tufte, 1996), ce qui est un peu inapproprié dans notre cas. À la différence des recherches de Robert et Gierl (2010) et Jang (2005) qui présentent les informations spécifiques sur la maitrise de chaque attribut accompagnées des réponses des candidats aux items, des réponses correctes ainsi qu'un sommaire des réponses, nous avons décidé de ne pas présenter ces informations dans nos rapports, étant donné qu'elles sont utiles seulement si les enseignants peuvent avoir accès à une copie de l'épreuve, ce qui n'est pas le cas pour le PIRLS 2011. De plus, le fait de ne pas présenter ces informations dans cette partie nous

laisse plus d'espace pour le résumé de la performance de l'élève ainsi que des pistes d'intervention, ce qui est aussi un avantage de nos rapports.

La dernière partie de nos rapports vise à présenter le résumé de la performance de l'élève et à proposer des pistes d'intervention pour améliorer ses points faibles. Dans nos rapports, cette partie a été rédigée avec soin dans un langage facile à comprendre et en tenant compte des principes de présentation des éléments textuels suggérés par Robert et Gierl (2010). De plus, les pistes d'intervention ont été proposées par les membres du panel d'experts en se basant sur les différents documents ministériels tels que la progression des apprentissages, le référentiel d'intervention en lecture pour les élèves de 10 à 15 ans. Ceci constitue une balise de référence importante pour assurer la cohérence entre les pistes d'intervention proposées pour le profil type et les progressions de l'apprentissage en lecture ainsi que la compréhension des rapports des enseignants, ce qui fait la différence entre nos rapports et ceux élaborés par Robert et Gierl (2010) et Jang (2005). Dans le cas de Robert et Gierl (2010), la proposition des pistes d'intervention reste encore très sommaire puisqu'ils ont seulement fait un résumé des points forts et des faiblesses de l'élève sans entrer en détail sur ce qu'il faudrait faire pour les améliorer. Ces éléments sont en outre absents dans les rapports de Jang (2005). Cependant, en plus des rapports individuels destinés à chaque élève, Jang (2005) a également produit un rapport destiné à l'enseignant qui résume les performances de tous les élèves du groupe, ceci peut être une idée intéressante à considérer lorsque nous voulons produire les rapports en grand nombre.

En ce qui concerne le guide d'interprétation, nous avons intégré des éléments proposés par Robert et Gierl (2010) portant sur une description très détaillée des habiletés évaluées et comment ces éléments sont reliés au test. Nous avons également intégré des informations qui permettent de savoir comme les scores diagnostiques ont été obtenus et quelles sont les directives pour interpréter les résultats du rapport. Cependant, nous avons décidé de ne pas mettre trop d'informations sur les modélisations diagnostiques pour ne pas nuire la compréhension des enseignants, ce qui est un peu différent du format de rapport proposé de Roberts et Gierl (2010). Le guide d'interprétation proposé par Jang (2005) ne contient que les définitions plus détaillées sur les neuf habiletés mesurées, car les directives pour l'interprétation des résultats ont été présentées dans la deuxième partie

du rapport en parallèle avec le graphique. Dans notre cas, nous avons ajouté une description plus détaillée sur les objectifs du rapport ainsi que le mode d'emploi pour lire les rapports avec des exemples plus concrets de l'interprétation des résultats. À notre avis, ces exemples avec des figures en différentes couleurs ou des pourcentages rendent le guide d'interprétation plus utile aux enseignants et assurent un équilibre entre les éléments textuels et non textuels. Ceci constitue un autre point fort de nos rapports en comparaison à ceux de Jang (2005) et de Roberts et Gierl (2010).

La dernière étape du cadre de référence de l'élaboration des rapports proposé par Zapata-Rivera, VanWinkle et Zwick (2012) et de Hambleton et Zenisky (2010) consiste à réaliser des évaluations externes de ces rapports auprès du public visé. Nous l'avons fait à l'aide d'un questionnaire administré en ligne auprès de 98 participants enseignants au primaire, conseillers pédagogiques et orthopédagogues. Nous discutons des résultats dans les lignes qui suivent.

5.1.4. Discussion sur les perceptions des enseignants au primaire, des conseillers pédagogiques et des orthopédagogues sur les formats de rapports évalués

À titre de rappel, l'instrument utilisé pour l'évaluation de trois formats de rapports a été conçu en adaptant les questionnaires de Zapata-Rivera, Vezzu et VanWinkle (2013) et de Vezzu, VanWinkle et Zapata (2012). Notre questionnaire comprend 55 questions portant sur quatre dimensions : la préférence, l'évaluation de la qualité, la compréhension des rapports et des informations sociodémographiques des participants. La partie de préférence contient quatre questions à trois choix de réponses sur le format de présentation préféré tandis que l'évaluation de la qualité des rapports se compose de trente-deux questions de type Likert à quatre choix de réponses sur les différentes parties du rapport. La compréhension des rapports contient neuf questions à choix multiples portant sur certains éléments spécifiques du profil de l'élève.

Une valeur ajoutée de notre étude est que nous avons réalisé des analyses factorielles pour définir la structure de 32 questions sur l'évaluation de la qualité des rapports. Ce genre d'analyses factorielles n'a pas été faite dans les recherches précédentes. Les résultats obtenus de ces analyses factorielles exploratoires suggèrent un regroupement de 30 items (avec la suppression des items 1.4 et 3.5) pour l'évaluation de

la qualité globale des rapports. Six sous-regroupements pour l'évaluation de différentes parties du rapport ont ensuite été suggérés, à savoir : évaluation des directives (3 items); évaluation de la présentation visuelle des graphiques (5 items); évaluation du contenu des graphiques (4 items); évaluation de la description du profil et des pistes d'intervention (6 items); évaluation du guide d'interprétation (8 items) ; évaluation de l'ensemble du rapport (4 items). Ces propositions de regroupement des items correspondent parfaitement à nos attentes, puisqu'ils ne sont pas très différents de la structure initiale du questionnaire. En effet, le fait de supprimer deux items parmi 32 items ne change pas beaucoup la qualité du questionnaire ni la quantité des informations obtenues. De plus, les coefficients d'Alpha de Cronbach pour le regroupement d'un facteur ainsi que des six sous-regroupements qui dépassent le seuil minimal requis (0,7) assurent bien la fidélité de notre instrument de mesure. Les lignes suivantes sont consacrées à la discussion des résultats obtenus en ce qui concerne la préférence, l'évaluation de la qualité et la compréhension des participants de trois formats de rapports élaborés en lien avec certaines variables sociodémographiques.

En ce qui concerne la préférence, les quatre questions permettent aux participants de choisir le format de rapport qu'ils préfèrent (question 1, le format qui permet de comprendre le mieux les niveaux de maitrise des habiletés de l'élève (question 2), le format qui permet de comprendre le positionnement de l'élève par rapport au groupe (question 3) et celui que les participants pensent que les enseignants vont préférer (question 4). Les résultats descriptifs révèlent que le format B est le plus choisi par les participants dans les quatre questions avec au moins 50 % des participants pour les questions 1, 2 et 4 et par 46,9 % pour la question 3. Vient ensuite le format C et finalement le format A. La préférence la plus grande pour le format B peut être expliquée par l'aspect familier du type de diagramme présenté. En effet, nous avons choisi un diagramme en barres juxtaposées (vertical) pour présenter le profil de l'élève. Ce type de diagramme est très courant et les participants peuvent le trouver souvent dans la vie quotidienne. Ce lien entre l'aspect familier du graphique utilisé dans les rapports et la préférence a été souligné dans la recherche de Zapata-Rivera, Vezzu et VanWinkle (2013). Une autre raison qui explique la préférence plus grande pour les formats B et C est que le profil de l'élève et celui du groupe sont présentés sur le même graphique au

lieu d'être présentés séparément dans deux graphiques comme dans le format A. Cette idée a été appuyée par le fait que les enseignants ont certaines hésitations dans l'interprétation des graphiques (Zapata-Rivera, Vezzu et VanWinkle, 2013). Ces hésitations feraient qu'ils ont tendance à choisir un format de rapport avec un type de graphique plus simple (format B) ou qui contient moins de graphiques (B et C).

Par ailleurs, bien que le nombre de graphiques soit le même dans les formats B et C, l'utilisation de deux axes différents pour les niveaux de maitrise des habiletés et les pourcentages de maitrise de chaque groupe dans le format C peuvent engendrer la confusion chez les participants, ce qui pourrait expliquer que ce format est moins choisi que le format B. Le format A contient deux graphiques qui ne portent pas nécessairement les mêmes types d'information, car l'un présente le profil de maitrise des habiletés de l'élève tandis que l'autre fournit les pourcentages des groupes d'élèves selon leur niveau de maitrise des habiletés. Afin de positionner l'élève par rapport au groupe, les participants doivent regarder les deux graphiques et situer l'élève par rapport à la couleur du groupe auquel il appartient. Le fait de présenter les résultats de l'élève et celui du groupe dans deux graphiques séparés rend plus difficile le positionnement de l'élève par rapport au groupe, ce qui explique le pourcentage le plus faible dans la question 3 pour le format A (14,3 %).

En ce qui concerne le lien entre la préférence et des variables contextuelles, les résultats du test chi 2 relèvent que la préférence des participants quant aux trois formats de rapports ne varie pas selon les postes occupés, l'ancienneté, le diplôme obtenu et le suivi ou non des cours en méthodes quantitatives. Les participants ont la plus grande préférence pour le format B, vient ensuite le format C, puis le format A. Ces résultats corroborent en grande partie ceux de la recherche de Zapata-Rivera, Vezzu et VanWinkle (2013). Les résultats de leur recherche reflètent qu'il n'y a pas de lien entre la préférence des rapports et l'ancienneté, le nombre de cours en statistiques suivi, le degré d'aisance avec les ordinateurs, le nombre de sessions de formation professionnelle en mesure, etc. Cependant, leur recherche montre que la préférence des participants varie selon le degré de confort avec les termes statistiques. Plus précisément, les participants qui sont plus familiers avec ces concepts préfèrent les formats de présentation plus abstraits comme les courbes tandis que ceux dont le degré de confort des termes statistiques est plus faible ont

tendance à choisir les types de graphiques concrets comme le diagramme avec des icônes empilées.

Par contre, notre recherche montre que la préférence des participants varie selon leur domaine de formation pour les questions 1 et 4. Les participants ayant suivi une formation en enseignement au primaire ou au préscolaire et en adaptation scolaire ont une préférence marquée pour le format B, suivi du format C et enfin du format A, tandis que les participants avec une formation en psychopédagogie et en d'autres domaines ont choisi d'abord le format A, puis le format B et enfin le format C. Ces résultats pourraient être expliqués par le fait que la question 1 porte sur la préférence personnelle sur les formats de rapports tandis que la question 4 porte sur le format de rapport qu'il pense que les autres enseignants préfèreront. Donc, dans la plupart des cas, les participants ont tendance à choisir le même format de rapport dans ces deux questions. C'est aussi une manière de renforcer leur croyance sur leur propre préférence, c'est-à-dire si le participant préfère le format A, il croit que d'autres enseignants auront le même choix. Parmi les explications possibles, une d'entre elles pourrait être liée aux caractéristiques du domaine de formation : les participants ayant une formation en psychopédagogie pourraient s'intéresser plus au processus d'apprentissage de chaque individu, ce qui fait qu'ils préfèrent que le profil de l'élève soit séparé de celui du groupe afin de mieux l'interpréter. Toutefois, jusqu'à présent, nous n'avons trouvé aucune recherche empirique qui appuie cette idée. Il serait donc souhaitable d'ajouter une question ouverte à chaque question de préférence afin de mieux comprendre la justification des participants comme cela a été fait dans la recherche de Zapata-Rivera, Vezzu et VanWinkle (2013).

En général, l'évaluation de la qualité des rapports varie selon le format de rapport, autrement dit, les résultats descriptifs suggèrent que l'évaluation de la qualité est davantage positive pour le format B, puis pour le format C et enfin pour le format A. Cependant, les tests d'Anova à un facteur suggèrent que les différences ne sont pas statistiquement significatives selon le format de rapport évalué. En ce qui concerne les postes occupés, les conseillers pédagogiques sont plus positifs quant à l'évaluation de la qualité des rapports, suivis des orthopédagogues et des enseignants au primaire. Toutefois, les résultats des tests d'Anova à un facteur montrent que les postes occupés par les participants n'influencent pas leurs perceptions sur la qualité des rapports.

Malgré le décalage dans le nombre des participants en provenance du Québec et ceux hors Québec, les résultats descriptifs suggèrent des tendances différentes dans les perceptions des participants quant à la qualité des rapports. Même s'ils sont peu nombreux, les participants non québécois sont plus positifs en ce qui concerne l'évaluation de la qualité des rapports, notamment dans l'évaluation de la description du profil et des pistes d'intervention, l'évaluation du guide d'interprétation et l'évaluation de l'ensemble du rapport. Cela peut s'expliquer par le caractère uniforme du programme de formation en lecture dans les différentes provinces au Canada, ce qui fait que même si nous nous basons sur les documents du CMEC et du MELS pour la reformulation des habiletés et pour la proposition des pistes d'intervention, les perceptions des participants non québécois ne sont pas moins positives que leurs collègues québécois. Ces tendances semblent inverses dans l'évaluation des directives, l'évaluation de la présentation et du contenu des graphiques, ce qui signifie que les participants québécois sont plus positifs que leurs collègues non québécois quant à l'évaluation de ces aspects. La raison qui l'explique est que nous avons travaillé avec un panel d'experts provenant du Québec, qui connaissent mieux la réalité du terrain. Ainsi, leurs propositions sur les directives, sur le choix du type de graphiques ainsi que sur le contenu à y mettre correspondent mieux aux attentes des participants québécois.

Malgré ces tendances observées dans l'échantillon, les résultats du test T suggèrent que les perceptions sur la qualité des rapports ne varient pas selon les régions de provenance, peut-être en raison du petit nombre de participants hors Québec. Nous avons fait le même exercice avec les variables l'ancienneté, le domaine de formation, le diplôme obtenu et le fait d'avoir suivi ou non des cours en méthodes quantitatives. Bien qu'il existe des tendances différentes dans l'évaluation de la qualité des rapports dans les sous-groupes, les résultats des tests inférentiels nous indiquent que ces différences ne sont pas statistiquement significatives. Cela veut dire que les perceptions des participants sur la qualité des rapports sont constantes peu importent les informations contextuelles.

Quant au lien entre l'évaluation de la qualité des rapports et la préférence des rapports, les résultats descriptifs suggèrent que le groupe de ceux qui évaluent le format de leur préférence a une évaluation plus positive de tous les aspects du rapport. Les résultats du test T confirment que les différences sont statistiquement significatives en ce

qui concerne l'évaluation de la qualité des rapports en général, l'évaluation des directives, l'évaluation de la présentation et du contenu des graphiques dans toutes les quatre questions de la préférence. Ces résultats peuvent s'expliquer par les différences dans la présentation visuelle ainsi que le contenu des graphiques de trois formats, ce qui fait que si le participant doit évaluer un type de graphique qu'il préfère et avec lequel il est familier, il est évident que ses perceptions seront plus positives. Par contre, les différences statistiquement significatives ne se trouvent pas dans l'évaluation de la description du profil et des pistes d'interprétation (Q1, Q2 et Q4), l'évaluation du guide d'interprétation (Q1, Q2, Q3 et Q4) et l'évaluation de l'ensemble du rapport (Q1 et Q2). La raison qui l'explique est que ces parties contiennent des éléments qui sont plus ou moins semblables dans les trois formats de rapports. Cela expliquerait que les perceptions des participants ne varient pas vraiment, peu importe le fait qu'ils évaluent leur format préféré ou pas. Ces résultats seraient intéressants à considérer lors de l'administration du questionnaire. La dernière partie de cette section est consacrée à la discussion des résultats de l'évaluation de la compréhension des participants sur les trois formats de rapports.

L'évaluation de la compréhension des participants se réalise en deux étapes. Dans un premier temps, nous avons fait des analyses descriptives avec le score total de neuf questions codées dichotomiquement (réponses correctes et incorrectes). Ces résultats descriptifs suggèrent que la compréhension que les participants ont du format A est quelque peu meilleure que celles qu'ils ont des formats B et C, bien que le format B soit préféré. Ceci peut être lié à la présentation du profil de l'élève et du groupe dans deux graphiques séparés permettant aux participants de mieux répondre aux questions portant sur ces deux profils. Les conseillers pédagogiques ont une meilleure compréhension des rapports que les orthopédagogues et les enseignants au primaire, ce qui va dans le même sens que les résultats de l'évaluation de la qualité des rapports. Autrement dit, ce sont également des conseillers pédagogiques qui présentent des perceptions plus positives, tandis que les enseignants au primaire sont, de tous, les moins positifs. Une des explications est que les conseillers pédagogiques pourraient être davantage familiers avec des présentations graphiques compte tenu de leur expérience avec les bilans d'études ou les recensions statistiques faites par les Commissions scolaires. Il serait donc intéressant

d'ajouter cette question sur le degré de familiarité des présentations graphiques ou la fréquence d'interactions avec rapports pour mieux comprendre ce lien.

Ces mêmes explications peuvent se trouver dans les groupes de différentes tranches d'âge et de l'ancienneté. Les participants âgés de 40 ans et plus et ceux qui sont plus expérimentés ont une meilleure compréhension que leurs collègues plus jeunes et moins expérimentés. Ces résultats sont logiques, car les participants plus expérimentés et plus âgés sont bien souvent ceux qui travaillent comme conseillers pédagogiques. Les répondants en provenance du Québec ont une meilleure compréhension des formats de rapports que ceux qui sont hors Québec. Ces résultats corroborent en grande partie des résultats obtenus sur l'évaluation de la qualité des rapports, surtout dans les parties où nous avons recours aux documents du MELS ou du CMEC pour les reformulations. Finalement, le fait d'avoir suivi des cours en méthodes quantitatives ou non au cours de cinq dernières années n'influence pas leur compréhension des rapports. Par contre, le groupe qui n'a pas suivi ce genre de cours manifeste une meilleure compréhension que leurs collègues qui l'ont suivi. Probablement, les cours de méthodes quantitatives portent sur des éléments plus sophistiqués en statistiques qui ne sont pas nécessairement utilisés dans l'interprétation de ces rapports. Toutefois, ce ne sont que des tendances observées dans notre échantillon avec les résultats descriptifs.

Les résultats de notre étude corroborent en partie ceux des recherches de Zwick, Zapata-Rivera et Hegarty (2014), de Zapata-Rivera et VanWinkle (2010) et de Zapata-Rivera, VanWinkle et Zwick (2012). Plus précisément, leurs recherches confirment que la compréhension des rapports ne varie pas selon les variables sociodémographiques telles que la familiarité avec les statistiques et le nombre d'années d'expérience des enseignants ainsi que la préférence des formats de rapports. Les études de Zapata-Rivera et VanWinkle (2010) et de Zapata-Rivera, VanWinkle et Zwick (2012) confirment également que la compréhension des rapports ne varie pas selon la présence ou l'absence du tutoriel (guide d'interprétation). Ceci nous amène à penser qu'il serait important pour nous de bien vérifier l'utilité de nos guides d'interprétation.

À la différence de notre recherche, celle de Zapata-Rivera, Vezzu et VanWinkle (2013) montre que les différences statistiquement significatives se trouvent entre la compréhension des rapports et les trois formats de rapports développés. Cependant, les

trois formats de présentations de graphiques sont tellement différents les uns des autres (courbe, boite à moustache et icônes empilés), qu'ils peuvent engendrer de différents niveaux de compréhension de ces formats. Dans notre cas, trois formats de présentation représentent certaines similitudes (même type de diagramme), ce qui pourrait expliquer que la compréhension des participants n'est pas grandement différente dans les trois formats évalués.

Dans un deuxième temps, afin de vérifier le lien entre la compréhension des rapports des participants et leurs perceptions sur la qualité de ces rapports, nous avons réalisé des tests non paramétriques de Mann-Whitney pour chacune des questions de la compréhension et l'évaluation de la qualité. L'intérêt est de savoir si l'évaluation de la qualité des rapports des participants est plus positive lorsqu'ils comprennent mieux les rapports ou pas. Nous n'avons pas utilisé la variable composite des neuf questions étant donné que les résultats des analyses factorielles ne nous suggèrent pas le regroupement approprié de ce facteur. Les résultats obtenus reflètent qu'il n'y a pas de lien entre l'évaluation de la qualité des rapports et la compréhension des questions 1, 2, 3, 4, 5, 8 et 9. Cependant, pour la question 6, l'évaluation des rapports varie dans les deux groupes en ce qui concerne les directives, la présentation des graphiques, la description du profil et des pistes d'intervention. Quant à la question 7, les résultats sont statistiquement significatifs pour l'évaluation de la qualité globale des rapports, l'évaluation des directives, l'évaluation de la description du profil et des pistes d'intervention et l'évaluation du guide d'interprétation.

À notre avis, la reformulation ainsi que la dépendance des questions de la compréhension sont des raisons qui expliquent les résultats statistiquement significatifs dans les questions 6 et 7 en ce qui concerne le lien entre l'évaluation de la qualité des rapports en général, d'une part et l'évaluation de certains aspects en particulier et la compréhension de ces questions, d'autre part. Par exemple, la réponse à la question 7 *« si on considère que le seuil de réussite est 0,5, lequel de ces énoncés est vrai?* » avec la proposition de quatre choix de réponse dépend de celle de la question 2 *« l'élève a-t-il bien maitrisé toutes les habiletés attendues?* » dont la réponse est Oui ou Non. Autrement dit, le choix de la réponse à la question 2 suggère celle à la question 7. Par contre, la question 6 *« Dans le graphique, quel est le niveau de maitrise de l'élève pour l'habileté*

Interpréter?» porte sur une habileté qui n'a pas été donnée comme exemple dans le guide d'interprétation, ce qui pourrait amener les participants à avoir de la difficulté à comprendre la question et, par conséquent, à choisir la bonne réponse Ces différences de tendances observées dans les réponses des participants justifient les liens qui sont statistiquement significatifs entre la compréhension de ces questions et l'évaluation de la qualité des rapports. Nous revenons sur la qualité de ces questions dans les limites de notre recherche.

5.2. Limites et recommandations

Une première limite de cette recherche émane du fait que l'élaboration de la matrice Q a été réalisée seulement avec un panel d'experts comprenant trois membres. Nous n'avons pas eu le moyen de vérifier ces habiletés identifiées avec les protocoles verbaux auprès des élèves en 4e année. L'identification des habiletés basée principalement sur les propositions des membres du panel d'experts et sur le cadre de référence de l'élaboration du test PIRLS 2011 pourrait engendrer le problème de la sur-spécification des habiletés, étant donné que les processus cognitifs pourraient être différents de ce qui s'est passé dans les têtes des élèves lors du test. Dans ce cas, les techniques du raffinement de la matrice Q telle que l'utilisation du test de Wald ou la méthode empirique proposée par de la Torre (2008) seront donc recommandées pour détecter les problèmes de la sous-spécification ou de la sur-spécification des attributs. Ceci constituerait également une piste importante dans les recherches futures permettant d'améliorer l'ajustement des modèles aux données ainsi que les paramètres de pseudo-chance et d'étourderie et d'augmenter donc la qualité diagnostique des items du PIRLS 2011.

Une deuxième limite renvoie au fait que lors de la discussion sur la qualité diagnostique des items du PIRLS 2011, les comparaisons ont été faites seulement avec les recherches en mathématiques ou en lecture réalisées avec d'autres tests étant donné qu'il n'existe pas encore des recherches réalisées avec PIRLS dans une même perspective diagnostique. Ce même problème a été mentionné dans la discussion des résultats de l'élaboration et de l'évaluation des rapports à cause d'un manque de recherche empirique sur le sujet. Nous n'avons comparé nos résultats avec les deux formats de rapports diagnostiques proposés par Jang (2005) et Roberts et Gierl (2010).

La troisième limite de l'étude réside dans le fait que nous avons travaillé avec deux panels d'experts comprenant différents membres en provenance de milieux éducatifs variés dans les deux phases de la recherche. Ceci constitue un avantage, mais aussi une limite de notre étude. Dans la phase de l'élaboration de la matrice Q, le panel d'experts se compose des membres ayant des expériences en lecture des tests à grande échelle et de l'approche diagnostique cognitive tandis que le panel 2 comprend des personnes qui connaissent mieux la réalité de la salle de classe en ce qui concerne l'enseignement et l'apprentissage de la lecture au primaire ainsi que la communication des résultats. C'est une des raisons qui explique pourquoi nous devions adapter la formulation des habiletés dans un langage accessible au panel 2. Cette étape de reformulation des habiletés en lecture a suscité des discussions des membres du panel 2 sur la nécessité d'une telle habileté pour une telle question. En effet, les membres du panel 1 ont d'abord identifié les attributs en anglais et les ont traduits en français afin de faciliter le travail du panel 2 : ceci a pour conséquence d'exercer une influence sur la compréhension que les membres du panel 2 ont des habiletés qu'ils ont identifiées. Il serait donc important que certains experts soient en même temps membres dans les deux panels afin de pouvoir garder le langage uniforme des habiletés dans ces deux phases de la recherche. D'un autre côté, le fait de travailler avec les experts différents en provenance des milieux éducatifs variés et dans deux langues différentes constitue également une force de notre projet sur laquelle nous revenons dans la section 5.3 sur les apports scientifiques et pratique de notre étude.

L'étape d'élaboration et d'évaluation des rapports diagnostiques présente également quelques limites. *Premièrement*, nous avons élaboré trois formats différents de rapports à partir d'un seul profil-type, ce qui est, selon nous, considéré comme une étude de cas. Ainsi, nous avons produit des graphiques avec Excel, ce qui fait qu'au moment de l'exportation de ces graphiques sur Survey Monkey pour l'évaluation des rapports, nous devions régler certains paramètres de l'image pour qu'ils ne soient pas trop lourds, tout en gardant la qualité visuelle acceptable des graphiques. Cependant, la qualité de ces images peut changer selon la grandeur des écrans d'ordinateurs ou la vitesse du Wifi que les participants utilisent pour répondre au questionnaire. Ce facteur pourrait influencer les évaluations des participants sur la qualité des rapports. Afin d'éviter ce problème, il est donc recommandé de penser à une étape de transfert entre l'élaboration des rapports et

l'évaluation des rapports pour conserver la qualité inchangeable des graphiques en utilisant des logiciels plus professionnels tels que Adobe Indesign, Shiny ou R Studio, etc. Cette étape est indispensable, lorsque nous voulons produire ces rapports en grand nombre.

Deuxièmement, la pré-validation du questionnaire n'a été réalisée qu'avec deux conseillers pédagogiques et un expert en mesure et évaluation. Bien que les commentaires obtenus nous semblent avoir été fortement utiles pour améliorer le questionnaire avant de le faire passer en grand nombre, nous aurions pu élargir notre pré-validation du questionnaire à des enseignants au primaire et des orthopédagogues afin de recevoir des rétroactions plus variées sur les questions conçues.

Troisièmement, bien que les résultats des analyses factorielles suggèrent d'excellents indices de la fiabilité et de la fidélité des questions de l'évaluation de la qualité des rapports, ils montrent également des problèmes de l'indépendance ainsi que la reformulation de nos questions sur la compréhension. Certaines questions ont fait l'objet d'exemples fournis dans le guide d'interprétation, ce qui fait que les participants peuvent facilement y trouver les bonnes réponses. Il aurait donc été souhaitable que cette partie du questionnaire soit retravaillée afin d'obtenir de meilleurs indices de fiabilité et de fidélité permettant de mieux faire le regroupement des facteurs sur la compréhension des rapports.

Quatrièmement, nous expliquons la fluctuation du nombre de réponses données par le fait que le questionnaire est jugé trop long par certains participants et que la qualité des images n'est pas toujours au rendez-vous.

*Cinquièmemen*t, lors de la passation du questionnaire, la question sur l'influence de la version du rapport évalué sur la préférence des participants a émergé. Cette idée pourrait être intéressante dans le choix d'une stratégie différente pour attribuer le format de rapport à évaluer aux participants. Faudrait-il laisser les participants évaluer leur format préféré ou inverser le processus, c'est-à-dire les questions sur l'évaluation de la qualité des rapports devraient être mises en premier lieu, ensuite les questions sur la préférence. Cette inversion permettrait d'étudier si la version du rapport à évaluer pourrait exercer une influence sur la préférence des rapports. Autrement dit, si les participants connaissent en avance le format de rapport à évaluer, choisiront-ils ce format

comme préférence? Cette idée est liée à la relation entre les perceptions des participants sur l'évaluation de la qualité des rapports et le fait qu'ils évaluent le format de leur préférence.

Finalement, la répartition des participants est inéquitable dans différents groupes sociodémographiques tels que les postes occupés, les régions de provenance, le genre des participants, l'ancienneté, le domaine de formation et le suivi ou non des cours en méthodes quantitatives. Cette répartition inéquitable des sous-groupes explique également que les postulats pour les tests de Chi 2, les tests T ainsi que ceux d'Anova à un facteur n'ont pas été parfaitement respectés dans certains cas. De plus, nous ne pouvons pas aller plus loin dans les analyses inférentielles pour la compréhension des rapports. Il aurait été utile d'améliorer nos stratégies de recrutement des participants en élargissant l'annonce dans différents milieux éducatifs telles que des commissions scolaires, des regroupements professionnels de conseillers pédagogiques et d'orthopédagogues dans d'autres provinces au Canada afin d'obtenir plus de participants avec une répartition plus équitable selon ces variables sociodémographiques. Malgré ces limites, notre recherche témoigne de plusieurs apports tant scientifiques que pratiques dont nous discuterons dans la partie suivante.

5.3. Apports scientifiques et pratiques et conclusion

L'objectif primordial de l'approche diagnostique cognitive (ADC) est de fournir des profils fins et détaillés sur les forces et les faiblesses cognitives des élèves à travers des rapports diagnostiques compréhensibles et interprétables. Cependant, malgré la grande demande d'informations diagnostiques de la part des enseignants et des apprenants, il y a peu d'épreuves conçues à visée diagnostique (Alderson, 2010; Jang, 2009; Lee et Sawaki, 2009). C'est pourquoi, les recherches empiriques sur cette approche ont été réalisées avec des tests à grande échelle tels que le TOEFL, le TOEFL iBT, le TOEIC, etc. Parmi les épreuves à grande échelle qui existe, le test PIRLS, qui a été élaboré avec un cadre de référence très poussé avec deux objectifs et quatre processus en lecture, présente donc un grand potentiel de diagnostic sur les forces et les faiblesses cognitives des élèves autres que les scores globaux.

Par ailleurs, une plus grande attention a été accordée à la conception des épreuves à grande échelle qui répondent aux attentes du public visé qu'à l'organisation et qu'à la communication des résultats (Goodman et Hambleton, 2004). La preuve est que les recherches sur l'élaboration et l'évaluation des rapports montrent que les enseignants rencontrent des difficultés dans l'interprétation des résultats à cause des jargons statistiques (Hambleton et al., 2002; Jaeger, 1998; Vezzu et al., 2012). De plus, ce processus ne souligne pas le rôle actif des enseignants lors de l'élaboration et de l'évaluation des rapports (Vezzu et al., 2012).

En partant de ces constats, notre recherche vise donc à modéliser des données de 4762 élèves canadiens ayant fait le PIRLS 2011 à visée diagnostique et cherche à élaborer et à évaluer des rapports diagnostiques issus des résultats des modélisations. Cette étude exploratoire apporte une contribution multiple à l'avancée des connaissances concernant les modélisations des données des tests à grande échelle avec les MCD, notamment avec le modèle G-DINA ainsi que l'élaboration et l'évaluation des rapports diagnostiques. Sur le plan scientifique, notre recherche est une des premières recherches qui suivent l'ensemble des cinq étapes de l'ADC proposée par Lee et Sawaki (2009), de l'élaboration de la matrice Q jusqu'à la communication et à la validation des rétroactions diagnostiques à travers des rapports. Il s'agit également de la première recherche qui adapte un cadre de référence de l'élaboration et l'évaluation des rapports (Zapata-Rivera, VanWinkle et Zwick, 2012) dans une approche diagnostique cognitive, ce qui n'existe pas encore dans les recherches précédentes menées dans une même perspective. Malgré la qualité diagnostique moyenne des items du PIRLS 2011, ce qui a été justifié par le fait que le test n'a pas été initialement conçu à visée diagnostique, les résultats obtenus des modélisations montrent donc la possibilité de recevoir des informations plus détaillées sur les forces et les faiblesses cognitives des élèves à travers l'enquête PIRLS et par l'intermédiaire de ces rapports diagnostiques. En effet, le fait que les habiletés sont maitrisées en gros autour de 50% est un argument en faveur du potentiel diagnostique du test.

Notre recherche repose sur une méthodologie très rigoureuse avec des étapes cohérentes et solides lors de la conception de la matrice Q ainsi que l'élaboration et l'évaluation des rapports avec le panel d'experts. Le lien entre les deux objectifs de

recherche a été assuré lorsque nous avons ajouté l'étape de la reformulation des habiletés en lecture dans un langage accessible aux enseignants en salle de classe en référant aux documents ministériels. Ce lien est également assuré avec le choix du profil-type le plus représentatif des élèves en salle de classe fait avec le panel d'expert 2.

Par ailleurs, notre étude a mis en avant l'aspect de la transversalité de l'étude dans les deux milieux américains et québécois, ce qui correspond tout à fait à l'échelle internationale du test PIRLS 2011, car il est rare, selon nous, d'avoir des experts américains et québécois dans une même étude. Le fait que nous avons travaillé avec les deux panels en provenance des milieux éducatifs variés et dans deux langues différentes permet de nous nous profiter de leur expertise et leurs riches expériences sur le terrain. Les résultats obtenus montrent que nous avons réussi à combiner adéquatement les deux visions, ce qui constitue un apport méthodologique très important de notre étude.

Au niveau pratique, nous sommes arrivée à concevoir trois formats de rapports qui sont fiables, interprétables, utilisables et accessibles avec un panel d'experts provenant de différents milieux éducatifs, ce qui a assuré la variété des commentaires obtenus lors du processus de l'élaboration et de l'évaluation des rapports. Ces rapports répondent bien aux différents critères d'un bon rapport suggérés dans la revue de la littérature. De plus, nous avons réussi à développer et valider un questionnaire sur l'évaluation des rapports diagnostiques, ce qui n'existe pas encore dans les recherches précédentes sur le sujet. Les résultats des analyses factorielles exploratoires montrent que la fiabilité et la fidélité de notre instrument sont très bien assurées. De plus, les résultats sur l'évaluation des rapports confirment qu'ils sont grandement appréciés par le public visé provenant de différents groupes sociodémographiques avec une préférence constante et une évaluation très positive ainsi qu'une très bonne compréhension des trois formats de rapports élaborés. Ces belles qualités de l'instrument que nous avons conçu et des trois formats de rapports élaborés suggèrent donc de l'excellent potentiel de leur applicabilité dans plusieurs milieux éducatifs différents. Les résultats obtenus du processus de l'évaluation des rapports nous ont permis de ressortir un format de rapport préféré des enseignants au primaire, des conseillers pédagogiques et des orthopédagogues (le format B). Par ailleurs, le fait que la préférence pour le format B résiste aux différences entre les divers groupes sociodémographiques nous semble un résultat vraiment intéressant, car cela nous fournit

une balise importante pour développer des rapports en grand nombre dans un contexte semblable. Plus précisément, en nous inspirant des démarches d'élaboration des rapports proposées et les résultats de cette étude, nous travaillons actuellement en collaboration avec la firme de testing privée de Brisson-Legris pour concevoir des rapports à partir des résultats d'un test diagnostique en mathématiques pour la formation professionnelle. À l'heure actuelle, nous avons réussi à générer les graphiques du format B de manière automatique à partir des résultats des modélisations G-DINA en utilisant le logiciel Python. La prochaine étape consistera à améliorer l'aspect visuel des rapports avant de les distribuer aux étudiants. Reste donc à savoir comment ces rapports seront appréciés auprès du public visé dans un contexte réel.

De nombreuses recherches dans la même perspective devront encore se pencher sur le raffinement de la matrice Q, permettant ainsi d'améliorer la qualité diagnostique des items du PIRLS 2011. Il faudra également trouver des logiciels plus professionnels pour la passation en ligne du questionnaire afin de conserver la qualité des images ainsi que des rapports. Des soucis concernant la qualité des questions de compréhension doivent être pris en compte afin d'assurer la fiabilité et la fidélité de ces questions. Finalement, les stratégies de recrutement doivent prendre en considération la répartition équitable des participants provenant de différents groupes sociodémographiques lors de l'évaluation des rapports.

Notre recherche contribue à établir le pont entre les résultats issus du test PIRLS 2011 et l'évaluation diagnostique cognitive en lecture grâce à des rapports diagnostiques compréhensibles et interprétables destinés aux enseignants visant ultimement à améliorer l'enseignement et l'apprentissage de la lecture au primaire.

BIBLIOGRAPHIE

Abrahamsen, E. P. & Shelton, K. C. (1989). Reading Comprehension in Adolescents with Learning Disabilities Semantic and Syntactic Effects. *Journal of Learning Disabilities*, *22*(9), 569-572.

Adams, M. & Bruce, B. (1982). Background knowledge and reading comprehension. *Reader meets author: Bridging the gap*, 2-25.

Adams, M.J. & Collins, A.M. (1979). A schema-theoretic view of reading . Dans Fredolle, R.O. (ed.) *Discourse Processing.' Multidisciplinary Perspectives.* Norwood, NJ: Ablex.

Alderson, J. C. (2005). *Assessing reading*. Stutgart, Denmark: Ernst Klett Sprachen.

Afflerbach, P. & Cho, B. (2009). Identifying and describing constructively responsive comprehension strategies in new and traditional forms of reading. Dans S.E. Israel, & G.G. Duffy (Eds.), *Handbook of research on reading comprehension* (pp. 69-90). New York: Routledge.

Alderson, J. C. (2010). "Cognitive diagnosis and q-matrices in language assessment": A commentary. *Language Assessment Quarterly*, *7*(2), 96-103.

Alderson, J. C. & Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a foreign language*, *5*(2), 253-270.

Alexander, P. A. & Jetton, T. L. (2000). Learning from text: A multidimensional and developmental perspective. *Handbook of reading research*, *3,* 285-310.

An, S. (2013). Schema Theory in Reading. *Theory and Practice in Language Studies*, *3*(1), 130-134.

Anderson, N. J., Bachman, L., Perkins, K. & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing, 8*(1), 41-66.

Anderson, R. C. & Pearson, P. D. (1984). A schema-theoretic view of basic processes in reading comprehension. Dans P. D. Pearson (Ed.), *Handbook of reading research* (Vol. 1, pp. 255-291). New York: Longman**.**

Anderson, R. C. & Pearson, P. D. (1988). A schema-theoretic view of basic processes in reading comprehension. *Interactive approaches to second language reading,* 37-55.

Anderson, R.C. & al. (1977). Frameworks for comprehending discourse. *American Educational Research Journal, 14*(4), 367-381

Aschbacher, P. R. & Herman, J. L. (1991). *Guidelines for effective score reporting*. National Center for Research on Evaluation, Standards, and Student Testing.

Baker, C. D. (1991). Literacy practices and social relations in classroom reading events. Dans C. Baker & A. Luke (Eds.), *Toward a critical sociology of reading pedagogy*. Philadelphia, PA: John Benjamins.

Barlett, F.C. (1932). *Remembering: A Study in Experimental and Social Psychology*. London: New Psychological Linguistics.

Barton, D. (1994). *Sustaining local literacies*. Clevedon: Multilingual Matters Education for Development.

Basokcu, T. O. (2014). Classification accuracy effects of Q-matrix validation and sample size in DINA and G-DINA models. *Journal of Education and Practice*, *5*(6), 220-230.

Beach, R. & Hynds, S. (1996). Research on response to literature. Dans R. Barr, M.L. Kamil, P. Mosenthal & P.D. Pearson (Eds.), *Handbook of reading research* (Vol. 2) (pp. 453-489). Mahwah, NJ: Lawrence Erlbaum Associates.

Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language*, 384-414.

Bloom, B. S., Hasting, J.T. & Madaus, G.F. (1971). *Handbook on formative and summative evaluation of student learning.* New York, États-Unis: Mcgraw-Hill.

Buck, G. (1990). *The testing of second language listening comprehension*. Thèse de doctorat inédite. Angleterre: Université de Lancaster.

Buck, G. & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: examining attributes of a free response listening test. *Language testing, 15*(2), 119-157.

Buck, G., Tatsuoka, K. & Kostin, I. (1997). The subskills of reading: rule-space analysis of a multiple-choice test of second language reading comprehension. *Language learning, 47*(3), 423-466.

Carrell, P. L. (1983). Some issues in studying the role of schemata, or background knowledge, in second language comprehension. *Reading in a foreign language*, *1*(2), 81-92.

Carrell, P. L., Devine, J. & Eskey, D. E. (Eds.). (1988). *Interactive approaches to second language reading*. Cambridge, NY: Cambridge University Press.

Carver, R. P. (1973). Reading as reasoning: Implications for measurement. *Assessment problems in reading.* Newark, DE: International Reading Association.

Carver, R. P. (1992). What do standardized tests of reading comprehension measure in terms of efficiency, accuracy, and rate?. *Reading Research Quarterly*, *27*(4), 347-359.

Chapelle, C., Grabe, W. & Berns, M. S. (1997). *Communicative language proficiency: Definition and implications for TOEFL 2000*. Princeton, NJ: Educational Testing Service.

Chen, J., Torre, J. & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*(2), 123-140.

Chiu, C., Seo, M. & Douglas, J. (2009). Cluster analysis for cognitive diagnosis: An application to the 2001 PIRLS reading assessment. *IERI Monograph Series: Issues and Methodologies in Large-scale Assessment, 2*, 137-159.

Clapham, C. (1996). *The development of IELTS: A study of the effect of background on reading comprehension* (Vol. 4). Cambridge, NY: Cambridge University Press.

Clay, M. M. (1991). *Becoming literate: The construction of inner control*. Portsmouth, NH: Heinemann

Cleveland, W. S. & McGill, R. (1985). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of American Statistical Association*, *79,* 531-554.

Conseil des Ministres de l'éducation du Canada (2011). *PPCE de 2010: Rapport de l'évaluation pancanadienne en mathématiques, en sciences et en lecture,* Toronto, Ontario: CMEC.

Cui, Y., Gierl, M. J. & Chang, H.-H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement, 49*(1), 19-38.

Davis, F. B. (1968). Research in comprehension in reading. *Reading Research Quarterly*, 499-545.

Davis, F. B. (1972). Psychometric research on comprehension in reading. *Reading Research Quarterly*, 628-678.

de la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement, 33*(3), 163-183.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179-

199.

de la Torre, J. & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *73*(4), 595-624.

Deschênes, A-J. (1986). *La compréhension, la production de textes et le développement de la pensée opératoire*. Thèse de doctorat inédite, Québec: Université de Laval.

DiBello, L. V., Roussos, L. A. & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. Dans C. V. Rao & S. Sinharay (Eds.), Handbook of statistics (Vol. 26, Psychometrics; pp. 979– 1027). Amsterdam, the Netherlands: Elsevier.

Drum, P. A., Calfee, R. C. & Cook, L. K. (1981). The effects of surface structure variables on performance in reading comprehension tests. *Reading Research Quarterly*, 486-514.

Dubin, F., Eskey, D. E., Grabe, W. & Savignon, S. (1986). *Teaching second language reading for academic purposes*. Reading, MA: Addison-Wesley.

Duke, N. (2004). The case for informational text. *Educational Leadership, 61*(6), 40-44.

Elley, W. B. (1992). *How in the World Do Students Read?; Warwick B. Elley: IEA Study of Reading Literacy*. International Association for the Evaluation of Educational Achievement.

Elley, W. B. (1994). *The IEA study of reading literacy: Achievement and instruction in thirty-two school systems.* Oxford, United Kingdom: Pergamon.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*(3), 380–396.

Fast, E. F. (2002). *A guide to effective accountability reporting: Designing public reports that effectively communicate accountability, assessment, and other quantitative education indicators in an easy understood format*. Washington, DC: Council of Chief State School Officers.

Field, A. (2009). Discovering Statistics Using SPSS. London, UK: Sage publication.

Figari. G. & Achouche. M., (2001). *L'activité évaluative réinterrogée-Regards scolaires et socioprofessionnels*. Bruxelles, Belgique: De Boeck.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76* (5), 378–382.

Flower, L. (1985). *Problem-solving strategies for writing (2nd ed.)*. New York, NY: Harcourt Brace Jovanovich.

Freedle, R. & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: Implications for construct validity. *Language Testing*, *10*(2), 134-169.

Fu, J. & Li, Y. (2007). *An integrated review of cognitively diagnostic psychometric models.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Galda, L. & Beach, R. (2001). Response to literature as a cultural activity. *Reading Research Quarterly, 36*(1), 64-73.

Gao, L. (2006). Toward a cognitive processing model of MELAB reading test item performance. *Spann Fellow Working Papers in Second or Foreign Language Assessment, 4*, 1-39.

Gao, L. & Rogers, W. T. (2010). Use of tree-based regression in the analyses of L2 reading test items. *Language Testing, 28*(2), 1-28.

Giasson, J. (1996). *La compréhension en lecture*. Bruxelles: De Boeck Supérieur.

Gierl, M. J. Cui, Y. & Hunka, S. (2008). Using connectionist models to evaluate examinees' response patterns to achievement tests. *Journal of Modern Applied Statistical Methods, 7*(1), 234-245.

Godman. K.S (1967). Reading: A psycholinguistic guessing game. *Journal of the reading specialist, 6*(1), 126-135.

Goldman, S. R. & Rakestraw, J. A. Jr. (2000). Structural aspects of constructing meaning from text. Dans M.L. Kamil, P. Mosenthal, P.D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. 3) (pp. 311-336). Mahwah, NJ: Lawrence Erlbaum Associates

Goodman, D. P. & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*(2), 145-220.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Gough, P. B. & Cosky, M. J. (1977). One second of reading again. *Cognitive theory*, *2,* 271-288.

Gough, P. B., Hoover, W. A. & Peterson, C. L. (1996). Some observations on a simple view of reading. Dans C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties* (pp. 1–13). Mahwah, NJ: Lawrence Erlbaum.

Grabe, W. (1991). Current developments in second language reading research. *TESOL*

quarterly, *25*(3)*,* 375-406.

Grabe, W. (2000). Reading research and its implications for reading assessment. *Fairness and validation in language assessment*, 226-262.

Grossmann, F. & Tauveron, C. (1999). Littératie, compréhension et interprétation des textes*., Repères*, *19,* 139-166.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*(4), 301-321.

Hambleton, R., Lissitz, R. & Schafer, W. (2002). How can we make NAEP and state test score reporting scales and reports more understandable. *Assessment in educational reform,* 192–205.

Hambleton, R. K. & Slater, S. (1997). Are NAEP executive summary reports understandable to policy makers and educators? (CSE Technical Report 430). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Hambleton, R. K. & Zenisky, A. (2010, Novembre). *Improvements to student score reporting: Steps and more use of suitable methodologies*. Invited presentation at the ETS Conference on Score Reporting, Princeton, NJ.

Hartz, S. M. (2002). *A bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Thèse de doctorat inédite. Champaign: University of Illinois at Urbana-Champaign.

Hattie, J. (2009). Visibly learning from reports: The validity of score reports. *Online Educational Research Journal*.

Heath, S. B. (1980). The functions and uses of literacy. *Journal of Communication*, *30*(1), 123-133.

Henson, R., Templin, J. & Douglas, J. (2007). Using efficient model based sum-scores for conducting skills diagnoses. *Journal of Educational Measurement, 44*(4), 361-376.

Hiebert, E. H. & Raphael, T. E. (1996). Perspectives from educational psychology in literacy and literacy learning and their extensions to school practice. *Handbook of educational psychology*, 550-602.

Horton, W. (1991). Overcoming chromophobia: A guide to the confident and appropriate use of color. *IEEE Transactions on Professional Communication*, *34,* 160-173.

Huang, H. Y. & Wang, W. C. (2014). The Random-Effect DINA Model. *Journal of Educational Measurement*, *51*(1), 75-97.

Hudson, T. (1982). The effects of induced schemata on the “short circuit” in L2 reading: Non-decoding factor in L2 reading performance. *Language learning*, *32*(1), 1-33.

Hudson, T. (1996). *Assessing second language academic reading from a communicative competence perspective: Relevance for TOEFL 2000.* Princeton, NJ: Educational Testing Service.

Huebner, A. & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement*, *71*(2), 407-419.

Huff, K. & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. Dans J.P. Leighton, M.J. Gierl (dir.), *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge, États-Unis: Presses universitaires de Cambridge.

Im, S. & Corter, J.E. (2011). Statistical consequences of attribute misspecification in the Rule-Space method. *Educationnal and psychological measurement*, *71* (4), 712-731

Impara, J. C., Divine, K. P., Bruce, F. A., Liverman, M. R. & Gay, A. (1991). Does interpretive test score information help teachers?. *Educational Measurement: Issues and Practice*, *10*(4), 16-18.

Irwin, J. W. (1986). *Teaching reading comprehension processes.* Englewood, New Jersey: Prentice-Hall.

Irwin, J. W. (1991). *Teaching reading comprehension processes.* Englewood, New Jersey: Prentice-Hall.

Jacobs, G. (1997). *Successful strategies for extensive reading*. Singapore: RELC.

Jaeger, R. M. (1998). *Reporting the results of the National Assessment of Educational Progress*. Washington, DC: American Institutes for Research.

Jang, E. E. (2010). Demystifying a Q-matrix for making diagnostic inferences about l2 reading skills: The author responds. *Language Assessment Quarterly, 6*(3), 210-238.

Jang, E.E. (2005). *A validity narrative: effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL*. Thèse de doctorat inédite. Champaign, IL: University d'Illinois.

Jang, E.E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability : Validity argument for fusion model application to LanguEdge assessment. *Language testing, 26*(1), 31-73.

Junker, B. W. & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and

connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258-272.

Kasai, M. (1997). *Application of the rule space model to the reading comprehension section of the test of English as a foreign language (TOEFL)*. Thèse de doctorat inédite, Champaign, IL: Université d'Illinois à Urbana-Champaign.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. *Psychological review*, *95*(2), 163-182.

Kintsch, W., & Kintsch, E. (2005). Comprehension.Dans S. Paris, & S. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 71-92). Mahwah, NJ: Lawrence Erlbaum Associates.

Kintsch, W., Welsch, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and language*, *29*(2), 133-159.

Kintsch, W., & Yarbrough, J. C. (1982). Role of rhetorical structure in text comprehension. *Journal of educational psychology*, *74*(6), 828-834.

Klesch, H. S. (2010). *Score reporting in teacher certification testing: A review, design, and interview/focus group study*. Thèse de doctorat inédite. Boston, MA: Université de Massachusetts Amherst.

Kline, P. (1979). *Psychometrics and psychology*. London: Acaderric Press.

Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing, 19* (2), 193-220.

Koda, K. (1994). Second language reading research: Problems and possibilities. *Applied psycholinguistics*, *15*(1), 1-28.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 154-186). New York: American Council on Education and Macmillan.

Kosslyn, S. M. (1994). Elements of graph design. New York, NY: Freeman. Kucer, S.B. (2005). *Dimensions of literacy: A conceptual base for teaching reading and writing in school settings* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Kunnan, A. J. & Jang, E. E. (2009). 32 diagnostic feedbacks in language assessment. *The handbook of language teaching*, 610-628.

Labrecque, M., Chuy, M., Brochu, P. & Houme, K. (2012). *PIRLS 2011, le contexte au Canada (Rapport numéro 35).* Toronto, Ontario: CMEC.

Landis, J. R. & Koch, G. G. (1977) The measurement of observer agreement for categorical

data. *Biometrics, 33,* 159-174.

Lafontaine, D. (2001). Quoi de neuf en littératie? Regard sur trente ans d'évaluation de la lecture. *Cahiers du Service de Pédagogie Expérimentale-Université de Liège*, *7,* 71-95.

Langer, J. A. (1990). The process of understanding: Reading for literary and informative purposes. *Research in the Teaching of English*, *24*(3), 229-260.

Langer, J. A. & Allington, R. L. (1992). Curriculum research in writing and reading. Dans P.Jackson (Ed.), *Handbook of research on curriculum* (pp 687-725). New York: Macmillan.

Laplante, L. (2011). L'évaluation diagnostique des difficultés d'apprentissage de la lecture. Dans M-J Berger et A. Desrochers (Dir.) *L'évaluation de la littératie.* Ottawa, Ontario : Presses de l'Université d'Ottawa.

Lee, Y-W et Sawaki, Y. (2011). Application of three cognitive diagnosis models to ESL reading and listening assessments, *Language Assessment Quarterly, 6*(3), 239-263.

Lee, Y.-W. & Sawaki, Y. (2009). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly, 6*(3), 172-189.

Legendre, R. (2005). *Dictionnaire actuel de l'éducation,* 3e édition. Montréal, Québec: Éditions Guérin.

Leighton, J. P. & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice*, *26*(2), 3-16.

Leong, C. K. (1984). Cognitive processing, language awareness, and reading in grade 2 and grade 4 children. *Contemporary Educational Psychology*, *9*(4), 369-383.

Leu, D.J., Kinzer, C.J., Coiro, J.L. & Cammack, D.W. (2004). Toward a theory of new literacies emerging from the internet and other information and communication technologies. Dans R.B. Ruddell and N.J. Unrau (Eds.), *Theoretical models and processes of reading* (5th ed.) (pp. 1570-1613). Newark, DE: International Reading Association.

Li, H. (2011). A cognitive diagnostic analysis of the MELAB reading test. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *9*, 17-46.

Li, H., & Suen, H. K. (2013). Constructing and Validating a Q-Matrix for Cognitive Diagnostic Analyses of a Reading Test. *Educational Assessment*, *18*(1), 1-25.

Loye, N. (2008). *Conditions d'élaboration de la matrice q des modèles cognitifs et impact*

sur sa validité et sa fidélité. Thèse de doctorat inédite, Ontario: Université d'Ottawa.

Loye, N. (2010). 2010, odyssée des modèles de classification diagnostique (MCD). *Mesure et évaluation en éducation*, *33*(3), 75-98.

Loye, N. (2011). Panorama des programmes d'enquêtes à large échelle. *Mesure et évaluation en éducation, 34*(2), 3-24.

Loye, N., & Barroso da Costa, C. (2013). Hiérarchiser les besoins de diagnostic en mathématique en FP à l'aide d'un modèle de Rasch. *Mesure et évaluation en éducation. 36* (2), 59-85.

Loye, N. & Lambert-Chan, J. (2016). Au coeur du développement d'une épreuve en mathématique dotée d'un potentiel diagnostique. *Mesure et évaluation en éducation. 39* (3), 29-57.

Lumley, T. (1993). The notion of subskills in reading comprehension tests: an EAP example. *Language testing, 10* (3), 211-234.

Ma, W., Iaconangelo, C. & de la Torre, J. (2016). Model Similarity, Model Selection, and Attribute Classification. *Applied Psychological Measurement, 40* (3), 1-18.

Ma, X. & Meng, Y. (2014). Towards personalized english learning diagnosis: Cognitive diagnostic modelling for EFL listening. *Asian journal of education and e-Learning, 2* (5), 336-348.

MacCallum, R.C., Widaman, K.F., Preacher, K.J. & Hong, S. (2001). Sample size in factor analysis: The role of model error, *Multivariate Behavioral Reserach, 36*(4), 611-637.

Mc Kenna, M. C. & Stahl, S.A. (2012). *Assessment for reading instruction.* New York, NY: The Guilford Press.

McClelland, J. L. & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological review*, *88*(5), 375.

McVee, M. B., Dunsmore, K. & Gavelek, J. R. (2005). Schema theory revisited. *Review of educational research*, *75*(4), 531-566.

Minsky, M. (1975). A framework for representing knowledge. Dans P. H. Winston (Ed.), *The psychology of computer vision* (pp. 211–277). New York: McGraw Hill.

Mislevy, R. J., Steinberg, L. S. & Almond, R. G. (2003). On the structure of educational assessment (with discussion). *Measurement: Interdisciplinary Research and Perspective*, *1*(1), 3-62.

Mullis, I. V., Martin, M. O., Kennedy, A. M., Trong, K. L. & Sainsbury, M. (2009). *PIRLS 2011 assessment framework*. Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement.

Nadeau. M-A., (1978). L'évaluation de l'apprentissage en milieu scolaire: un modèle d'évaluation continue. *Revue des sciences de l'éducation*, *4*(2), 205-221.

Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research, 64*(4), 575-603.

OCDE (2011). *Résultats du PISA 2009: Surmonter le milieu social : L'égalité des chances et l'équité du rendement de l'apprentissage (Volume II),* Paris, Éditions OCDE.

Otero, J. & Kintsch, W. (1992). Failures to detect contradictions in a text: What readers believe versus what they read. *Psychological Science*, *3*(4), 229-235.

Palincsar, A. & Duke, N. (2004). The role of text and text-reader interactions in young children's reading development and achievement. *The Elementary School Journal, 105*(2), 183-197.

Perfetti, C. A. (1988). Verbal efficiency in reading ability. Dans C. A. Daneman, Meredyth (Ed); Mackinnon, GE (Ed); Waller, T. Gary (Ed). *Reading research: Advances in theory and practice*. San Diego, CA: Academic Press.

Perfetti, C. & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific studies of reading, 18,* 22-37.

Pettersson, R. (2002). Information design: An introduction. Philadelphia, PA: John Benjamins Publishing.

Pierre, R. (1992). La compréhension de textes écrits face au rehaussement des standards de littératie. *Scientia paedagogica experimentalis*, *29*(1), 3-21.

Pressley, M. (2006, April). *What the future of reading research could be*. Paper presented at the meeting of the International Reading Association's Reading Research, Chicago, Illinois.

Pressley, M. & Gaskins, I. W. (2006). Metacognitively competent reading comprehension is constructively responsive reading: how can such reading be developed in students? *Metacognition and Learning, 1*(1), 99-113.

Purves, A. C. (1973). *Literature Education in Ten Countries: An Empirical Study. International Studies in Evaluation II*. Stockholm: Almqvist & Wiksell.

Roberts, M. R. & Gierl, M. J. (2010). Developing score reports for cognitive diagnostic

assessments. Educational Measurement: *Issues and Practice, 29*(3), 25-38.

Ravand, H., Barati, H. & Widhiarso, W. (2013). Exploring Diagnostic Capacity of a High Stakes Reading Comprehension Test: A Pedagogical Demonstration. *Iranian Journal of Language Testing*, *3*(1), 11-37.

Rayner, K., Fischer, M. H., & Pollatsek, A. (1998). Unspaced text interferes with both word identification and eye movement control. *Vision Research*, *38*(8), 1129-1144.

Roegiers, V. (2004). *L'école et l'évaluation : Des situations pour évaluer les compétences des élèves*. Bruxelles : De Boeck.

Roussos, L. A., Templin, J. L. & Henson, R. A. (2007). Skills diagnosis using IRT-based latent class models. *Journal of Educational Measurement, 44,* 293-311.

Rost, D. H. (1993). Assessing different components of reading comprehension: fact or fiction?. *Language Testing*, *10*(1), 79-92.

Ruddell, R.B. & Unrau, N.J. (Eds.). (2004). Theoretical models and processes of reading (5th ed.). Newark, DE: International Reading Association.

Rumelhart, D. E. (1976). *Understanding and summarizing brief stories*. University of California, San Diego: Center for Human Information Processing,.

Rumelhart, D. E. (1978). *Schemata: The building blocks of cognition*. University of California, San Diego: Center for Human Information Processing,

Rupp, A. A., Ferne, T. & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: a cognitive processing perspective. *Language testing, 23*(4), 441-474.

Ryan, J. M. (2006). Practices, issues, and trends in student test score reporting. Dans S.M. Downing & T.M. Haladyna (eds). *Handbook of test development (pp 677–710)*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ryan, J. M. (2003). *An analysis of item mapping and test reporting strategies*. Greensboro, NC: South Carolina Department of Education.

Samuels, S. J. & Kamil, M. L. (1984). 7 Models of the reading process. *Handbook of reading research*, *1*, 185.

Sawaki, Y., Kim, H.-J. & Gentile, C. (2009). Q-matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly, 6*(3), 190-209.

Scallon, G. (1988). *L'évaluation formative des apprentissages* (Vol. 2). Québec, Québec:

Presses universitaires de Laval.

Scallon, G. (2000). *L'évaluation formative* (Vol. 2). Saint-Laurent, Québec: Éditions du Renouveau Pédagogique.

Scott, H. S. (1998). *Cognitive diagnostic perspectives of a second language reading test*. Thèse de doctorat inédite. Champaign, IL: University d'Illinois à Urbana-Champaign.

Sfard, A. (1998). On two metaphors for learning and the dangers of choosing just one. *Educational researcher*, *27*(2), 4-13.

Shah, P., Mayer, R. E., & Hegarty, M. (1999). Graphs as aids to knowledge construction: Signaling techniques for guiding the process of graph construction. *Journal of Educational Psychology, 91* (4), 690-702.

Sinharay, S. (2004). Experiences with Markov Chain Monte Carlo convergence assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics, 29* (4), 461-488.

Sinharay, S., Puhan, G. & Haberman, S. J. (2010). Reporting diagnostic scores in educational testing : Temptations, pitfalls, and some solutions. *Multivariate Behavioral Research, 45*(3), 553-573.

Smith, F. (1971). *Understanding Reading*. New York: Holt, Rinehart and Winston.

Snow, C. (2002). *Reading for Understanding. Towards an R&D Program in Reading Comprehension* (No. MR-1465-DERI). RAND CORP SANTA MONICA CA.

Spearritt, D. (1972). Identification of subskills of reading comprehension by maximum likelihood factor analysis. *ETS Research Bulletin Series*, *1972*(1), i-24.

Spiro, R. J., Bruce, B. C. & Brewer, W. F. (Eds.). (1980). *Theoretical issues in reading comprehension: Perspectives from cognitive psychology, linguistics, artificial intelligence, and education*. New York NY: Routledge.

Spooner, A. L. R., Baddeley, A. D. & Gathercole, S. E. (2004). Can reading accuracy and comprehension be separated in the Neale Analysis of Reading Ability? *British Journal of Educational Psychology, 74* (2), 187-204.

St-Pierre, M. C., Dalpé, V., Lefebvre, P., Giroux, C., Dalpé, M. C. S. P. V. & Céline, P. L. G. (2011). *Difficultés de lecture et d'écriture*. Québec: Presses universitaires du Québec.

Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading research quarterly*, *16*(1),

32-71.

Stothard, S. E. & Hulme, C. (1992). Reading comprehension difficulties in children: The role of language comprehension and working memory skills. *Reading and Writing, 4* (3), 245-256.

Street, B.V. (2001). Literacy empowerment in developing societies. Dans L. Verhoeven, & C. Snow (Eds.), Literacy and motivation: Reading engagement in individuals and groups (pp. 71-94). Mahwah, NJ: Lawrence Erlbaum.

Svetina, D., Gorin, J. S. & Tatsuoka, K. K. (2011). Defining and comparing the reading comprehension construct: A cognitive-psychometric modeling approach. *International Journal of Testing, 11*(1), 1-23.

Templin, J. & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological methods, 11*(3), 287.

Templin, J. & al. (2010). Diagnostic measurement: Theory, methods, and applications. New York, NY: Presses de Guilford.

Tharp, R. G. & Gallimore, R. (1988). *Rousing minds to life: Teaching, learning, and schooling in social context*. Cambridge, NY: Cambridge University Press.

Thorndike, R. L. (1973). Reading as reasoning. *Reading Research Quarterly*, *9*(2), 135-147.

Tjoe, H. & de la Torre, J. (2013a). Designing cognitively-based proportionnal reasonning problems as an application of modern psychological measurement models. *Journal of Mathematics education, 6*(2), 17-26.

Tjoe, H. & de la Torre, J. (2014). The identification and validation process of proportionnal reasonning attributes: an application of a cognitive diagnosis modeling framework. *Mathematics educational research journal, 26*(2), 237-255

Tufte, E. R. (1983). *The visual display of quantitative information.* Cheshire, CT: Graphics Press.

Tufte, E. R. (1996). *Visual explanations.* Cheshire, CT: Graphics Press.

Underwood, J. S., Zapata-Rivera, D. & VanWinkle, W. (2007). *Growing pains: Teachers using and learning to use IDMS* (Research Memorandum 08–07). Princeton, NJ: Educational Testing Service.

van Dijk, T.A. & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York, NY: Academic Press.

von Davier, M. (2008). A general diagnostic model applied to language testing data.

Princeton, NJ: ETS.

von Davier, M. & Yamamoto, K. (2004, Octobre). *A class of models for cognitive diagnosis.* Paper presented at the 4th Spearman Conference, Philadelphia, PA.

Vezzu, M., VanWinkle, W. & Zapata-Rivera, D. (2012). *Designing and evaluating an interactive score report for students*. Princeton, NJ: ETS.

Wainer, H. (1997). *Visual revelations*. New York, NY: Copernicus Books.

Wainer, H. (2005). *Graphic discovery*. Princeton, NJ: Princeton University Press.

Wainer, H., Hambleton, R. K. & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement*, *36* (4), 301-335.

Walter, P. (1999). Defining literacy and its consequences in the developing world. *International Journal of Lifelong Education, 18* (1), 31-48.

Wang, C., & Gierl, M. J. (2011). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in critical reading. *Journal of Educational Measurement*, *48*(2), 165-187.

Weaver, C.A. & Kintsch, W. (1996). Expository text. Dans R. Barr, M.L. Kamil, P. Mosenthal, & P.D. Pearson (Eds.), *Handbook of reading research* (Vol. 2) (pp. 230-245). Mahwah, NJ: Lawrence Erlbaum Associates.

Weir, C., Huizhong, Y. & Yan, J. (2000). *An empirical investigation of the componentiality of L2 reading in English for academic purposes* (Vol. 12). Cambridge: Cambridge University Press.

Wharton, C. & Kintsch, W. (1991). An overview of construction-integration model: a theory of comprehension as a foundation for a new cognitive architecture. *ACM SIGART Bulletin*, *2*(4), 169-173.

Yang, X. & Embretson, S. E. (2007). Construct validity and cognitive diagnostic assessment. Dans J.P. Leighton, M.J. Gierl (dir.), *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge, États-Unis: Presses universitaires de Cambridge.

Zapata-Rivera, D. (2011, April). *Designing score reports that help teachers make instructional decisions*. Paper presented the meeting of the Educational Research Association, New Orleans, LA.

Zapata-Rivera, D., VanWinkle, W. & Zwick, R. (2012). Applying Score Design

Principles in the Design of Score Reports for CBAL™ Teachers. *Research Memorandum No. RM-12-20)*. Princeton: Educational Testing Service.

Zapata-Rivera, D., Vezzu, M. & VanWinkle, W. (2013). Exploring Teachers' Understanding of Graphical Representations of Group Performance. *Research Memorandum No. RM-12-20)*. Princeton: Educational Testing Service.

Zapata-Rivera, J. D. & Katz, I. R. (2014). Keeping your audience in mind: applying audience analysis to the design of interactive score reports. *Assessment in Education: Principles, Policy & Practice*, *21*(4), 442-463.

Zenisky, A. L., Hambleton, R. K. & Sireci, S. G. (2009). Getting the message out: An evaluation of NAEP score reporting practices with implications for disseminating test results. *Applied Measurement in Education, 22*(4), 359-375.

Zwick, R., Zapata-Rivera, D. & Hegarty, M. (2014). Comparing graphical and verbal representations of measurement error in test score reports. *Educational Assessment*, *19*(2), 116-138.

ANNEXE 1-RÉSULTATS DESCRIPTIFS

Tableau 1.1. Analyses descriptives des parties 1 et 2 du questionnaire

Partie	Items	Moyenne	Écart type	N
1.Des directives	1.1- Les informations fournies sur l'élève sont suffisantes pour personnaliser le rapport.	3,13	0,731	68
	1.2- Les directives pour lire le rapport sont définies avec clarté	3,04	0,721	68
	1.3- Ces directives sont utiles pour mon interprétation du rapport	3,00	0,712	68
	1.4- J'aimerais que les objectifs du rapport diagnostique soient ajoutés dans les directives du rapport	3,21	0,682	68
2.Présentation du profil	2.1-Les graphiques présentés sont faciles à comprendre.	2,97	0,880	68
	2.2- Dans les graphiques, l'utilisation de différentes couleurs selon les niveaux de maitrise des habiletés est pertinente pour ma compréhension	3,18	0,791	68
	2.3-La taille des graphiques facilite ma lecture.	3,32	0,742	68
	2.4-Les légendes sont claires pour mon interprétation des graphiques	2,90	0,831	68
	2.5-Les légendes sont utiles pour mon interprétation des graphiques	3,12	0,723	68
	2.6-Les chiffres présentés dans les barres facilitent le positionnement de l'élève par rapport au groupe	3,12	0,838	68
	2.7-Le positionnement de l'élève par rapport au groupe permet de mieux identifier les pistes d'intervention	2,96	0,854	68
	2.8-Les niveaux de maitrises des habiletés de l'élève sont faciles à ceux du groupe	2,81	0,815	68
	2.9-J'ai confiance aux résultats diagnostiques présentés	3,04	0,656	68

Tableau 1.2. Analyses descriptives des parties 3 et 4 du questionnaire

Partie	Items	Moyenne	Écart-type	N
3. Description du profil	3.1-La description du profil est utile pour mon interprétation du graphique.	3,31	0,605	68
	3.2-La description du profil correspond à ma compréhension des graphiques	3,29	0,575	68
	3.3-Les pistes d'intervention pour l'élève (individuel) sont utiles pour mon enseignement	3,32	0,584	68
	3.4-Les pistes d'intervention pour l'élève (individuel) sont utiles pour mon évaluation	3,06	0,667	68
	3.5-J'aimerais bien avoir des pistes d'intervention pour le groupe	3,15	0,778	68
	3.6-Dans l'ensemble, le vocabulaire utilisé est facile à comprendre	3,22	0,709	68
	3.7-Les paragraphes sont bien formulés	3,19	0,629	68
	4.1-Les objectifs du rapport diagnostique sont présentés avec clarté.	3,35	0,641	68
	4.2-La description des habiletés facilite la compréhension du rapport	3,32	0,657	68
4. Guide d'interprétation	4.3-Le mode d'emploi pour lire les graphiques est utile pour ma compréhension du rapport	3,13	0,689	68
	4.4-Les exemples utilisés dans cette partie facilitent mon interprétation des graphique	3,28	0,666	68
	4.5-La quantité des informations fournies est suffisante pour comprendre le profil de l'élève	3,12	0,659	68
	4.6-Dans l'ensemble, le vocabulaire utilisé est facile à comprendre	3,19	0,652	68
	4.7-Les paragraphes sont bien formulés	3,19	0,652	68
	4.8-Dans l'ensemble, le rapport présenté a un aspect visuel qui me plaît	2,91	0,728	68
	4.9-Ce rapport est plus détaillé que le rapport fait par un professionnel (par exemple, les orthopédagogues, les conseillers pédagogiques, etc.)	2,46	0,818	68
	4.10-J'aimerais avoir accès à ce format de rapport pour mes élèves	3,13	0,644	68
	4.11-J'ai confiance dans mon interprétation du profil de mon élève	3,25	0,608	68
	4.12-Le profil de l'élève présenté dans ce rapport diagnostique correspond à celui de certains de mes élève	2,97	0,712	68

Tableau 1.3. Statistiques descriptives de l'évaluation de la qualité des rapports

	N	Minimum	Maximum	Moyenne	Écart-type
Évaluation de la qualité des rapports	68	1,57	4,00	3,1098	0,43701
Évaluation des directives	76	1,33	4,00	3,0395	0,63003
Évaluation de la présentation des graphiques	73	1,20	4,00	3,0932	0,67006
Évaluation du contenu des graphiques	73	1,50	4,00	3,0000	0,63191
Évaluation de la description du profil et des pistes d'intervention	72	2,00	4,00	3,2384	0,45864
Évaluation du guide d'interprétation	68	1,00	4,00	3,2298	0,52795
Évaluation de l'ensemble du rapport	68	1,00	4,00	2,8676	0,52453

Tableau 1.4. Résultats descriptifs de l'évaluation des rapports selon le format évalué

Groupe		Évaluation de la qualité des rapports	Évaluation des directives	Évaluation de la présentation des graphiques	Évaluation du contenu des graphiques	Évaluation de la description du profil et des pistes d'intervention	Évaluation du guide d'interprétation	Évaluation de l'ensemble du rapport
Format A	Moyenne	3,0772	3,1111	3,1300	2,8125	3,2833	3,1776	2,8684
	N	19	21	20	20	20	19	19
	Écart type	,37615	0,55109	0,58138	0,65331	0,37502	0,47747	0,51619
	Minimum	2,27	2,00	2,20	2,00	2,67	2,00	2,00
	Maximum	4,00	4,00	4,00	4,00	4,00	4,00	4,00
Format B	Moyenne	**3,1333**	3,1733	3,0522	3,2391	3,2121	3,2500	2,8571
	N	21	25	23	23	22	21	21
	Écart type	0,57879	0,60949	0,78443	0,74421	0,50705	0,68122	0,63033
	Minimum	1,57	1,67	1,20	1,50	2,00	1,00	1,00
	Maximum	3,97	4,00	4,00	4,00	4,00	4,00	4,00
Format C	Moyenne	3,1143	2,8778	3,1000	2,9417	3,2278	3,2500	2,8750
	N	28	30	30	30	30	28	28
	Écart type	0,36089	0,68079	0,65126	0,46740	0,48440	0,43964	0,45896
	Minimum	2,47	1,33	2,00	2,00	2,33	2,25	2,00
	Maximum	3,93	4,00	4,00	4,00	4,00	4,00	4,00

Tableau 1.5. Résultats descriptifs de l'évaluation des rapports selon les postes occupés

Vous êtes:		Évaluation de la qualité des rapports	Évaluation des directives	Évaluation de la présentation des graphiques	Évaluation du contenu des graphiques	Évaluation de la description du profil et des pistes d'intervention	Évaluation du guide d'interprétation	Évaluation de l'ensemble du rapport
Enseignant au primaire	Moyenne	3,0790	3,0533	3,0208	2,9479	3,1944	3,2038	2,8370
	N	46	50	48	48	48	46	46
	Écart type	0,46382	0,63303	0,70287	0,62091	0,47057	0,56725	0,56573
	Minimum	1,57	1,67	1,20	1,50	2,00	1,00	1,00
	Maximum	4,00	4,00	4,00	4,00	4,00	4,00	4,00
Orthopédagogue	Moyenne	3,0867	2,8333	3,1500	2,9375	3,3194	3,2625	2,7250
	N	10	12	12	12	12	10	10
	Écart type	0,29740	0,67420	0,60378	0,56533	0,44072	0,38392	0,39878
	Minimum	2,77	1,33	2,40	2,25	2,67	2,88	2,25
	Maximum	3,60	3,33	4,00	4,00	4,00	4,00	3,50
Conseiller pédagogique	Moyenne	3,2472	3,1667	3,3077	3,2500	3,3333	3,3021	3,1042
	N	12	14	13	13	12	12	12
	Écart type	0,43099	0,58104	0,59226	0,71443	0,43809	0,50130	0,39107
	Minimum	2,80	2,00	2,40	2,00	3,00	2,75	2,50
	Maximum	3,97	4,00	4,00	4,00	4,00	4,00	4,00

Tableau 1.6. Résultats descriptifs de l'évaluation des rapports selon les régions de provenance

Statistiques de groupe					
	Régions	N	Moyenne	Écart type	Moyenne erreur standard
Évaluation de la qualité des rapports	Hors-Québec	6	3,1278	0,52974	0,21627
	Québec	62	3,1081	0,43212	0,05488
Évaluation des directives	Hors-Québec	6	2,9444	0,74287	0,30327
	Québec	70	3,0476	0,62499	0,07470
Évaluation de la présentation des graphiques	Hors-Québec	6	2,9000	0,65422	0,26708
	Québec	67	3,1104	0,67356	0,08229
Évaluation du contenu des graphiques	Hors-Québec	6	2,9167	0,62583	0,25550
	Québec	67	3,0075	0,63659	0,07777
Évaluation de la description du profil et des pistes d'intervention	Hors-Québec	6	3,3333	0,58689	0,23960
	Québec	66	3,2298	0,44985	0,05537
Évaluation du guide d'interprétation	Hors-Québec	6	3,2917	0,55715	0,22746
	Québec	62	3,2238	0,52943	0,06724
Évaluation de l'ensemble du rapport	Hors-Québec	6	3,1250	0,60725	0,24791
	Québec	62	2,8427	0,51461	0,06536

Tableau 1.7. Résultats descriptifs de l'évaluation des rapports selon les tranches d'âges

	Age	N	Moyenne	Écart-type	Moyenne erreur standard
Évaluation de la qualité des rapports	20-39 ans	30	3,1422	0,34662	0,06328
	40 ans et plus	38	3,0842	0,50013	0,08113
Évaluation des directives	20-39 ans	35	3,0286	0,64864	0,10964
	40 ans et plus	41	3,0488	0,62165	0,09709
Évaluation de la présentation des graphiques	20-39 ans	33	3,1212	0,63234	0,11008
	40 ans et plus	40	3,0700	0,70682	0,11176
Évaluation du contenu des graphiques	20-39 ans	33	2,8939	0,56607	0,09854
	40 ans et plus	40	3,0875	0,67594	0,10688
Évaluation de la description du profil et des pistes d'intervention	20-39 ans	33	3,2879	0,45122	0,07855
	40 ans et plus	39	3,1966	0,46653	0,07470
Évaluation du guide d'interprétation	20-39 ans	30	3,2792	0,44125	0,08056
	40 ans et plus	38	3,1908	0,59042	0,09578
Évaluation de l'ensemble du rapport	20-39 ans	30	2,8750	0,42926	0,07837
	40 ans et plus	38	2,8618	0,59473	0,09648

Tableau 1.8. Résultats descriptifs de l'évaluation des rapports selon l'ancienneté

Expérience		Évaluation de la qualité des rapports	Évaluation des directives	Évaluation de la présentation des graphiques	Évaluation du contenu des graphiques	Évaluation de la description du profil et des pistes d'intervention	Évaluation du guide d'interprétation	Évaluation de l'ensemble du rapport
Moins de 10 ans	Moyenne	3,1500	2,9733	3,0174	2,9565	3,2899	3,3375	2,8750
	N	20	25	23	23	23	20	20
	Écart-type	0,35467	0,70000	0,57813	0,60138	0,43871	0,43886	0,49003
	Minimum	2,47	1,33	2,00	2,00	2,50	2,75	2,00
	Maximum	4,00	4,00	4,00	4,00	4,00	4,00	4,00
11-20 ans	Moyenne	3,0848	3,0370	3,0629	2,9143	3,2095	3,2071	2,8929
	N	35	36	35	35	35	35	35
	Écart-type	0,46900	0,57888	0,73847	0,61220	0,47373	0,57598	0,53990
	Minimum	1,57	1,67	1,20	1,50	2,00	1,00	1,00
	Maximum	3,97	4,00	4,00	4,00	4,00	4,00	4,00
20 ans et plus	Moyenne	3,1154	3,1556	3,2800	3,2667	3,2262	3,1250	2,7885
	N	13	15	15	15	14	13	13
	Écart-type	0,48926	0,65304	0,64054	0,69093	0,47863	0,52787	0,56685
	Minimum	2,27	2,00	2,00	2,00	2,67	2,00	2,00
	Maximum	3,93	4,00	4,00	4,00	4,00	4,00	4,00

Tableau 1.9. Résultats descriptifs de l'évaluation des rapports selon le diplôme obtenu

Diplôme		Évaluation de la qualité des rapports	Évaluation des directives	Évaluation de la présentation des graphiques	Évaluation du contenu des graphiques	Évaluation de la description du profil et des pistes d'intervention	Évaluation du guide d'interprétation	Évaluation de l'ensemble du rapport
Baccalauréat	Moyenne	3,1144	3,0882	3,0970	3,0379	3,1970	3,2583	2,8250
	N	30	34	33	33	33	30	30
	Écart-type	0,44787	0,56443	0,64638	0,60635	0,45349	0,58991	0,61989
	Minimum	1,57	2,00	1,60	2,00	2,00	1,00	1,00
	Maximum	3,93	4,00	4,00	4,00	4,00	4,00	4,00
Microprogramme ou DESS	Moyenne	3,0708	3,0784	3,1059	2,9559	3,1979	3,1719	2,8594
	N	16	17	17	17	16	16	16
	Écart-type	0,34488	0,46442	0,71105	0,72476	0,38112	0,28822	0,36479
	Minimum	2,53	2,00	2,00	2,00	2,67	2,75	2,50
	Maximum	4,00	4,00	4,00	4,00	4,00	4,00	4,00
Maitrise et doctorat	Moyenne	3,1318	2,9467	3,0783	2,9783	3,3261	3,2330	2,9318
	N	22	25	23	23	23	22	22
	Écart-type	0,49564	0,80323	0,70256	0,62118	0,51854	0,58794	0,49511
	Minimum	2,07	1,33	1,20	1,50	2,33	2,00	2,00

Tableau 1.10. Résultats descriptifs de l'évaluation des rapports selon le domaine de formation

Domaine de formation		Évaluation de la qualité des rapports	Évaluation des directives	Évaluation de la présentation des graphiques	Évaluation du contenu des graphiques	Évaluation de la description du profil et des pistes d'intervention	Évaluation du guide d'interprétation	Évaluation de l'ensemble du rapport
Enseignement au primaire	Moyenne	3,1186	3,1111	3,1676	3,0541	3,2130	3,1838	2,8309
	N	34	39	37	37	36	34	34
	Écart-type	0,49937	0,58406	0,72650	0,70989	0,47578	0,64052	0,56649
	Minimum	1,57	2,00	1,20	1,50	2,00	1,00	1,00
	Maximum	3,93	4,00	4,00	4,00	4,00	4,00	4,00
Enseignement en adaptation scolaire	Moyenne	3,0067	2,7879	2,9364	2,8409	3,2197	3,2313	2,7500
	N	20	22	22	22	22	20	20
	Écart-type	0,27543	0,61330	0,57698	0,46640	0,43152	0,35651	0,40555
	Minimum	2,47	1,33	2,00	2,25	2,50	2,88	2,00
	Maximum	3,60	4,00	4,00	4,00	4,00	4,00	3,50
Psychopédagogie et autre	Moyenne	3,2357	3,2222	3,1429	3,1071	3,3333	3,3393	3,1250
	N	14	15	14	14	14	14	14
	Écart-type	0,45469	0,69769	0,65365	0,63332	0,47592	0,43696	0,51655
	Minimum	2,47	1,67	2,00	2,00	2,50	2,75	2,25
	Maximum	4,00	4,00	4,00	4,00	4,00	4,00	4,00

Tableau 1.11. Résultats descriptifs de l'évaluation des rapports selon le suivi ou non des cours en méthodes quantitatives

	Pendant les cinq dernières années, avez-vous suivi des cours en méthodes quantitatives?	N	Moyenne	Écart-type	Moyenne erreur standard
Évaluation de la qualité des rapports	Oui	13	3,3154	0,38094	0,10565
	Non	55	3,0612	0,43833	0,05910
Évaluation des directives	Oui	16	3,1875	0,77907	0,19477
	Non	60	3,0000	0,58545	0,07558
Évaluation de la présentation des graphiques	Oui	15	3,2400	0,55652	0,14369
	Non	58	3,0552	0,69564	0,09134
Évaluation du contenu des graphiques	Oui	15	3,2667	0,53841	0,13902
	Non	58	2,9310	0,64003	0,08404
Évaluation de la description du profil et des pistes d'intervention	Oui	15	3,5111	0,43400	0,11206
	Non	57	3,1667	0,44096	0,05841
Évaluation du guide d'interprétation	Oui	13	3,4231	0,41938	0,11632
	Non	55	3,1841	0,54372	0,07332
Évaluation de l'ensemble du rapport	Oui	13	2,9808	0,51500	0,14284
	Non	55	2,8409	0,52784	0,07117

Tableau 1.12. Résultats descriptifs de l'évaluation des rapports et la préférence des participants.

	Questions		Évaluation de la qualité des rapports	Évaluation des directives	Évaluation de la présentation des graphiques	Évaluation du contenu des graphiques	Évaluation de la description du profil et des pistes d'intervention	Évaluation du guide d'interprétation	Évaluation de l'ensemble du rapport
Q1	Oui	Moyenne	**3,2688**	**3,1944**	**3,5000**	**3,2059**	**3,3182**	**3,3320**	**2,9922**
		Écart-type	0,38808	0,57113	0,53541	0,61997	0,45331	0,47900	0,48975
	Non	Moyenne	2,9685	2,9000	2,7385	2,8205	3,1709	3,1389	2,7569
		Écart-type	0,43424	0,65459	0,56968	0,59313	0,45801	0,55884	0,53614
Q2	Oui	Moyenne	**3,2387**	**3,2000**	**3,4364**	**3,1742**	**3,3021**	**3,3024**	**2,9677**
		Écart-type	0,42252	0,60607	0,57327	0,60752	0,48163	0,48603	0,51536
	Non	Moyenne	3,0018	2,9024	2,8100	2,8563	3,1875	3,1689	2,7838
		Écart-type	0,42470	0,62470	0,61427	0,62246	0,43883	0,55996	0,52428
Q3	Oui	Moyenne	**3,3108**	**3,2059**	**3,5188**	**3,2578**	**3,3656**	**3,3629**	**3,0323**
		Écart-type	0,38426	0,56862	0,49477	0,60069	0,45830	0,49408	0,48624
	Non	Moyenne	2,9414	2,9048	2,7610	2,7988	3,1423	3,1182	2,7297
		Écart-type	0,41047	0,65139	0,60037	0,58689	0,44027	0,53598	0,52168
Q4	Oui	Moyenne	**3,2758**	**3,2072**	**3,4743**	**3,1929**	**3,3382**	**3,3523**	**3,0076**
		Écart-type	0,38408	0,56847	0,54898	0,61562	0,46145	0,48559	0,49008
	Non	Moyenne	2,9533	2,8803	2,7421	2,8224	3,1491	3,1143	2,7357
		Écart-type	0,43078	0,65108	0,57687	0,60098	0,44315	0,54686	0,52840

Tableau 1.13. Résultats descriptifs de l'évaluation de la qualité et la compréhension des rapports

Question	Réponses		Évaluation						
		Statistiques	De la qualité	Directives	Présentation des graphiques	Contenu des graphiques	Descrip -tion du profil	Guide inter- prétation	Ensemble du rapport
1-Combien d'habiletés l'élève doit-il maitriser pour le test?	Incorrectes	Moyenne	3,0556	2,8333	2,8667	**3,1667**	3,0278	3,2500	**3,0000**
		Écart-type	0,3879	0,8628	0,7659	0,5400	0,3234	0,5000	0,6123
		Moyenne	**3,1368**	**3,0936**	**3,1404**	2,9912	**3,2719**	**3,2632**	2,8684
		Écart-type	0,4408	0,5833	0,6675	0,6338	0,4814	0,5259	0,5156
2-L'élève a-t-il bien maitrisé toutes les habiletés attendues?	Incorrectes	Moyenne	**3,1667**	**3,1667**	3,1000	**3,1250**	3,000	**3,3750**	**3,1250**
		Écart-type	0,1414	1,1785	0,7071	0,1767	0,000	0,5303	0,1767
	Correctes	Moyenne	3,1226	3,0645	**3,1065**	2,9960	**3,2473**	3,2540	2,8629
		Écart-type	0,4403	0,5970	0,6782	0,6313	0,4797	0,5200	0,5292
3-L'élève a-t-il bien maitrisé l'habileté «Repérer des informations explicites » ?	Incorrectes	Moyenne	2,7000	2,3000	2,4000	2,7500	2,7500	3,0000	2,6250
		Écart-type	0,3299	0,9428	0,5656	0,0000	0,3535	0,0000	0,5303
	Correctes	Moyenne	**3,1376**	**3,0914**	**3,1290**	**3,0081**	**3,2554**	**3,2661**	**2,8790**
		Écart-type	0,4318	0,5888	0,6678	0,6304	0,4711	0,5227	0,5254
4- L'élève a-t-il bien maitrisé l'habileté «Compréhension globale » ?	Incorrectes	Moyenne	3,1000	2,7777	2,9333	2,9167	3,2222	**3,4167**	**2,9167**
		Écart-type	0,4176	0,3849	0,5773	0,3818	0,6938	0,3818	0,6291
	Correctes	Moyenne	**3,1251**	**3,0820**	**3,1148**	**3,0041**	**3,2404**	3,2500	2,8689
		Écart-type	0,4378	0,6136	0,6805	0,6332	0,4689	0,5234	0,5236
5-Le niveau de maitrise de l'élève est-il identique pour les cinq habiletés?	Incorrectes	Moyenne	3,0033	3,0000	2,8000	**3,3750**	3,1667	2,8750	**3,1250**
		Écart-type	0,0471	0,4714	0,2828	0,1767	0,0000	0,1767	0,1767
	Correctes	Moyenne	**3,1269**	**3,0660**	**3,1161**	2,9879	**3,2419**	**3,2702**	2,8629
		Écart-type	0,4404	0,6129	0,6809	0,6279	0,4816	0,5196	0,5292
6-Dans le graphique, quel est le niveau de maitrise de l'élève pour l'habileté « Interpréter »?	Incorrectes	Moyenne	2,6667	2,1111	2,2000	2,8333	2,6667	3,2083	2,4167
		Écart-type	0,3464	0,3849	0,3464	0,1443	0,2886	0,4732	0,5204
	Correctes	Moyenne	**3,1464**	**3,1148**	**3,1508**	**3,0082**	**3,2678**	**3,2602**	**2,8934**
		Écart-type	0,4273	0,5769	0,6548	0,6556	0,4647	0,5220	0,5173
7-Si on considère que le seuil de réussite est 0,5, lequel de ces énoncés est vrai?	Incorrectes	Moyenne	2,6667	2,6190	2,6571	2,7143	2,7143	2,7321	2,4643
		Écart-type	0,7081	0,6506	0,9217	0,6362	0,6434	0,9197	0,9290
	Correctes	Moyenne	**3,1801**	**3,1228**	**3,1614**	**3,0351**	**3,3041**	**3,3224**	**2,9211**
		Écart-type	0,3586	0,5829	0,6250	0,6168	0,4120	0,4131	0,4385
8-Dans le graphique sur le profil de maitrise des habiletés du groupe, pour l'habi -leté « compréhension globale », quel groupe a le pourcentage le plus élevé?	Incorrectes	Moyenne	3,0641	2,8462	2,8923	2,8269	**3,2692**	**3,3173**	2,8654
		Écart-type	0,3730	0,7529	0,6563	0,4828	0,5294	0,4070	0,5554
	Correctes	Moyenne	**3,1392**	**3,1242**	**3,1608**	**3,0441**	3,2320	3,2426	**2,8725**
		Écart-type	0,4500	0,5576	0,6729	0,6494	0,4643	0,5433	0,5205
9-Pour l'habileté « faire des inférences», le pourcen- tage des élèves ayant les niveaux de maitrise de 0,71 à 1 est-il le plus élevé?	Incorrectes	Moyenne	3,0604	3,0625	2,8875	**3,0156**	3,1979	3,2109	2,8125
		Écart-type	0,6088	0,5737	0,8196	0,7098	0,5748	0,7688	0,7041
	Correctes	Moyenne	**3,1451**	**3,0694**	**3,1792**	2,9948	**3,2535**	**3,2734**	**2,8906**
		Écart-type	0,3634	0,6225	0,6094	0,5980	0,4416	0,4098	0,4551

ANNEXE 2-VÉRIFICATION DES POSTULATS DE L'ANOVA ET DU TEST T

Afin de vérifier la normalité des variables testées, nous avons fait des tests Kolmogorov-Smirnov pour un échantillon, les résultats obtenus (tabbleau 2.1) suggèrent que l'évaluation de la qualité des rapports, l'évaluation de la présentation des graphiques suivent la loi normale. Par contre, ce postulat n'a pas été respecté pour l'évaluation des directives, l'évaluation du contenu des graphiques, l'évaluation de la description du profil et des pistes d'intervention ainsi que l'évaluation du guide d'interprétation et l'évaluation de l'ensemble du rapport. Toutefois, les tests d'Anova sont robustes à la violation de la normalité si les distributions des sous-groupes se ressemblent. Afin de le vérifier, les tests de Kruskal-Wallis ont été réalisés et les résultats suggèrent que les distributions des sous-groupes se ressemblent pour tous ces facteurs. Nous considérons que le postulat sur la normalité est donc respecté.

Tableau 2.1. Résultats des tests Kolmogorov-Smirnov pour un échantillon pour vérifier de la normalité.

Test Kolmogorov-Smirnov pour un échantillon

		Évaluation de la qualité des rapports	Évaluation des directives	Évaluation de la présentation des graphiques	Évaluation du contenu des graphiques	Évaluation de la description du profil et des pistes d'intervention	Évaluation du guide d'interprétation	Évaluation de l'ensemble du rapport
N		68	76	73	73	72	68	68
Paramètres normaux[a,b]	Moyenne	3,1098	3,0395	3,0932	3,0000	3,2384	3,2298	2,8676
	Écart-type	0,43701	0,63003	0,67006	0,63191	0,45864	0,52795	0,52453
Différences les plus extrêmes	Absolue	0,099	0,172	0,104	0,133	0,171	0,170	0,136
	Positif	0,069	0,130	0,103	0,133	0,171	0,123	0,136
	Négatif	-0,099	-0,172	-0,104	-0,094	-0,107	-0,170	-0,109
Statistiques de test		0,099	0,172	0,104	0,133	0,171	0,170	0,136
Sig. asymptotique (bilatérale)		0,099[c]	0,000[c]	0,050[c]	0,003[c]	0,000[c]	0,000[c]	0,003[c]

ANNEXE 3-RÉSULTATS DE L'ANOVA À UN FACTEUR

Tableau 3.1. Évaluation de la qualité des rapports selon le format de rapport évalué

	F	p	Somme des carrés	ddl	Carré-moyen	F	p	Eta-carré partiel
Évaluation de la qualité globale des rapports	1,237	0,297	0,032	2	,016	0,021	0,921	0,003
Évaluation des directives	0,626	0,537	1,340	2	0,670	1,720	0,186	0,045
Évaluation de la présentation des graphiques	0,624	0,539	0,067	2	0,034	0,073	0,930	0,002
Évaluation du contenu des graphiques	5,492	0,006	2,120	2	1,060	2,787	0,068	0,074
Évaluation de la description du profil et des pistes d'intervention	0,918	0,404	0,059	2	0,029	0,137	0,872	0,004
Évaluation du guide d'interprétation	0,521	0,597	0,072	2	0,036	0,125	0,882	0,004
Évaluation de l'ensemble du rapport	0,685	0,508	0,004	2	0,002	0,007	0,993	0,000

Tableau 3.2. Évaluation de la qualité des rapports selon l'ancienneté

	F	p	Somme des carrés	ddl	Carré-moyen	F	p	Eta-carré partiel
Évaluation de la qualité globale des rapports	0,870	0,424	0,055	2	0,027	0,139	0,870	0,004
Évaluation des directives	0,774	0,465	0,312	2	0,156	0,386	0,681	0,010
Évaluation de la présentation des graphiques	0,945	0,394	0,688	2	0,344	0,761	0,471	0,021
Évaluation du contenu des graphiques	0,471	0,626	1,367	2	0,684	1,748	0,182	0,048
Évaluation de la description du profil et des pistes d'intervention	0,064	0,938	0,092	2	0,046	0,214	0,808	0,006
Évaluation du guide d'interprétation	0,004	0,996	0,393	2	0,196	0,698	0,501	0,021
Évaluation de l'ensemble du rapport	0,104	0,901	0,105	2	0,052	0,186	0,831	0,006

Tableau 3.3. Évaluation de la qualité des rapports selon le diplôme obtenu

	F	p	Somme des carrés	ddl	Carré-moyen	F	p	Éta-carré partiel
Évaluation de la qualité globale des rapports	1,187	0,312	0,036	2	0,018	0,091	0,913	0,003
Évaluation des directives	4,686	0,012	0,332	2	0,161	0,399	0,672	0,011
Évaluation de la présentation des graphiques	0,364	0,696	0,008	2	0,004	0,009	0,991	0,000
Évaluation du contenu des graphiques	0,981	0,380	0,091	2	0,046	0,112	0,895	0,003
Évaluation de la description du profil et des pistes d'intervention	1,686	0,193	0,260	2	0,130	0,611	0,546	0,017
Évaluation du guide d'interprétation	3,572	0,034	0,078	2	0,039	0,137	0,872	0,004
Évaluation de l'ensemble du rapport	2,609	0,081	0,146	2	0,073	0,260	0,772	0,008

Tableau 3.4. Évaluation de la qualité des rapports selon leur domaine de formation

	F	p	Somme des carrés	ddl	Carré-moyen	F	p	Eta-carré partiel
Évaluation de la qualité globale des rapports	2,819	0,067	0,437	2	0,219	1,150	0,323	0,034
Évaluation des directives	0,204	0,816	2,094	2	1,047	2,761	0,070	0,070
Évaluation de la présentation des graphiques	1,393	0,255	0,780	2	0,390	0,866	0,425	0,024
Évaluation du contenu des graphiques	3,188	0,047	0,826	2	0,413	1,035	0,361	0,029
Évaluation de la description du profil et des pistes d'intervention	0,104	0,902	0,157	2	0,079	0,367	0,694	0,011
Évaluation du guide d'interprétation	1,913	0,156	0,240	2	0,120	0,423	0,637	0,013
Évaluation de l'ensemble du rapport	0,425	0,655	1,250	2	0,625	2,364	0,102	0,068

ANNEXE 4-RÉSULTATS DU TEST T DE STUDENT

Tableau 4.1. Évaluation de la qualité des rapports selon les régions de provenance

		Test de Levene sur l'égalité des variances		Test t pour égalité des moyennes		
		F	Sig.	t	ddl	Sig. (bilatéral)
Évaluation de la qualité des rapports	Hypothèse de variances égales	0,429	0,515	0,105	66	0,917
Évaluation des directives	Hypothèse de variances égales	0,060	0,807	-0,383	74	0,703
Évaluation de la présentation des graphiques	Hypothèse de variances égales	0,591	0,445	-0,735	71	0,465
Évaluation du contenu des graphiques	Hypothèse de variances égales	0,115	0,736	-0,335	71	0,739
Évaluation de la description du profil et des pistes d'intervention	Hypothèse de variances égales	0,416	0,521	0,527	70	0,600
Évaluation du guide d'interprétation	Hypothèse de variances égales	0,324	0,571	0,299	66	0,766
Évaluation de l'ensemble du rapport	Hypothèse de variances égales	0,236	0,629	1,264	66	0,211

Tableau 4.2. Évaluation de la qualité des rapports selon les tranches d'âges

		Test de Levene sur l'égalité des variances		Test t pour égalité des moyennes		
		F	Sig.	t	ddl	Sig. (bilatéral)
Évaluation de la qualité des rapports	Hypothèse de variances égales	2,323	0,132	0,541	66	0,591
Évaluation des directives	Hypothèse de variances égales	0,004	0,951	-0,138	74	0,890
Évaluation de la présentation des graphiques	Hypothèse de variances égales	0,020	0,887	0,323	71	0,748
Évaluation du contenu des graphiques	Hypothèse de variances égales	1,779	0,187	-1,309	71	0,195
Évaluation de la description du profil et des pistes d'intervention	Hypothèse de variances égales	0,017	0,898	0,840	70	0,404
Évaluation du guide d'interprétation	Hypothèse de variances égales	0,322	0,572	0,683	66	0,497
Évaluation de l'ensemble du rapport	Hypothèse de variances égales	1,661	0,202	0,102	66	0,919

Tableau 4.3. . Évaluation de la qualité des rapports selon le suivi ou non des cours en méthodes quantitatives

		Test de Levene sur l'égalité des variances		Test t pour égalité des moyennes		
		F	Sig.	t	ddl	Sig. (bilatéral)
Évaluation de la qualité des rapports	Hypothèse de variances égales	0,036	0,850	1,924	66	0,059
Évaluation des directives	Hypothèse de variances égales	1,648	0,203	1,059	74	0,293
Évaluation de la présentation des graphiques	Hypothèse de variances égales	0,793	0,376	0,952	71	0,345
Évaluation du contenu des graphiques	Hypothèse de variances égales	0,939	0,336	1,865	71	0,066
Évaluation de la description du profil et des pistes d'intervention	Hypothèse de variances égales	0,078	0,781	2,700	70	0,009
Évaluation du guide d'interprétation	Hypothèse de variances égales	0,103	0,750	1,481	66	0,143
Évaluation de l'ensemble du rapport	Hypothèse de variances égales	0,030	0,863	0,863	66	0,391

Tableau 4.4. Évaluation de la qualité des rapports selon la préférence

		Test de Levene sur l'égalité des variances		Test t pour égalité des moyennes		
		F	Sig.	t	ddl	Sig. (bilatéral)
Évaluation globale de la qualité des rapports	Q1	0,031	0,860	2,991	66	**0,004**
	Q2	0,231	0,633	2,296	66	**0,025**
	Q3	0,035	0,852	3,804	66	**0,000**
	Q4	0,006	0,941	3,250	66	**0,002**
Évaluation des directives	Q1	0,050	0,823	2,079	74	**0,041**
	Q2	0,014	0,907	2,098	74	**0,039**
	Q3	0,019	0,891	2,119	74	**0,037**
	Q4	0,051	0,822	2,326	74	**0,023**
Évaluation de la présentation des graphiques	Q1	0,101	0,751	5,858	71	**0,000**
	Q2	0,105	0,747	4,468	71	**0,000**
	Q3	0,166	0,685	5,770	71	**0,000**
	Q4	0,213	0,645	5,544	71	**0,000**
Évaluation du contenu des graphiques	Q1	0,130	0,719	2,711	71	**0,008**
	Q2	0,000	0,992	2,196	71	**0,031**
	Q3	0,077	0,782	3,282	71	**0,002**
	Q4	0,061	0,805	2,601	71	**0,011**
Évaluation de la description du profil et des pistes d'intervention	Q1	0,067	0,797	1,366	70	0,176
	Q2	0,213	0,275	1,054	70	0,295
	Q3	0,478	0,492	2,094	70	**0,040**
	Q4	0,492	0,485	1,773	70	0,081
Évaluation du guide d'interprétation	Q1	0,005	0,943	1,520	66	0,133
	Q2	0,001	0,977	1,039	66	0,303
	Q3	0,222	0,639	1,942	66	0,056
	Q4	0,168	0,683	1,893	66	0,063
Évaluation de l'ensemble du rapport	Q1	0,031	0,860	1,881	66	0,064
	Q2	0,135	0,715	1,452	66	0,151
	Q3	0,001	0,972	2,456	66	**0,017**
	Q4	0,000	0,990	2,196	66	**0,032**

ANNEXE 5-DOCUMENTS DE TRAVAIL AVEC LE PANEL D'EXPERTS 2

Tableau 5.1. Répartition de réponse de l'élève (profil 4) présenté au panel d'experts

Partie 1			Partie 2		
Items	**Réponses**	**Habiletés**	**Items**	**Réponses**	**Habiletés**
1 (Q1)	CO	CG; FI	17 (Q1)	CO	IE
2 (Q2	IC	IN; VS	18 (Q2)	AB	IE; IN
3 (Q3)	CO	IE	19 (Q3	CO	IE
4 (Q4)	IC	IN; FI	20 (Q4)	IC	IE; CG; IN
5 (Q5)	CO	IE; FI; VS	21 (Q5)	IC	IN; FI
6 (Q6)	CO	IE	22 (Q6)	CO	IE
7 (Q7)	CO	IE	23 (Q7)	IC	IN; FI
8 (Q8)	CO	FI	24 (Q8A)	AB	IN; VS
9 (Q9)	IC	IN; VS	25 (Q8B)	AB	IN; VS
10 (Q10)	CO	IN; FI	26 (Q8C)	AB	IN; VS
11 (Q11)	CO	IN; FI	27 (Q9)	CO	IN; FI
12 (Q12)	CO	IN; FI	28 (Q10)	AB	IN; VS
13 (Q13)	CO	CG; FI	29 (Q11)	IC	IE
14 (Q14)	CO	CG; IN; VS	30 (Q12)	AB	CG; VS
15 (Q15)	AB	IE; CG; IN; FI; VS	31 (Q13A)	AB	IE; IN
16 (Q16)	CO	CG; IN; VS	32 (Q13B)	AB	IE; IN
			33 (Q13C)	AB	IE; IN
			34 (Q13ABC)	AB	IE; IN
			35 (Q14)	AB	IE

CO=Réponse correcte; PC= Partiellement correcte; IC=Réponse incorrecte; AB=Abandonner; IE=Informations explicites; CG=Compréhension globale; IN=Interprétation; FI=Faire des inférences; VS=Vocabulaire et syntaxe

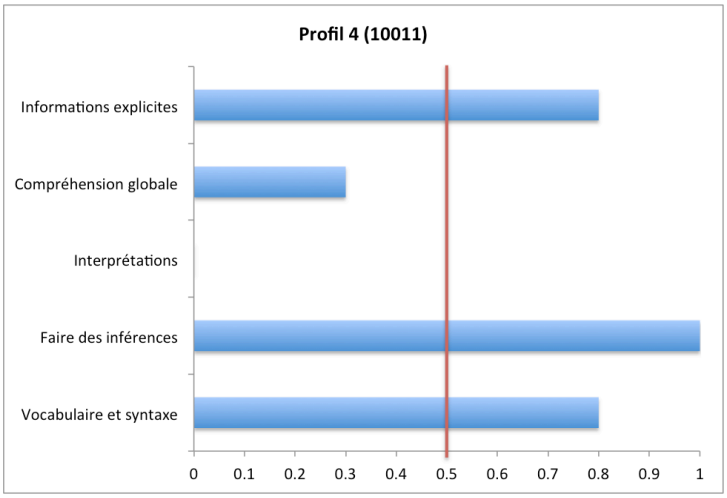


Figure 5.1. Exemple d'un profil d'élève présenté au panel d'experts.

Tableau 5.2. Fiche d'évaluation des rapports fournis au panel d'experts.

	Liste de vérification	Vos commentaires ou suggestions
Fond	Titre	
	Partie 1 (Identification de l'élève)	
	Partie 2 (Profil de maitrise des habiletés	
	Partie 4 (Guide d'interprétation des résultats)	
	Légendes	
Forme	Disposition des parties	
	Taille des graphiques	
	Format et taille des polices	
	Couleur des graphiques	
	Couleur du background	

ANNEXE 6-QUESTIONNAIRE DE L'ÉVALUATION DES RAPPORTS

LETTRE D'INFORMATION ET DE CONSENTEMENT à l'attention des participants

Vous êtes invité à participer à un projet de recherche. Avant d'accepter, veuillez prendre le temps de lire ce document présentant les conditions de participation au projet. N'hésitez pas à poser toutes les questions à la chercheuse dont les coordonnées se trouvent en bas de la page.

A) RENSEIGNEMENTS AUX PARTICIPANTS

1. Participation à la recherche

Votre participation à la recherche consiste à répondre à un questionnaire sur l'évaluation des rapports diagnostiques destinés aux enseignants. Le questionnaire se compose de 24 questions à choix de réponse portant sur la qualité des rapports et sur votre compréhension de leur contenu, ainsi que sur des informations sociodémographiques et nous évaluons le temps pour répondre au questionnaire à environ 30 minutes.

2. Risques et inconvénients

Outre le temps consacré au questionnaire, il n'y a pas de risque ni d'inconvénient particulier à participer à ce projet. Vous pourrez à tout moment mettre fin à votre participation.

3. Avantages et bénéfices

Il n'y a pas d'avantage particulier à participer à ce projet. Vous contribuerez cependant à une meilleure compréhension sur les regards que portent des enseignants au primaire, des étudiants, des conseillers pédagogiques et des orthopédagogues sur les formats de rapports diagnostiques élaborés.

4. Confidentialité

Les renseignements personnels que vous nous donnerez demeureront confidentiels. L'adresse IP des participants sera enregistrée. Cependant, aucune information permettant de vous identifier d'une façon ou d'une autre ne sera publiée. Chaque participant à la recherche se verra attribuer un code et seules la chercheuse et son équipe pourront connaitre son identité. Les données seront conservées dans un lieu sûr et seront détruites sept ans après la fin du projet.

5. Compensation

Aucune compensation n'est prévue pour votre participation à la recherche.

6. Droit de retrait

Votre participation est entièrement volontaire. Vous pouvez vous retirer de cette recherche à tout moment sur simple avis verbal, sans préjudice et sans devoir justifier votre décision. Si vous décidez de vous retirer de la recherche, vous pouvez communiquer avec la chercheure, au numéro de téléphone indiqué ci-dessous en lui fournissant votre adresse IP. Les informations qui auront été recueillies au moment du retrait, seront détruites et votre retrait n'aura aucune conséquence sur votre cheminement scolaire ou votre carrière professionnelle.

7. Coordonnées

Pour toute question relative à l'étude, ou pour vous retirer de la recherche, veuillez communiquer avec **Dan Thanh Duong Thi** , par tel : [redacted] ou par courriel: [redacted]

Pour toute préoccupation sur vos droits ou sur les responsabilités des chercheurs concernant votre participation à ce projet, veuillez contacter le Comité plurifacultaire d'éthique de la recherche par courriel: [redacted] ou par téléphone: [redacted] consulter le site Web http://recherche.umontreal.ca/participants.

Toute plainte relative à votre participation à cette recherche peut être adressée à l'ombudsman de l'Université de Montréal, par tel: [redacted] ou par courriel: [redacted]

B) CONSENTEMENT

Déclaration du participant

- Je comprends que je peux prendre mon temps pour réfléchir avant de donner mon accord ou non à participer à la recherche.
- Je peux poser des questions à l'équipe de recherche et exiger des réponses satisfaisantes.
- Je comprends qu'en participant à ce projet de recherche, je ne renonce à aucun de mes droits ni ne dégage les chercheurs de leurs responsabilités.

* 1. **Après avoir lu les informations sur la recherche et avoir réfléchi à son implication, consentez-vous à y participer ?**

◯ Oui, je consens à participer à cette recherche

◯ Non, je ne consens pas à participer à cette recherche

Je souhaite avoir une copie écrite de mon consentement signée par la chercheuse (veuillez inscrire une adresse courriel).

Identification

* 2. **Vous êtes:**

- ◯ Étudiant(e) en enseignement au primaire
- ◯ Enseignant(e) au primaire
- ◯ Conseiller(ère) pédagogique
- ◯ Orthopédagoque

* 3. **Quelle est votre région de provenance?**

- ◯ Québec
- ◯ Ontario
- ◯ Nouveau-Brunswick
- ◯ Autre province du Canada
- ◯ Autre pays

* 4. **De quel sexe êtes-vous?**

- ◯ Féminin
- ◯ Masculin

* 5. **À quel groupe d'âge appartenez-vous?**

- ◯ Moins de 20 ans
- ◯ 20 à 29 ans
- ◯ 30 à 39 ans
- ◯ 40 à 49 ans
- ◯ 50 à 59 ans
- ◯ 60 ans et plus

* 6. **Combien d'années d'expérience comptez-vous en enseignement?**

- ◯ Moins d'un an (stage, suppléances)
- ◯ De 1 à 5 ans
- ◯ 6 à 10 ans
- ◯ 11 à 15 ans
- ◯ 16 à 20 ans
- ◯ 21 à 25 ans
- ◯ Plus de 25 ans

* 7. **Quel est le niveau de scolarité le plus élevé que vous ayez complété ou en voie d'obtention?**

- ◯ 1er cycle (baccalauréat)
- ◯ 2e cycle (microprogramme ou DESS)
- ◯ 2e cycle (maîtrise)
- ◯ 3e cycle (doctorat)
- ◯ Autre (veuillez préciser)

* 8. **Quel est votre domaine de formation?**

- ◯ Enseignement au primaire et préscolaire
- ◯ Enseignement en adaption scolaire
- ◯ Psychopédagogie
- ◯ Autre (veuillez préciser)

* 9. **Pendant les cinq dernières années, avez-vous suivi des cours en méthodes quantitatives?**

- ◯ Oui
- ◯ Non

Section 1. Préférence des rapports

Avec les trois formats de rapports diagnostiques affichés à l'écran, choisissez un format qui correspond le plus à votre préférence

* 10. **Veuillez cliquer sur les formats de rapports suivants pour les visualiser.**

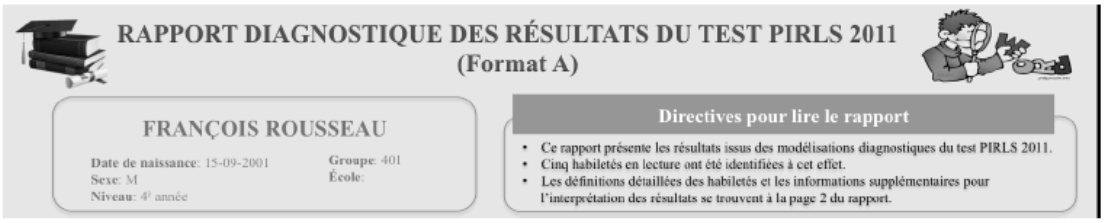
Format A Format B Format C

	Format A	Format B	Format C
1-Parmi les trois formats de rapports présentés, lequel préférez-vous?	◯	◯	◯
2-Lequel, parmi ces trois formats, vous permet de comprendre le mieux **les niveaux de maitrise des habiletés de l'élève**?	◯	◯	◯
3-Lequel, parmi ces trois formats, vous permet de comprendre **le positionnement de l'élève par rapport au groupe**?	◯	◯	◯
4-Lequel, parmi ces trois formats, pensez-vous que les enseignants vont préférer?	◯	◯	◯

Section 2. Évaluation de la qualité du rapport

Dans cette section, vous donnerez vos opinions sur les parties distinctes du format A. Veuillez cliquer sur le Format A pour visualiser le rapport.

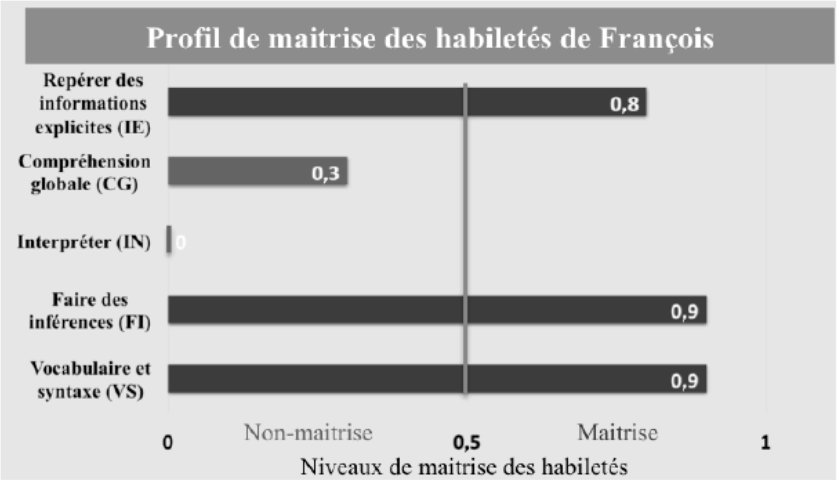
* 11. **Indiquez vos opinions sur l'identification de l'élève et les directives pour lire le rapport ci-dessous.**

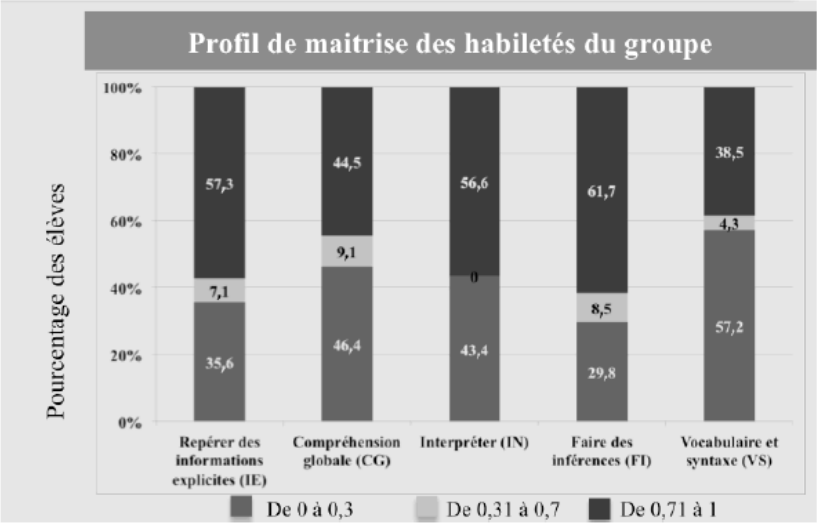


	Très en accord	En accord	En désaccord	Très en désaccord
1- Les informations fournies sur l'élève sont suffisantes pour personnaliser le rapport.	○	○	○	○
2- Les directives pour lire le rapport sont définies avec clarté.	○	○	○	○
3- Ces directives sont utiles pour mon interprétation du rapport.	○	○	○	○
4- J'aimerais que les objectifs du rapport diagnostique soient ajoutés dans les directives du rapport.	○	○	○	○

Section 2. Évaluation de la qualité du rapport (suite)

* 12. **Indiquez vos opinions sur le profil de l'élève et celui du groupe**





	Très en accord	En accord	En désaccord	Très en désaccord
1-Les graphiques présentés sont faciles à comprendre.	○	○	○	○

	Très en accord	En accord	En désaccord	Très en désaccord
2- Dans les graphiques, l'utilisation de différentes couleurs selon les niveaux de maitrise des habiletés est pertinente pour ma compréhension.	◯	◯	◯	◯
3-La taille des graphiques facilite ma lecture.	◯	◯	◯	◯
4-Les légendes sont claires pour mon interprétation des graphiques.	◯	◯	◯	◯
5-Les légendes sont utiles pour mon interprétation des graphiques.	◯	◯	◯	◯
6-Les chiffres présentés dans les barres facilitent le positionnement de l'élève par rapport au groupe.	◯	◯	◯	◯
7-Le positionnement de l'élève par rapport au groupe permet de mieux identifier les pistes d'intervention.	◯	◯	◯	◯
8-Les niveaux de maitrises des habiletés de l'élève sont faciles à comparer avec ceux du groupe.	◯	◯	◯	◯
9-J'ai confiance aux résultats diagnostiques présentés.	◯	◯	◯	◯

Section 2. Évaluation de la qualité du rapport (suite)

* 13. **Indiquez vos opinions sur les descriptions du profil de l'élève et les pistes d'intervention**

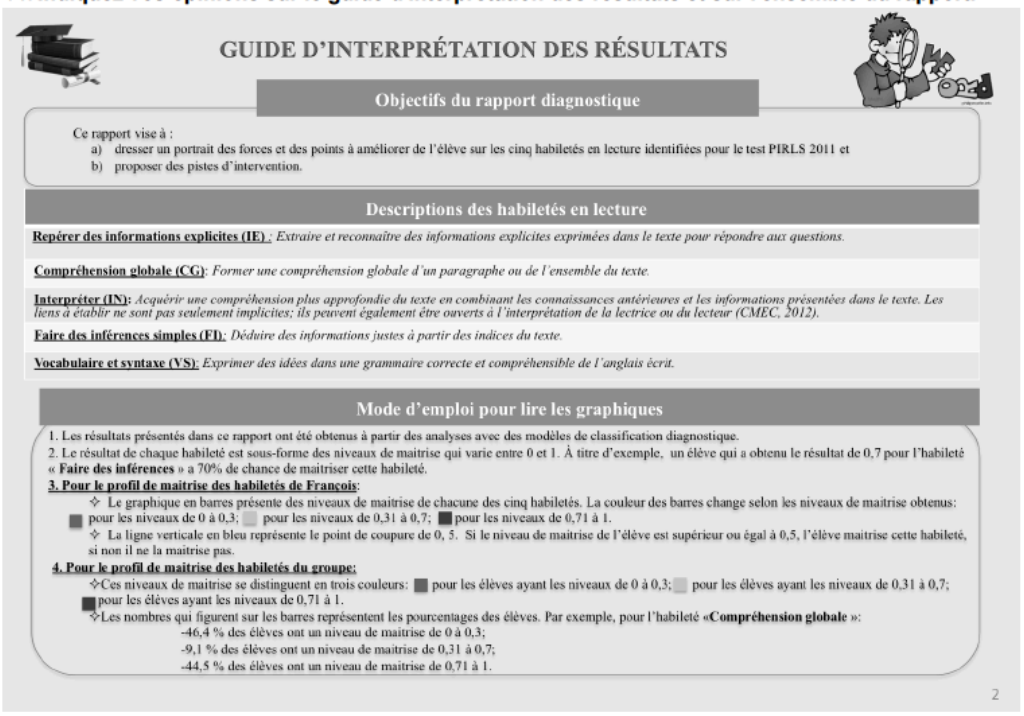
1. Descriptions du profil de l'élève
François démontre une compréhension adéquate des faits ou des événements présentés de manière explicite (IE) et implicite (FI), d'un texte courant et d'un texte littéraire. Il est capable de formuler des réponses aux questions de façon adéquate (VS).
François ne démontre pas une compréhension globale (CG) et approfondie des textes lui permettant de formuler des interprétations plausibles (IN).
2. Pistes d'intervention:
Amener l'élève à :-Cerner l'information importante dans les phrases plus longues et plus complexes (CG).
-Défendre son interprétation personnelle en donnant des raisons (IN).
-Surligner les éléments d'information qui appuient son interprétation (IN).
-Vérifier dans le texte s'il n'y pas de contradiction avec l'interprétation retenue (IN).

N.B: Ces pistes d'intervention ont été proposées à partir de différents documents ministériels tels que la progression des apprentissages; le référentiel d'intervention en lecture pour les élèves de 10 à 15 ans.

	Très en accord	En accord	En désaccord	Très en désaccord
1-La description du profil est utile pour mon interprétation du graphique.	○	○	○	○
2-La description du profil correspond à ma compréhension des graphiques.	○	○	○	○
3-Les pistes d'intervention pour l'élève (individuel) sont utiles pour mon enseignement.	○	○	○	○
4-Les pistes d'intervention pour l'élève (individuel) sont utiles pour mon évaluation.	○	○	○	○
5-J'aimerais bien avoir des pistes d'intervention pour le groupe	○	○	○	○
6-Dans l'ensemble, le vocabulaire utilisé est facile à comprendre.	○	○	○	○
7-Les paragraphes sont bien formulés.	○	○	○	○

Section 2. Évaluation de la qualité du rapport (suite)

* 14. **Indiquez vos opinions sur le guide d'interprétation des résultats et sur l'ensemble du rapport.**



GUIDE D'INTERPRÉTATION DES RÉSULTATS

Objectifs du rapport diagnostique

Ce rapport vise à :
a) dresser un portrait des forces et des points à améliorer de l'élève sur les cinq habiletés en lecture identifiées pour le test PIRLS 2011 et
b) proposer des pistes d'intervention.

Descriptions des habiletés en lecture

Repérer des informations explicites (IE) : *Extraire et reconnaître des informations explicites exprimées dans le texte pour répondre aux questions.*

Compréhension globale (CG): *Former une compréhension globale d'un paragraphe ou de l'ensemble du texte.*

Interpréter (IN): *Acquérir une compréhension plus approfondie du texte en combinant les connaissances antérieures et les informations présentées dans le texte. Les liens à établir ne sont pas seulement implicites; ils peuvent également être ouverts à l'interprétation de la lectrice ou du lecteur (CMEC, 2012).*

Faire des inférences simples (FI): *Déduire des informations justes à partir des indices du texte.*

Vocabulaire et syntaxe (VS): *Exprimer des idées dans une grammaire correcte et compréhensible de l'anglais écrit.*

Mode d'emploi pour lire les graphiques

1. Les résultats présentés dans ce rapport ont été obtenus à partir des analyses avec des modèles de classification diagnostique.
2. Le résultat de chaque habileté est sous-forme des niveaux de maitrise qui varie entre 0 et 1. À titre d'exemple, un élève qui a obtenu le résultat de 0,7 pour l'habileté « **Faire des inférences** » a 70% de chance de maitriser cette habileté.
3. **Pour le profil de maitrise des habiletés de François**:
   - ✧ Le graphique en barres présente des niveaux de maitrise de chacune des cinq habiletés. La couleur des barres change selon les niveaux de maitrise obtenus: pour les niveaux de 0 à 0,3; pour les niveaux de 0,31 à 0,7; pour les niveaux de 0,71 à 1.
   - ✧ La ligne verticale en bleu représente le point de coupure de 0, 5. Si le niveau de maitrise de l'élève est supérieur ou égal à 0,5, l'élève maitrise cette habileté, si non il ne la maitrise pas.
4. **Pour le profil de maitrise des habiletés du groupe:**
   - ✧Ces niveaux de maitrise se distinguent en trois couleurs: pour les élèves ayant les niveaux de 0 à 0,3; pour les élèves ayant les niveaux de 0,31 à 0,7; pour les élèves ayant les niveaux de 0,71 à 1.
   - ✧Les nombres qui figurent sur les barres représentent les pourcentages des élèves. Par exemple, pour l'habileté **«Compréhension globale »**:
     - -46,4 % des élèves ont un niveau de maitrise de 0 à 0,3;
     - -9,1 % des élèves ont un niveau de maitrise de 0,31 à 0,7;
     - -44,5 % des élèves ont un niveau de maitrise de 0,71 à 1.

2

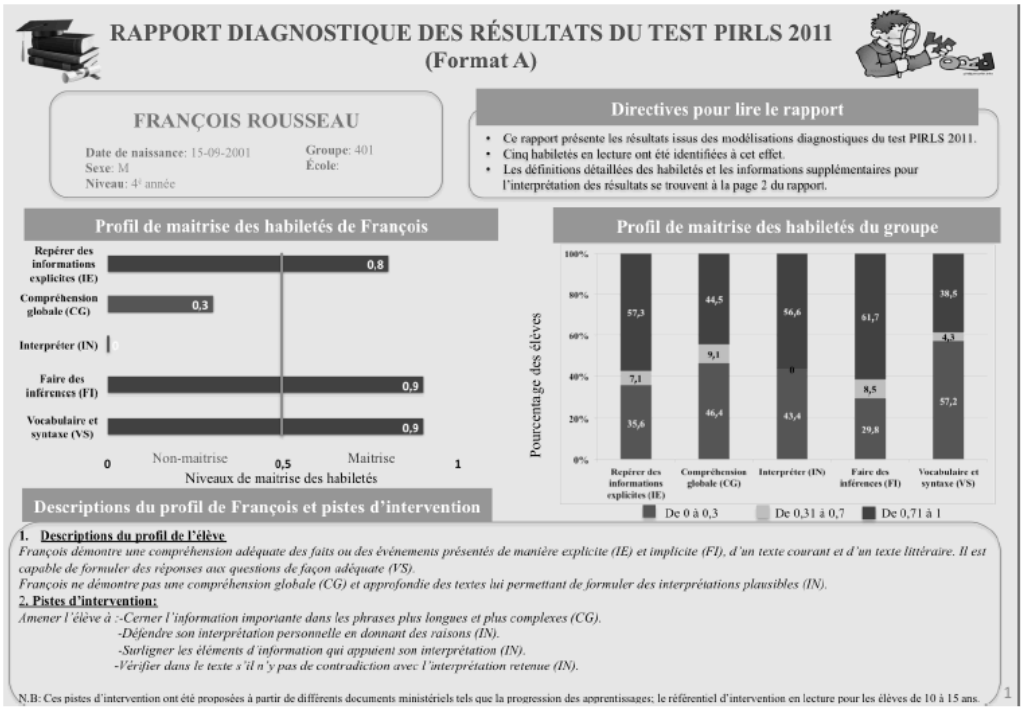
	Très en accord	En accord	En désaccord	Très en désaccord
1-Les objectifs du rapport diagnostique sont présentés avec clarté.	○	○	○	○
2-La description des habiletés facilite la compréhension du rapport.	○	○	○	○
3-Le mode d'emploi pour lire les graphiques est utile pour ma compréhension du rapport.	○	○	○	○
4-Les exemples utilisés dans cette partie facilitent mon interprétation des graphiques.	○	○	○	○

	Très en accord	En accord	En désaccord	Très en désaccord
5-La quantité des informations fournies est suffisante pour comprendre le profil de l'élève.	◯	◯	◯	◯
6-Dans l'ensemble, le vocabulaire utilisé est facile à comprendre.	◯	◯	◯	◯
7-Les paragraphes sont bien formulés.	◯	◯	◯	◯
8-Dans l'ensemble, le rapport présenté a un aspect visuel qui me plaît.	◯	◯	◯	◯
9-Ce rapport est plus détaillé que le rapport fait par un professionnel (par exemple, les orthopédagoques, les conseillers pédagogiques, etc.).	◯	◯	◯	◯
10-J'aimerais avoir accès à ce format de rapport pour mes élèves.	◯	◯	◯	◯
11-J'ai confiance dans mon interprétation du profil de mon élève.	◯	◯	◯	◯
12-Le profil de l'élève présenté dans ce rapport diagnostique correspond à celui de certains de mes élèves.	◯	◯	◯	◯

15. **Quelles sont vos suggestions pour améliorer ce format de rapport diagnostique?**

Section 3. Compréhension du rapport

* 16. **Veuillez choisir une réponse qui correspond le plus à votre compréhension du format A ci-dessous.**



Combien d'habiletés l'élève doit-il maitriser pour le test?

- ◯ a- Trois
- ◯ b- Quatre
- ◯ c- Cinq

* 17. **L'élève a t-il maitrisé toutes les habiletés attendues?**

- ◯ Oui
- ◯ Non

* 18. **L'élève a-t-il bien maitrisé l' habileté "*Repérer des informations explicites*" ?**

- ◯ Oui
- ◯ Non

* 19. **L'élève a-t-il bien maitrisé l'habileté "*Compréhension globale*"?**

- ◯ Oui
- ◯ Non

* 20. **Le niveau de maitrise de l'élève est-il identique pour les cinq habiletés?**

- ◯ Oui
- ◯ Non

* 21. **Dans le graphique, quel est le niveau de maitrise de l'élève pour l'habileté "*Interpréter*"?**

- ◯ 0
- ◯ 0,3
- ◯ 0,8
- ◯ 0,9
- ◯ 0,9

* 22. **Si on considère que le seuil de réussite est 0,5, lequel de ces énoncés est*vrai* ?**

- ◯ L'élève a maitrisé 1 sur 5 habiletés
- ◯ L'élève a maitrisé 2 sur 5 habiletés
- ◯ L'élève a maitrisé 3 sur 5 habiletés
- ◯ L'élève a maitrisé 4 sur 5 habiletés
- ◯ L'élève a maitrisé toutes les habiletés
- ◯ L'élève n'a maitrisé aucune des habiletés

* 23. **Dans le graphique sur *le profil de maitrise des habiletés du groupe* pour l'habileté "*Compréhension globale*", quel groupe a le pourcentage le plus élevé?**

- ◯ Groupe de 0 à 0.3
- ◯ Groupe de 0.31 à 0.7
- ◯ Groupe de 0.71 à 1

* 24. **Pour l'habileté "*faire des inférences*', le pourcentage des élèves ayant les niveaux de maitrise de *0.71 à 1* est-il le plus élevé?**

- ◯ Oui
- ◯ Non

Veuillez cliquer sur "Terminé" pour enregistrer vos réponses. Merci d'avoir répondu à notre sondage.

ANNEXE 7-APPROBATION ÉTHIQUE

Université de Montréal

Comité plurifacultaire d'éthique de la recherche

22 mars 2016

Madame Dan Thanh Duong Thi
Candidate au doctorat
Administration et fondements de l'éducation - Faculté des sciences de l'éducation

OBJET: Reconnaissance d'une approbation éthique

Mme Dan Thanh Duong Thi,

Le *Comité plurifacultaire d'éthique de la recherche (CPER)* a étudié le projet de recherche intitulé « Modélisation et élaboration des rapports à visée diagnostique des données du test PIRLS 2011. » et a délivré le certificat d'éthique demandé suite à la satisfaction des exigences précédemment émises.

Notez qu'il y apparaît une mention relative à un suivi annuel et que le certificat comporte une date de fin de validité. En effet, afin de répondre aux exigences éthiques en vigueur au Canada et à l'Université de Montréal, nous devons exercer un suivi annuel auprès des chercheurs et étudiants-chercheurs.

De manière à rendre ce processus le plus simple possible et afin d'en tirer pour tous le plus grand profit, nous avons élaboré un court questionnaire qui vous permettra à la fois de satisfaire aux exigences du suivi et de nous faire part de vos commentaires et de vos besoins en matière d'éthique en cours de recherche. Ce questionnaire de suivi devra être rempli annuellement jusqu'à la fin du projet et pourra nous être retourné par courriel. La validité de l'approbation éthique est conditionnelle à ce suivi. Sur réception du dernier rapport de suivi en fin de projet, votre dossier sera clos.

Il est entendu que cela ne modifie en rien l'obligation pour le chercheur, tel qu'indiqué sur le certificat d'éthique, de signaler au CPER tout incident grave dès qu'il survient ou de lui faire part de tout changement anticipé au protocole de recherche.

Nous vous prions d'agréer, Madame, l'expression de nos sentiments les meilleurs,

[redacted]

Tiiu Poldma, Présidente
Comité plurifacultaire d'éthique de la recherche (CPER)
Université de Montréal

TP/RS/rs

c.c. Gestion des certificats, BRDV
Nathalie Loye, Professeure agrégée, Administration et fondements de l'éducation - Faculté des sciences de l'éducation
Lucie Lefrançois
p.j. Certificat CPER-15-136-D

adresse postale
3744 Jean-Brillant, B-430-8
C.P. 6128, succ. Centre-ville
Montréal QC H3C 3J7
www.cper.umontreal.ca

Téléphone : 514-343-6111 poste 1896
cper@umontreal.ca

Université de Montréal

Comité plurifacultaire d'éthique de la recherche

N° de certificat
CPER-15-136-D

CERTIFICAT D'APPROBATION ÉTHIQUE

Le Comité plurifacultaire d'éthique de la recherche (CPER), selon les procédures en vigueur, en vertu des documents qui lui ont été fournis, a examiné le projet de recherche suivant et conclu qu'il respecte les règles d'éthique énoncées dans la Politique sur la recherche avec des êtres humains *de l'Université de Montréal.*

Projet	
Titre du projet	**Modélisation et élaboration des rapports à visée diagnostique des données du test PIRLS 2011.**
Étudiante requérant	**Dan Thanh Duong Thi** [redacted] Candidate au doctorat, Administration et fondements de l'éducation - Faculté des sciences de l'éducation Université de Montréal
Financement	
Organisme	CRSH
Programme	Subvention de développement de partenariat
Titre de l'octroi si différent	--
Numéro d'octroi	890-2011-0011
Chercheur principal	Nathalie Loye
No de compte	--

Approbation reconnue	
Approbation émise par	non
Certificat:	s.o.

MODALITÉS D'APPLICATION

Tout changement anticipé au protocole de recherche doit être communiqué au CPER qui en évaluera l'impact au chapitre de l'éthique.

Toute interruption prématurée du projet ou tout incident grave doit être immédiatement signalé au CPER.

Selon les règles universitaires en vigueur, un suivi annuel est minimalement exigé pour maintenir la validité de la présente approbation éthique, et ce, jusqu'à la fin du projet. Le questionnaire de suivi est disponible sur la page web du CPER.

[redacted]

Tiiu Poldma, Présidente
Comité plurifacultaire d'éthique de la recherche
Université de Montréal

22 mars 2016
Date de délivrance

1 avril 2017
Date de fin de validité

9

Université de Montréal

Comité plurifacultaire d'éthique de la recherche

28 Juin 2017

Madame Dan Thanh Duong Thi
Candidate au doctorat
Administration et fondements de l'éducation - Faculté des sciences de l'éducation

OBJET: Approbation éthique (renouvellement)

Mme Dan Thanh Duong Thi,

Le *Comité plurifacultaire d'éthique de la recherche (CPER)* a étudié votre demande de renouvellement pour le projet de recherche intitulé « Modélisation et élaboration des rapports à visée diagnostique des données du test PIRLS 2011. » et a délivré le certificat d'éthique demandé suite à la satisfaction des exigences qui prévalent. Vous trouverez ci-joint une copie numérisée de votre certificat; copie également envoyée à votre directeur/directrice de recherche et à la technicienne en gestion de dossiers étudiants (TGDE) de votre département.

Notez qu'il y apparaît une mention relative à un suivi annuel et que le certificat comporte une date de fin de validité. En effet, afin de répondre aux exigences éthiques en vigueur au Canada et à l'Université de Montréal, nous devons exercer un suivi annuel auprès des chercheurs et étudiants-chercheurs.

De manière à rendre ce processus le plus simple possible et afin d'en tirer pour tous le plus grand profit, nous avons élaboré un court questionnaire qui vous permettra à la fois de satisfaire aux exigences du suivi et de nous faire part de vos commentaires et de vos besoins en matière d'éthique en cours de recherche. Ce questionnaire de suivi devra être rempli annuellement jusqu'à la fin du projet et pourra nous être retourné par courriel. La validité de l'approbation éthique est conditionnelle à ce suivi. Sur réception du dernier rapport de suivi en fin de projet, votre dossier sera clos.

Il est entendu que cela ne modifie en rien l'obligation pour le chercheur, tel qu'indiqué sur le certificat d'éthique, de signaler au CPER tout incident grave dès qu'il survient ou de lui faire part de tout changement anticipé au protocole de recherche.

Nous vous prions d'agréer, Madame, l'expression de nos sentiments les meilleurs,

[redacted]

Raphaëlle Stenne, conseillère en éthique de la recherche
Comité plurifacultaire en éthique de la recherche (CPER)
Université de Montréal
JP/RS/rs
c.c. Gestion des certificats, BRDV, Nathalie Loye, Professeure agrégée, Administration et fondements de l'éducation - Faculté des sciences de l'éducation Lucie Lefrançois
p.j. Certificat CPER-15-136-D(1)

adresse postale
3744 Jean-Brillant. B-430-8
C.P. 6128, succ. Centre-ville
Montréal QC H3C 3J7
www.cper.umontreal.ca

Téléphone : 514-343-6111 poste 1896
cper@umontreal.ca

Université de Montréal

Comité plurifacultaire d'éthique de la recherche

N° de certificat
CPER-15-136-D(1)

CERTIFICAT D'APPROBATION ÉTHIQUE

- 1er renouvellement -

Le Comité plurifacultaire d'éthique de la recherche (CPER), selon les procédures en vigueur et en vertu des documents relatifs au suivi qui lui a été fournis conclu qu'il respecte les règles d'éthique énoncées dans la Politique sur la recherche avec des êtres humains *de l'Université de Montréal*

Projet	
Titre du projet	**Modélisation et élaboration des rapports à visée diagnostique des données du test PIRLS 2011.** Modifications apportées lors du suivi fait en 2017: les étudiants de 3e et 4e année n'ont pas été visés par la recherche et des orthopédagogues ainsi que des conseillers pédagogiques ont participé au projet de recherche.
Étudiante requérant	**Dan Thanh Duong Thi** [redacted] Candidate au doctorat, Administration et fondements de l'éducation - Faculté des sciences de l'éducation Université de Montréal
Financement	
Organisme	CRSH
Programme	Subvention de développement de partenariat
Titre de l'octroi si différent	--
Numéro d'octroi	890-2011-0011
Chercheur principal	Nathalie Loye
No de compte	--

MODALITÉS D'APPLICATION

Tout changement anticipé au protocole de recherche doit être communiqué au CPER qui en évaluera l'impact au chapitre de l'éthique. Toute interruption prématurée du projet ou tout incident grave doit être immédiatement signalé au CPER.

Selon les règles universitaires en vigueur, un suivi annuel est minimalement exigé pour maintenir la validité de la présente approbation éthique, et ce, jusqu'à la fin du projet. Le questionnaire de suivi est disponible sur la page web du CPER.

[redacted]

Raphaëlle Stenne, conseillère en éthique de la recherche
Comité plurifacultaire d'éthique de la recherche
Université de Montréal

28 juin 2017 Date de délivrance du renouvellement ou de la réémission*	**1er juillet 2018** Date du prochain suivi
22 mars 2016 Date du certificat initial	**1er juillet 2018** Date de fin de validité

*Le présent renouvellement est en continuité avec le précédent certificat

adresse postale
3744 Jean-Brillant. B-430-8
C.P. 6128, succ. Centre-ville
Montréal QC H3C 3J7
www.cper.umontreal.ca

Téléphone : 514-343-6111 poste 1896
cper@umontreal.ca