

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

APPROCHE NOVATRICE DE VALIDATION DE MESURE DE
COMPRÉHENSION EN LECTURE

MÉMOIRE
PRÉSENTÉ
COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN ÉDUCATION

PAR
AUDREY WAGENER

AVRIL 2021

UNIVERSITÉ DU QUÉBEC À MONTRÉAL
Service des bibliothèques

Avertissement

La diffusion de ce mémoire se fait dans le respect des droits de son auteur, qui a signé le formulaire *Autorisation de reproduire et de diffuser un travail de recherche de cycles supérieurs* (SDU-522 – Rév.10-2015). Cette autorisation stipule que «conformément à l'article 11 du Règlement no 8 des études de cycles supérieurs, [l'auteur] concède à l'Université du Québec à Montréal une licence non exclusive d'utilisation et de publication de la totalité ou d'une partie importante de [son] travail de recherche pour des fins pédagogiques et non commerciales. Plus précisément, [l'auteur] autorise l'Université du Québec à Montréal à reproduire, diffuser, prêter, distribuer ou vendre des copies de [son] travail de recherche à des fins non commerciales sur quelque support que ce soit, y compris l'Internet. Cette licence et cette autorisation n'entraînent pas une renonciation de [la] part [de l'auteur] à [ses] droits moraux ni à [ses] droits de propriété intellectuelle. Sauf entente contraire, [l'auteur] conserve la liberté de diffuser et de commercialiser ou non ce travail dont [il] possède un exemplaire.»

REMERCIEMENTS

Je tiens à exprimer mes plus sincères remerciements à mon directeur, M. Éric Dion, sans qui ce mémoire n'aurait pu être aussi abouti. Sa rigueur scientifique, sa vivacité intellectuelle, ses talents de vulgarisateur ainsi que ses nombreux travaux en compréhension de la lecture m'ont grandement inspirée et m'ont amenée à donner le meilleur de moi-même. Je suis très reconnaissante de tous les échanges que nous avons eus et de toutes les opportunités scientifiques et professionnelles qu'il m'a permis de vivre.

Je remercie également Mme Elizabeth Olivier, chercheure post-doctorale au Laboratoire de recherche sur la synergie substantive et méthodologique, Département de psychologie, Université Concordia, pour sa précieuse aide dans la réalisation des analyses factorielles confirmatoires.

Je souhaite remercier, pour leur généreuse coopération, les sept enseignantes d'expérience qui ont gentiment répondu à mon appel.

Également, je tiens à remercier la Fondation de l'UQÀM, la Dre Véronneau-Troutman, la Fondation de la Fédération des écoles normales du Québec en sciences de l'éducation, la Fondation J.-A DeSève et la RBC Banque Royale pour l'octroi de bourses qui m'ont permis d'alléger mon quotidien étudiant et m'ont encouragée à persévérer.

Finalement, ce mémoire n'aurait jamais pu exister sans le soutien indéfectible de mon conjoint, merci !

DÉDICACE

À mes parents, à mon conjoint et à
tous ceux qui ont participé de près ou
de loin à la réalisation de ce travail,
ma plus profonde reconnaissance pour
votre soutien et vos encouragements.

AVANT-PROPOS

« À moins d’avoir confiance en la fiabilité de la règle, si vous utilisez l’instrument pour mesurer une table, vous risquez de plutôt utiliser la table pour mesurer la règle » (propos de Wittgenstein cités par Taleb, 2005).

La règle de Wittgenstein nous amène à la raison de ce mémoire. Évalue-t-on ce que l’on prétend évaluer ? En d’autres termes, évaluons-nous ce qui est le plus pertinent ? Et comment détermine-t-on cette pertinence ? C’est à partir de ces questionnements que la présente recherche a pris forme. Elle s’inscrit dans un processus réflexif d’une équipe de recherche dirigée par mon directeur, Éric Dion, dans le cadre d’une étude CRSH menée dans 34 classes primaires en milieu défavorisé.

Cette équipe de recherche s’est notamment questionnée sur les meilleures façons d’enseigner la compréhension des textes informatifs aux jeunes lecteurs de 3^e année. Après avoir rédigé 120 courts textes retraçant la vie d’explorateur de Samuel de Champlain, elle a mis en place 3 conditions dans lesquelles les élèves devaient lire ces textes en tutorat par les pairs (Wagener et Dion, 2018). Lors du prétest, posttest et des évaluations en cours d’intervention, la compréhension a été évaluée à l’aide d’un rappel de texte. La problématique de la validité de ce test de compréhension a été soulevée et a mené à l’élaboration de ce présent projet. Notre méthodologie reprend les données collectées lors du prétest qui a eu lieu en décembre 2018.

TABLES DES MATIÈRES

AVANT-PROPOS	iv
LISTE DES FIGURES.....	vii
LISTE DES TABLEAUX.....	viii
RÉSUMÉ	ix
INTRODUCTION	1
CHAPITRE I PROBLÉMATIQUE	2
1.1 Le défi de la compréhension en lecture	2
1.2 Définitions théoriques de la compréhension en lecture	3
1.3 L'importance et les défis de l'évaluation de la compréhension	5
CHAPITRE II CADRE CONCEPTUEL	11
2.1 Concepts psychométriques de fidélité et de validité.....	11
2.2 La validité de critère	13
2.3 La validité de construit	13
2.4 La validité de contenu et d'apparence	14
2.5 La validité sociale	15
2.6 La validité d'utilisateur.....	16
2.7 Hypothèses.....	21
CHAPITRE III MÉTHODOLOGIE	23
3.1 Participants	23
3.2 Instruments	24
3.3 Procédure pour l'évaluation des élèves	29

3.4	Analyses.....	30
CHAPITRE IV RÉSULTATS		31
4.1	Analyses préliminaires.....	31
4.2	Dimensionnalité et pondération des items conservés	37
4.3	Statistiques descriptives sur les impressions des enseignantes	41
4.4	Lien entre la fluidité, la cotation du rappel et l'impression des enseignantes ...	43
4.5	Calcul des scores de compréhension	44
CHAPITRE V DISCUSSION.....		47
ANNEXE A <i>LES DANGERS DE L'OCÉAN</i> , VERSION ORIGINALE.....		56
ANNEXE B <i>LES DANGERS DE L'OCÉAN</i> , VERSION ADAPTÉE.....		58
ANNEXE C <i>L'HIVER</i>		59
ANNEXE D ANALYSE EN PROPOSITIONS DES <i>DANGERS DE L'OCÉAN</i>		60
ANNEXE E ANALYSE EN PROPOSITIONS DE <i>L'HIVER</i>		61
ANNEXE F FORMULAIRE DE CONSENTEMENT DES ENSEIGNANTS EXPERTS		62
APPENDICE A LETTRE DE CONSENTEMENT PARENTAL		65
RÉFÉRENCES.....		69

LISTE DES FIGURES

Figure	Page
4.1 Modèle de mesure finale pour <i>Les dangers de l'océan</i>	39
4.2 Modèle de mesure finale pour <i>L'hiver</i>	41
4.3 Lien entre la fluidité, le score au rappel et l'impression moyenne des enseignantes pour <i>Les dangers de l'océan</i>	44
4.4 Lien entre la fluidité, le score au rappel et l'impression moyenne des enseignantes pour <i>L'hiver</i>	44
4.5 Chiffrier permettant de calculer le score au rappel du texte <i>Les dangers de l'océan</i>	45

LISTE DES TABLEAUX

Tableau	Page
3.1 Grille de correction pour <i>Les dangers de l'océan</i>	27
3.2 Grille de correction pour <i>L'hiver</i>	28
4.1 Niveau de difficulté (et erreur standard) des items du rappel du texte <i>Les Dangers de l'océan</i>	33
4.2 Corrélations tétrachoriques pour <i>Les Dangers de l'océan</i>	34
4.3 Niveau de difficulté (et erreur standard) des items du rappel du texte <i>L'hiver</i>	35
4.4 Corrélations tétrachoriques pour <i>L'hiver</i>	36
4.5 Impression moyenne des quatre enseignantes sur la qualité du rappel des deux textes Grille de correction pour <i>Les dangers de l'océan</i>	42

RÉSUMÉ

Ce mémoire examine une nouvelle procédure de validation des mesures de compréhension en lecture. Nous avons mis à l'essai une procédure novatrice dite de validité d'utilisateur en examinant les rappels de deux textes informatifs complétés par 306 élèves de 3^e année de milieux défavorisés. Notre objectif était d'évaluer dans quelle mesure un score basé sur une cotation formelle du rappel reflète adéquatement la compétence d'un élève selon les normes en vigueur dans le milieu scolaire. Pour ce faire, nous avons élaboré des grilles pour coter le rappel et, de manière parallèle, nous avons demandé à quatre enseignantes de donner leurs impressions, à l'aveugle et de manière indépendante, sur le quart de ces rappels en utilisant une échelle de cotation globale. L'analyse factorielle confirmatoire a été utilisée pour dériver des scores de notre cotation formelle et mettre ces scores en relation avec les impressions des enseignantes. Les résultats indiquent que les enseignantes ont toutes exprimé un point de vue similaire sur la qualité des rappels et que leurs points de vue correspondaient étroitement à nos scores, appuyant ainsi la validité d'utilisateur de ces derniers. De plus, comme attendu selon une théorie généralement acceptée en lecture (Gough et Tunmer, 1986; Hoover et Gough, 2018), les élèves avec une bonne fluidité ont obtenu en général des scores plus élevés, un argument supplémentaire lié à la validité de construit. Nous examinons, en discussion, le rôle de la validité d'utilisateur dans l'élaboration d'évaluations appropriées pour la recherche-intervention ou pour la pratique normale.

Mots clés : compréhension, lecture, validation, textes informatifs

INTRODUCTION

La compréhension en lecture est un phénomène complexe et relativement difficile à cerner. Selon les théories qui ont examiné en détails la nature de ce phénomène, le rappel est possiblement la meilleure façon de l'évaluer. Cependant, malgré la transparence de cette façon de procéder, rien ne garantit qu'elle génère des scores valides (c.-à-d. qui représentent vraiment le niveau de compréhension), notamment parce qu'il est relativement difficile de scorer le rappel.

Afin de contourner ce problème, nous proposons de mettre à l'essai une procédure novatrice de validation de tests.

CHAPITRE I

PROBLÉMATIQUE

1.1 Le défi de la compréhension en lecture

La compréhension en lecture, un processus d'extraction et d'élaboration de sens (RAND, 2002), est aussi importante que difficile pour plusieurs élèves du primaire. Afin de comprendre, ces derniers doivent être en mesure de lire correctement le texte à un débit raisonnable (Hosp et Fuchs, 2005), mais aussi d'identifier les idées présentées dans le texte et de les mettre en lien, ce qui implique de saisir la signification des mots et de percevoir la logique dans l'enchaînement des mots et des phrases (Irwin, 2007). Pour ajouter à la difficulté de la tâche, même les textes les plus simples contiennent plusieurs sous-entendus. En fait, plusieurs informations ou liens essentiels à la compréhension ne sont pas explicités et le lecteur doit les inférer, notamment en faisant appel à ses propres connaissances (Hirsch, 2003 ; Kintsch, 1986). Le lecteur du 1^{er} ou du 2^e cycle du primaire qui ne parvient pas à réaliser toutes ces opérations est susceptible d'être peu motivé à s'investir dans une activité qui est pour lui dépourvue de sens (ex. : Morgan *et al.*, 2008).

Pour ajouter aux défis de la compréhension, tous les textes ne sont pas organisés ou structurés de la même façon et ne présentent pas tous le même genre d'information. Il existe en effet plusieurs types de textes qui comportent chacun leurs défis particuliers. Le texte narratif est certainement le type de texte avec lequel les lecteurs débutants sont

le plus familier. Il s'agit d'un texte qui présente une séquence d'événements en respectant une structure convenue (ex. : situation initiale, éléments déclencheurs; Mandler et Johnson, 1977). Typiquement, le texte narratif décrit un ou plusieurs personnages fictifs qui évoluent dans un temps et un lieu donné, par exemple dans la série Harry Potter (ex. : Rowling, 2016). Le texte argumentatif décrit, quant à lui, un point de vue sur un sujet donné, alors que le texte poétique mise sur la sonorité et la polysémie des mots (ex. : Nelligan, 1998). Quant au texte informatif, il se distingue par le fait qu'il présente un contenu non fictionnel dont le but premier est de transmettre des informations organisées et hiérarchisées sur un phénomène ou un événement, le plus souvent en utilisant un vocabulaire précis et relativement avancé (Yopp et Yopp, 2012). Le lecteur du primaire peut, par exemple, rencontrer ce dernier type de texte dans les encyclopédies en ligne (ex. : Vikidia) et dans les magazines jeunesse (ex. : Les Débrouillards). L'organisation du texte informatif peut prendre différentes formes qui ne respectent pas celle, plus familière, du texte narratif (Meyer et Ray, 2011; Williams *et al.*, 2014). En particulier, certains textes informatifs sont organisés sous forme de narration (ex. : un texte décrivant les actions d'un personnage historique), alors que d'autres non (ex. : un texte décrivant les causes des tornades; voir Meyer et Ray, 2011).

1.2 Définitions théoriques de la compréhension en lecture

Que signifie comprendre un texte, de type informatif ou autre? Parmi les théories à ce sujet, celle proposée par Kintsch (ex. : Kintsch et Kintsch, 2005, Chap. 3) est possiblement celle qui retient le plus l'attention présentement (ex. : Language and Reading Research Consortium, 2016). Selon cette théorie, un lecteur comprend en s'élaborant un modèle de situation, c'est-à-dire une représentation mentale du phénomène, des notions ou de la suite d'événements décrits dans le texte (voir aussi Kintsch et van Dijk, 1978). Plutôt que de mémoriser entièrement le texte, le lecteur doit se créer une représentation schématique du contenu de ce dernier en fonction, notamment, de ses capacités de synthèse et de ses connaissances sur le sujet du texte

(pour un point de vue similaire, voir Graesser *et al.*, 1994). Pour ce faire, il doit notamment, selon Kintsch et Kintsch, identifier les relations de cause à effet suggérées par le texte et aller au-delà du contenu présenté explicitement en faisant des inférences à l'aide de ses connaissances antérieures. Lorsque la compréhension est adéquate, le lecteur s'est représenté l'information de manière organisée en l'intégrant à ses connaissances antérieures, ce qui fait en sorte qu'il est en mesure d'utiliser plus tard cette information (ex. : pour discuter d'un sujet ou pour résoudre un problème). Pour les lecteurs experts familiers avec le sujet du texte (ex. : les étudiants de niveau collégial), le modèle de situation (la représentation) s'élabore automatiquement. Pour le lecteur du primaire, la tâche peut être plus difficile, en particulier lorsque le sujet est nouveau.

Pour élaborer leur théorie, Kintsch et Kintsch (2005) ont surtout étudié des lecteurs avancés, c'est-à-dire des lecteurs avec une vaste expérience de l'écrit qui lisent de manière fluide et qui possèdent un vocabulaire et des connaissances antérieures étendus (ex. : Rawson et Kintsch, 2005). van den Broek et collègues se sont intéressés plus spécifiquement au lecteur du primaire (voir van den Broek et Kendeou, 2017). Au même titre que Kintsch et Kintsch, ces derniers chercheurs proposent que la compréhension implique l'élaboration d'une représentation mentale organisée du texte. Ils considèrent néanmoins que la représentation du lecteur du primaire est souvent incomplète et inexacte, en plus de contenir des intrusions créées par l'activation de connaissances antérieures non pertinentes. Williams (1993) a aussi remarqué que plusieurs lecteurs en difficulté n'arrivaient pas à comprendre, ne serait-ce que minimalement les textes, apparemment parce que leur lecture hésitante et fragmentaire des mots active des connaissances antérieures ou des points de vue personnels sans liens avec le sujet du texte (voir aussi Williams, 1991). Pour des chercheurs comme van den Broek, Williams et leurs collègues, l'évaluation de la compréhension du lecteur du primaire doit établir dans quelle mesure sa représentation est élaborée (sans

nécessairement être complète) et près du contenu du texte, c'est-à-dire exempte d'intrusions.

1.3 L'importance et les défis de l'évaluation de la compréhension

S'il n'est pas simple pour le lecteur du primaire de comprendre un texte, il n'est pas simple non plus d'évaluer son degré de compréhension puisque sa représentation mentale du contenu du texte n'est pas directement observable (Fletcher, 2006). L'évaluateur doit donc inférer le degré de compréhension atteint par le lecteur du primaire (RAND, 2002). Cette évaluation peut être réalisée, par exemple, en questionnant directement le lecteur sur les relations de cause à effet suggérées par le texte (ex. : Williams *et al.*, 2007) ou en utilisant le rappel, c'est-à-dire en lui demandant, de manière plus ouverte, de décrire le contenu du texte (ex. : Linderholm et van den Broek, 2002). Fréquemment utilisée en recherche (ex. : van den Broek *et al.*, 2011), cette dernière forme d'évaluation permet, en principe (ex. : en l'absence de difficultés d'expression), d'accéder à la représentation mentale que s'est créée le lecteur, c'est-à-dire ce qui est considéré comme le produit final de la compréhension. Comme le soulignent entre autres Morrow (1988) et Klingner (2004), en plus de permettre d'évaluer le degré de compréhension, un examen du rappel offert par le lecteur pourrait permettre d'identifier certains problèmes de compréhension, notamment une difficulté à inférer les liens logiques entre différents éléments du texte ou à organiser ces éléments en un tout cohérent. À noter que si le rappel génère une information apparemment riche sur la compréhension du lecteur, ce type d'évaluation est long et difficile à corriger de manière raisonnablement objective et uniforme (Reed et Vaughn, 2012), ce qui explique probablement pourquoi il est rarement utilisé dans des contextes où le temps de correction est une considération importante, par exemple en classe (Morrow, 1988).

Bien qu'elle puisse être difficile, l'évaluation de la compréhension est néanmoins importante d'un point de vue éducatif. Elle permet, entre autres, d'établir les besoins

des élèves, d'évaluer l'efficacité d'une méthode d'enseignement (ex. : Dion, Roux, Landry *et al.*, 2011) et de sanctionner les études (Desrochers et Saint-Aubin, 2008). Si les problèmes de compréhension ne sont pas identifiés raisonnablement tôt, c'est tout le cheminement scolaire de l'élève qui est susceptible d'être affecté. L'élève est effectivement rapidement appelé à lire (et à comprendre) pour apprendre de nouvelles notions, non seulement en français, mais également, par exemple, en univers social et en mathématiques (ex. : Fuchs, Gilbert *et al.*, 2018). De manière plus générale, la compréhension est la finalité de la lecture (Perfetti *et al.*, 2005) et l'évaluation doit notamment refléter l'atteinte de cet objectif. En un sens, il ne peut y avoir une lecture véritable en l'absence de compréhension.

Le test standardisé est une solution souvent proposée pour évaluer la compréhension en lecture (Pearson et Hamm, 2005, Chap. 2). Il s'agit d'une évaluation administrée et corrigée de manière uniforme et dont le score est interprété en fonction d'une norme établie à partir d'un échantillon représentatif de la population d'intérêt. L'évaluation est souvent réalisée en individuel, par un évaluateur qualifié. À titre d'exemple, le *Test de rendement individuel de Weschler* (WIAT-II; Weschler, 2008) présente au lecteur débutant ou avancé de courts passages de deux à trois phrases devant être lus en silence. Une fois la lecture complétée, le lecteur doit répondre à des questions ouvertes portant sur chaque passage. Certaines questions portent sur une information présente de manière littérale dans le texte alors que d'autres requièrent de faire une inférence. Dans tous les cas, une réponse courte est attendue (Fletcher, 2006). L'évaluation est interrompue lorsque le lecteur évalué répond incorrectement à un nombre déterminé de questions. L'évaluateur corrige les réponses en utilisant des critères précis et le score, qui peut être exprimé sous différentes formes, est interprété en fonction d'une norme qui permet notamment de déterminer le niveau scolaire atteint en compréhension par le lecteur évalué (pour plus de détails sur la correction, voir Weschler, 2008).

Le test standardisé présente des avantages évidents. Tout d'abord, la subjectivité de l'évaluateur (ex. : en ce qui concerne le comportement de l'élève en classe) n'intervient pas, en principe, ce qui fait en sorte que le score à ce type de test est équitable dans la mesure où il dépend uniquement de la performance lors de l'évaluation (voir cependant Fuchs et Fuchs, 1986; Johnston, 1984; Solórzano, 2008). La correction est simple, étant donné la brièveté des réponses attendues et la précision des critères. De plus, la standardisation de la passation et de la correction implique que les scores des lecteurs évalués ont une signification uniforme et que ces scores peuvent facilement être comparés entre eux, notamment pour situer les lecteurs par rapport à un groupe de référence (c.-à-d. des élèves du même âge ou du même niveau). Finalement, le recours à des tests standardisés (c.-à-d. à une métrique commune) pour évaluer l'efficacité de différentes méthodes d'enseignement simplifie la comparaison de ces méthodes dans le cadre de recensions méta-analytiques (Fuchs, Gilbert *et al.*, 2018).

Les tests standardisés ne sont cependant pas sans limites. Comme le souligne Froese-Germain (1999), ces tests ne sont pas élaborés par les enseignants des élèves qui sont évalués. Leur contenu ne reflète donc pas nécessairement ce qui a été enseigné à ces derniers ou les attentes spécifiques en termes d'apprentissage. De plus, les normes établies pour un test standardisé sont inévitablement influencées par la composition de l'échantillon à partir duquel elles ont été établies. De telles normes pourraient, par conséquent, se révéler inéquitables pour certains élèves, par exemple ceux issus de minorités, qui auraient été sous-représentés dans l'échantillon normatif (Simon, 2011, Chap. 10). En outre, les scores aux tests standardisés sont parfois considérés comme interchangeables, ce qui n'est pas nécessairement le cas, même si ces tests mesurent nominalement tous le même construit, c'est-à-dire la compréhension. Keenan et collègues (2008) ont comparé les scores de lecteurs évalués à l'aide de quatre tests de mesure de la compréhension en lecture parmi les plus utilisés en recherche et en milieu scolaire aux États-Unis. Les chercheurs ont observé qu'un lecteur pouvait obtenir des scores de compréhension substantiellement différents d'un test à l'autre (voir aussi

Fletcher, 2006). Il importe ainsi de garder en tête que le score de compréhension obtenu par un lecteur est toujours fonction, au moins en partie, du test utilisé pour l'évaluer, même lorsqu'il s'agit d'un test standardisé. À titre d'exemple, Keenan et collègues (2008) ont remarqué que les élèves répondent parfois correctement à un test standardisé en devinant les réponses en fonction de la formulation des questions, sans faire preuve de compréhension véritable et, vraisemblablement, sans s'élaborer un modèle du contenu du texte. Le problème découle en partie de la brièveté des réponses demandées, notamment en comparaison avec les tests qui utilisent le rappel.

Toujours en termes de limites, le recours aux tests standardisés comporte un risque de réification, c'est-à-dire de transposition rigide d'un concept abstrait (ex. : l'intelligence) en un fait observable particulier (ex. : le score à un test de quotient intellectuel; pour une discussion, voir Gould, 1981). En d'autres termes, l'utilisateur de test ne doit pas associer trop étroitement « la compréhension » au score à un test standardisé. Le risque pourrait découler, en partie, du recours à des normes. Affirmer, par exemple, qu'un élève démontre une compréhension de niveau « fin 3^e année du primaire » sur la base de la performance à un test standardisé peut donner à l'affirmation une impression de caractère définitif qui n'est pas entièrement justifiée. Il n'a pas été établi, en fait, que les tests standardisés représentent la meilleure façon d'évaluer la compréhension (Fuchs *et al.*, 2001; Pearson et Hamm, 2005, Chap. 2). D'ici à ce qu'une telle démonstration soit réalisée, il est pertinent de considérer des formes alternatives d'évaluation de la compréhension, notamment sur des considérations plus théoriques et moins logistiques (c.-à-d. la facilité de passation et de correction).

De plus, indépendamment des avantages et des limites, recourir aux tests standardisés pour évaluer la compréhension n'est pas vraiment une option pour le milieu scolaire québécois francophone. En général, très peu de tests d'habiletés scolaires ont été normés en recourant à des échantillons québécois (Cormier *et al.*, 2006). Établir et maintenir à jour des normes sont des opérations dispendieuses, ce qui explique

probablement pourquoi plusieurs maisons d'édition préfèrent simplement traduire et adapter des tests conçus pour des populations anglophones (pour une recension, voir Berger et Desrochers, 2011). Dans sa recension des tests publiés entre 1980 et 2014, Monetta (2015) a repéré seulement trois tests standardisés évaluant la compréhension en lecture auprès des élèves québécois : le *Test de rendement pour francophones* (Sarrazin, 1995), *La forme noire* (Maeder, 2010) et le WIAT-II (Weschler, 2008). Le *Test de rendement pour les francophones* permet de mesurer la compréhension de texte, mais il est probablement trop compliqué pour plusieurs élèves du primaire (Cormier *et al.*, 2006). Dans un même ordre d'idées, *La forme noire* inclut un test de compréhension, mais ce dernier est approprié seulement pour les élèves de 9 à 12 ans. Quant au WIAT-II, des normes québécoises existent pour une clientèle de 6 à 17 ans. Ce dernier test représente un outil d'évaluation diagnostique généraliste qui constitue une bonne porte d'entrée pour évaluer les capacités d'un élève. Cependant, malgré une passation relativement aisée, son coût élevé et une correction chronophage limitent grandement son utilisation dans les écoles québécoises.

Étant donné l'absence ou l'accès limité aux tests standardisés, les praticiens québécois utilisent des outils maison ou des extraits de matériel scolaire qui génèrent des scores dont la précision et la pertinence sont indéterminées (Simon, 2011, Chap. 10). Une telle situation est potentiellement problématique puisque rien ne garantit qu'un test qui évalue en principe la compréhension le fasse correctement. La compréhension est un construit plus complexe et difficile à cerner, par exemple, que la conscience phonologique ou le décodage (van den Broek *et al.*, 2009) et les directives proposées aux enseignants en matière d'évaluation sont vagues. Le ministère de l'Éducation et de l'Enseignement supérieur (MEES) recommande essentiellement d'évaluer la compréhension des éléments « significatifs », « explicites et implicites », « pertinents d'un texte » (ministère de l'Éducation et de l'Enseignement supérieur, 2011). Les concepts issus de la psychométrie pourraient être utiles pour orienter l'évaluation de la compréhension.

L'objectif général de ce mémoire est d'explorer une approche novatrice pour examiner la pertinence (validité) de tests de compréhension en lecture. En lien avec les modèles théoriques courants, cette approche sera appliquée à des tests évaluant l'élaboration du modèle mental par le biais du rappel. Finalement, étant donné la rareté relative des travaux portant spécifiquement sur la compréhension des textes informatifs par les lecteurs du primaire, en particulier avant la 3^e année, nous avons décidé de considérer l'évaluation de la compréhension de ce style littéraire.

CHAPITRE II

CADRE CONCEPTUEL

Nous proposons une procédure novatrice pour examiner la validité (pertinence) des tests de compréhension en lecture (pour une définition du concept, voir la section 1.2 de la problématique). Cette procédure s'appuie sur une réinterprétation des concepts issus de la psychométrie. Afin de justifier notre procédure, nous procédons ici à un examen détaillé de ces concepts, en particulier ceux liés à la validité.

2.1 Concepts psychométriques de fidélité et de validité

La psychométrie est le champ de recherche s'intéressant à la théorie et à la pratique de la mesure des caractéristiques humaines, incluant les aptitudes cognitives et l'apprentissage (Crocker et Algina, 2008). Comme ces caractéristiques ne sont pas nécessairement observables directement, il est important de bien les cerner sur le plan conceptuel (Desrochers *et al.*, 2011, Chap. 2). Ce travail de réflexion n'est toutefois pas suffisant. En fait, la psychométrie repose sur l'idée que le score observé (la performance) ne correspond pas forcément au score réel (la compétence; Laurier *et al.*, 2005). Il est ainsi important de démontrer que le score observé, et par extension le test, présente certaines caractéristiques, en particulier la fidélité et la validité. En termes généraux, il faut montrer, données à l'appui, que le test génère un score précis et pertinent.

Par définition, plus un score est fidèle ou précis, plus sa marge d'erreur est modeste, autrement dit, plus le score observé est près du score réel (Carpenter et Paris, 2005, Chap. 12). Encore une fois, comme le score réel ne peut être observé, la fidélité du score observé doit être inférée. La fidélité est souvent opérationnalisée en termes de consistance ou de fidélité test-retest (AERA, APA et NCME, 2014). Plus spécifiquement, un test sera considéré comme fidèle si les élèves obtiennent des scores comparables lorsqu'ils complètent deux versions équivalentes de ce test dans un court laps de temps (ex. : Tolar et *al.*, 2012). L'écart entre les scores aux deux passations correspond alors à la marge d'erreur (exprimée sous forme d'un coefficient de corrélation).

Il est important de noter qu'un test peut être précis ou fidèle sans être pertinent. Pour prendre un exemple à caractère historique, le fondateur de la psychométrie, Francis Galton (1822-1911), a tenté d'utiliser le temps de réaction lors de l'exécution d'une tâche pour évaluer la qualité du fonctionnement neuronal des personnes, ce qui serait appelé aujourd'hui l'intelligence (voir Galton, 1890). Cependant, cette mesure plutôt simple, même si elle est probablement fidèle (ou précise), n'est apparemment pas pertinente au sens où elle n'est pas vraiment en lien avec l'intelligence. Il n'est donc pas approprié d'utiliser un score reflétant fidèlement le temps de réaction observable pour inférer l'intelligence de la personne (Crocker et Algina, 2008; Laveault et Grégoire, 2014).

En plus d'être fidèle, un score doit donc être pertinent ou valide, c'est-à-dire mesurer ce qu'il prétend mesurer (Crocker et Algina, 2008). Selon la terminologie présentement en cours, le score doit permettre de faire des inférences valides (AERA, APA et NCME, 2014). Par exemple, le score (observé) à un test réputé mesurer les capacités d'apprentissage est valide seulement si ce score permet d'inférer correctement les réelles capacités d'apprentissage de l'élève. Établir la validité du score à un test est un

processus complexe dans lequel une variété d'arguments, ou de preuves, peut être invoquée.

2.2 La validité de critère

Établir la validité de critère représente un des arguments les plus évidents en ce qui concerne la pertinence d'un score. De ce point de vue, un score observé est considéré valide s'il est étroitement associé à un critère ou, autrement dit, à un autre score dont la validité est elle-même très bien établie ou transparente. À titre d'exemple, le poids d'un enfant (en kilogrammes) est parfois évalué en demandant aux parents cette information (Statistique Canada, 2009). Dans ce cas, il s'agirait de démontrer l'existence d'une correspondance étroite entre l'estimation des parents (score observé) et le poids de l'enfant tel qu'évalué à l'aide d'un pèse-personne précis (un critère correspondant étroitement à la réalité). Cet exemple illustre l'importance d'examiner la validité puisqu'il s'est avéré que les parents surestiment substantiellement, en moyenne, le poids des enfants (Shields *et al.*, 2008). Tel que mentionné précédemment, les nombreux travaux sur la compréhension en lecture n'ont pas permis d'isoler un critère reflétant de manière claire et transparente ce construit, probablement en raison de la complexité de ce dernier (ex. : Pearson et Hamm, 2005, Chap. 2; Kintsch et Kintsch, 2005, Chap. 3). Par conséquent, la validité des mesures de compréhension ne peut être établie en invoquant des arguments de validité de critère (voir aussi Fuchs *et al.*, 2001).

2.3 La validité de construit

En l'absence de critère, il est possible d'invoquer d'autres types d'arguments, notamment ceux relevant de la validité de construit (Cronbach et Meehl, 1955, Laurier *et al.*, 2005). Selon ce dernier type d'argument, un score est valide lorsqu'il est conforme à un modèle théorique, en particulier lorsqu'il entretient avec d'autres scores

les relations dictées par la théorie. Dans le cas de la compréhension, le modèle de Gough et Tunmer (1986) stipule notamment l'existence d'un lien entre la fluidité et la compréhension; une lecture raisonnablement rapide et exacte (c.-à-d. fluide; Fuchs *et al.*, 2001) étant nécessaire pour prendre connaissance des idées ou informations présentées dans le texte (Language and Reading Research Consortium, 2015). En ce sens, un score de compréhension valide devrait, entre autres, être corrélé avec la fluidité (ex. : Hosp et Fuchs, 2005). La difficulté réside ici dans le fait que les arguments de validité de construit sont convaincants dans la mesure où ces arguments s'appuient sur une théorie bien établie. Dans des disciplines relativement jeunes comme la psychologie ou l'éducation, ce type de théorie est rare. Même les théories les plus fréquemment invoquées continuent à faire l'objet de vives controverses scientifiques. Considérons, par exemple, le cas de la motivation intrinsèque (c.-à-d. la propension à réaliser une activité en l'absence de renforcement externe). Le fait qu'il existe des instruments raisonnablement bien conçus pour évaluer ce construit (ex. : McAuley *et al.*, 1989) n'empêche pas certains chercheurs de sérieusement remettre en question l'existence même de la motivation intrinsèque (ex. : Akin-Little *et al.*, 2004; Reiss, 2005). Bien qu'intéressants, les arguments liés à la validité de construit ne peuvent donc pas être considérés comme définitifs dans le domaine de la mesure en éducation ou en psychologie.

2.4 La validité de contenu et d'apparence

Heureusement, d'autres types d'arguments peuvent être invoqués pour soutenir la validité d'un test. Ces types d'arguments reposent sur l'opinion d'experts du domaine ou d'utilisateurs quant au contenu du test (Crocker et Algina, 2008). En ce qui concerne les arguments liés à la validité de contenu, la validité d'un score apparaît raisonnablement crédible lorsque des experts considèrent que le contenu du test est représentatif du concept qui est en principe évalué (AERA, APA et NCME, 2014). Par exemple, Parmar et collègues (1996) se sont penchés sur la validité de contenu de tests

utilisés pour évaluer les compétences en mathématiques d'élèves en difficulté d'apprentissage. Plus spécifiquement, les chercheurs se sont intéressés à la correspondance entre les items du test et le curriculum scolaire. Dans ce cas, la validité de contenu a été jugée limitée puisque le calcul arithmétique était, selon les experts, surreprésenté dans le test. Quant aux arguments liés à la validité apparente, ils reposent sur un principe similaire sauf qu'il s'agit, dans ce cas, de considérer le point de vue des utilisateurs (ex. : parents, élèves) plutôt que des experts du domaine. Considérer la validité apparente revient à établir dans quelle mesure le test est crédible aux yeux des utilisateurs. La démarche est pertinente puisque si ces derniers ne considèrent pas le test crédible, ils pourraient ne pas accorder d'importance au score ou même ne pas répondre au test avec suffisamment de soin (ex. : Chan *et al.*, 1997).

En général, les arguments reposant sur la validité de contenu ou d'apparence sont considérés comme utiles, mais insuffisants, pour appuyer la validité d'un score. Comme le soulignent Crocker et Algina (2008), il est possible que les experts soient biaisés dans leur jugement. Dans un même ordre d'idées, plusieurs spécialistes de la psychométrie considèrent que l'opinion des utilisateurs peut reposer sur des considérations vagues et superficielles, ces dernières entretenant possiblement des liens ténus avec l'utilité réelle du test (pour une description de la controverse entourant la validité apparente, voir Sartori, 2010).

2.5 La validité sociale

Les limites associées aux arguments reposant sur la validité de contenu ou d'apparence découlent en partie de la procédure utilisée pour recueillir l'opinion des experts ou des utilisateurs. Typiquement, les deux groupes doivent donner leur opinion en considérant uniquement le contenu du test, sans avoir la possibilité d'examiner un échantillon représentatif de réponses au test. Il ne serait donc pas étonnant, dans ce contexte, que les experts ne détectent pas toujours les items biaisés ou d'un niveau de difficulté

inapproprié ou que les utilisateurs omettent de considérer des éléments importants du test (ex. : le format de réponse). Bien qu'ils concernent davantage les interventions que les tests (voir cependant Bagnato *et al.*, 2014), les arguments liés à la validité sociale pourraient permettre de contourner ce problème.

Une intervention est considérée comme socialement valide lorsque les utilisateurs estiment qu'elle permet d'atteindre des objectifs importants en recourant à des moyens acceptables (Wolf, 1978; Snodgrass *et al.*, 2018). À titre d'exemple, Rubow et collègues (2018) ont démontré qu'une stratégie de gestion de classe, le jeu du bon comportement, permettait de réduire la fréquence des conduites perturbatrices dans les classes spécialisées pour élèves en troubles du comportement. Afin d'appuyer la validité sociale de cette intervention, les chercheurs ont demandé aux enseignants et aux élèves d'évaluer la pertinence de l'intervention en termes d'efficacité, de facilité d'utilisation et d'équité. L'intervention apparaît ici valide socialement dans la mesure où les deux groupes d'utilisateurs l'ont considérée comme pertinente. Dans le cas de la validité sociale, l'opinion des utilisateurs ne repose pas sur une description de l'intervention (ex. : dans un manuel), mais plutôt sur une observation directe de son utilisation et de ses résultats, sans référence à un critère ou à une théorie.

2.6 La validité d'utilisateur

Bien que cela n'ait pas encore été fait, il apparaît possible d'adapter les principes de la validité sociale aux tests, notamment à ceux évaluant la compréhension. Nous proposons ici un nouveau type d'argument reposant sur ce que nous appelons la validité d'utilisateur qui serait un hybride entre la validité sociale et la validité d'apparence. Pour établir ce type de validité, il faut déterminer dans quelle mesure un score établi en utilisant les réponses à un test reflète adéquatement la compétence de l'élève selon les normes informelles du milieu dans lequel il évolue, par exemple la classe. De manière analogue à ce qui est fait pour établir la validité sociale d'une intervention, l'expert ou

l'utilisateur serait appelé à considérer l'utilisation du test, incluant les réponses de l'élève, pour émettre son opinion.

Sans utiliser l'appellation en tant que telle, Dion et ses collègues ont commencé à invoquer des arguments liés à la validité d'utilisateur dans le cadre de travaux utilisant de nouveaux tests. Pour donner un exemple ne concernant pas la lecture ou la compréhension, Afia et collègues (2019) ont mené des entrevues structurées auprès d'adolescents afin notamment de recueillir de l'information sur le vécu familial de ces derniers au cours de l'année précédente. Les chercheurs ont élaboré une grille de cotation du contenu de l'entrevue permettant d'établir un score reflétant la qualité de l'encadrement parental offert à l'adolescent. Même si la grille de cotation (c.-à-d. le « test ») a été élaborée en s'inspirant des théories les plus récentes sur l'encadrement parental, la validité du score n'était pas garantie. Dans le but d'établir cette dernière, les chercheurs ont demandé à deux psychologues scolaires de lire un sous-échantillon de comptes-rendus d'entrevue et d'ordonner les adolescents selon la qualité d'encadrement parental reçu (du moins bien encadré au mieux encadré). Les deux psychologues scolaires ont effectué cette cotation de manière indépendante et sans consulter les scores attribués par les chercheurs. Comme les rangs attribués par les deux psychologues étaient fortement corrélés ($r = ,92$), un rang moyen a été calculé pour chaque adolescent. Ce rang moyen était aussi fortement corrélé ($r = ,82$) avec le score attribué par les chercheurs à l'aide de la grille de cotation, ce qui appuie la validité d'utilisateur de cette dernière. Bien qu'ils ne l'aient pas fait, Afia et collègues auraient pu considérer aussi le point de vue de parents ou d'adolescents pour établir la validité de leur score.

Il est à noter qu'Afia et collègues (2019) n'ont pas tenté de baliser le point de vue des psychologues, préférant leur laisser le soin de juger de la qualité de l'encadrement parental en se fiant à leur opinion spontanée ou brute. En d'autres termes, plutôt que de fournir aux psychologues une grille d'analyse précise, Afia et collègues ont

demandé à ces dernières de répondre en fonction du cadre de référence relativement informel qu'elles appliquaient dans leur pratique. Les psychologues ont ainsi été considérées comme des observatrices culturellement informées sensibles à une variété de nuances concernant le vécu d'adolescents appartenant à un groupe connu (dans ce cas, des adolescents de milieux défavorisés vivant dans la grande région montréalaise). Cette approche n'est pas sans précédent. Elle est notamment similaire à celle utilisée par Krokoff et collègues (1989) pour étudier les émotions exprimées lors d'interactions conjugales filmées en laboratoire. À l'instar d'Afia et collègues, ces chercheurs ont décidé de se fier au jugement brut d'observateurs pour détecter des émotions exprimées parfois de manière subtile par les conjoints (voir aussi Coan et Gottman, 2007). En l'absence de critères précis pour encadrer le jugement des experts ou des utilisateurs, il est nécessaire d'établir que le jugement qu'ils expriment n'est pas entièrement idiosyncratique. Il faut donc démontrer que les différents experts ou utilisateurs consultés émettent des jugements similaires sur les réponses des personnes évaluées. En ce sens, l'échantillon d'évaluateurs utilisé par Afia et collègues ($n = 2$) apparaît de taille minimale pour estimer le degré de convergence dans les jugements.

D'autres travaux réalisés par Dion et ses collègues concernent directement la lecture ou la compréhension. Roux et collègues (2015) ont évalué l'efficacité d'une nouvelle forme d'enseignement de la compréhension en utilisant notamment un test d'identification des idées principales d'un texte (un aspect important de la compréhension de textes narratifs). Les chercheurs ont dérivé un score en utilisant des critères de cotation précis. Afin d'établir la validité de ce nouveau test, les chercheurs ont présenté ce dernier (texte et consignes), ainsi que les réponses des élèves, à quatre enseignantes d'expérience. Roux et collègues ont aussi décidé de se fier au jugement brut de ces enseignantes en leur demandant d'utiliser une échelle d'impression plutôt générale (0 = « L'idée [formulée par l'élève] est absente ou diluée dans des détails à 2 = « L'idée est complète et n'inclut pas d'éléments superflus »). Comme dans l'étude d'Afia et collègues (2019), les enseignantes ont donné leurs impressions sans se

consulter entre elles, en ne référant pas aux scores attribués par les chercheurs et à l'aveugle (c.-à-d. sans disposer d'information sur les élèves). Les quatre enseignantes ont en général exprimé des impressions similaires (intraclasse = ,70) sur les réponses de chacun des élèves et leurs impressions étaient fortement corrélées ($r = ,83$) avec le score donné par les chercheurs, ce qui appuie la validité d'utilisateur du test mesurant la capacité à identifier les idées principales d'un texte.

Dans le même ordre d'idées, Michaud et collègues (2017) ont évalué l'efficacité d'une nouvelle approche d'enseignement du décodage des mots en demandant à de jeunes lecteurs en difficulté de lire des mots affichés à l'écran d'ordinateur et en corrigeant leurs réponses enregistrées audionumériquement à l'aide de critères très précis. Les chercheurs ont demandé à quatre orthopédagogues d'écouter les enregistrements, sans se consulter et à l'aveugle, en utilisant une échelle d'impression générale (0 = « le mot n'est pas lu du tout » à 5 = « le mot est lu parfaitement ») pour évaluer la lecture de chacun des mots par chacun des élèves. En général, les orthopédagogues ont exprimé des impressions similaires (intraclasse = ,95) et ces dernières étaient fortement corrélées avec les scores donnés par les chercheurs ($r = ,97$). Notons qu'en plus d'appuyer la validité du test de décodage utilisé par les chercheurs, le fait d'avoir consulté des orthopédagogues facilite le transfert de connaissances puisqu'il est possible de dire aux praticiens qu'ils auraient abouti essentiellement aux mêmes conclusions que les chercheurs s'ils avaient eux-mêmes coté les réponses des élèves (Wagener et Dion, 2018).

Finalement, Arcand et collègues (2014) se sont intéressés au lien entre la prosodie (intonation) et la compréhension du texte narratif chez le lecteur débutant de 2^e année du primaire. Les chercheurs ont utilisé un nouveau test de compréhension qu'ils ont validé dans le cadre de leur étude. Plus spécifiquement, ils ont demandé à chaque élève de lire un texte à voix haute. La compréhension a été évaluée par le biais du rappel : l'élève devait raconter à l'assistant « de quoi parlait le texte » sans voir ce dernier. Le

rappel (enregistré et retranscrit) a été utilisé pour attribuer un score reflétant à la fois la présence des éléments importants du texte, ainsi que la séquence dans laquelle ces éléments étaient mentionnés. En lien avec les théories sur la compréhension (van den Broek et Kendeou, 2017), ce score reflète dans quelle mesure le rappel est complet et structuré correctement. Afin de valider ce score, les chercheurs ont demandé à quatre enseignantes du primaire d'utiliser une échelle globale (de 1 = « Compréhension début 1^{re} année » à 3,5 = « Compréhension de niveau milieu 3^e année ») pour donner leurs impressions, indépendamment et à l'aveugle, de la compréhension de chaque élève. Comme dans les études de Roux et collègues (2015) et de Michaud et collègues (2017), les quatre enseignantes ont généralement exprimé des impressions similaires (intraclasse = ,84). Leurs impressions moyennes étaient aussi fortement corrélées ($r = ,84$) avec le score attribué par les chercheurs, ce qui suggère qu'il est possible d'utiliser des arguments liés à la validité d'utilisateur pour appuyer la pertinence d'un test de compréhension du texte narratif.

L'étude réalisée dans le cadre du présent mémoire visera à généraliser et à approfondir les résultats d'Arcand et collègues (2014). À l'instar de ces derniers, nous examinerons la validité d'un test de compréhension s'adressant à des lecteurs relativement peu avancés du primaire. Toutefois, alors qu'Arcand et collègues se sont intéressés à l'évaluation de la compréhension du texte narratif, nous considérons plutôt l'évaluation de la compréhension du texte informatif. La différence est potentiellement importante puisque, de l'avis de plusieurs (ex. : Williams *et al.*, 2004; Reed et Vaughn, 2012; PIRLS, 2017) les élèves en général, et ceux du primaire en particulier, comprennent moins bien le texte informatif que le texte narratif. Il est donc possible que leur compréhension fragmentaire du texte informatif soit particulièrement difficile à évaluer. Relativement peu de recherches se sont cependant intéressées à l'évaluation de la compréhension du texte informatif, notamment lorsque le rappel est utilisé (Lorch et van den Broek, 1997; Coté *et al.*, 1998).

Il s'agit d'une lacune importante. Il est nécessaire d'élaborer et de valider des tests de compréhension du texte informatif puisque cette compétence semble jouer vraisemblablement un rôle central dans la réussite scolaire. Dès la deuxième moitié du primaire, les élèves sont en effet appelés à lire des textes informatifs afin de réaliser des apprentissages, par exemple en lien avec l'univers social (voir Duke, 2004; Wagener et Dion, 2018). Cependant, plusieurs d'entre eux ont de la difficulté à apprendre en lisant, ce qui pourrait expliquer ce qui est appelé « la plongée de la quatrième année » (*fourth-grade slump*; Chall et Jacobs, 2003), c'est-à-dire la diminution drastique des scores aux tests de lecture observée chez les élèves de milieux défavorisés à partir du moment où des textes informatifs commencent à être fréquemment utilisés en classe. En renforçant la capacité du milieu scolaire à détecter les difficultés de compréhension du texte informatif chez les jeunes élèves, il pourrait être possible de mieux prévenir l'échec scolaire, notamment en milieu défavorisé.

2.7 Hypothèses

L'objectif de la présente étude est d'examiner l'utilité d'une procédure novatrice pour valider des tests de compréhension de textes informatifs destinés à des lecteurs de 3^e année du primaire. Afin d'illustrer et d'explorer l'application de cette procédure, la compréhension en lecture de deux textes est évaluée. Bien que ces deux textes soient informatifs sur le plan du contenu, un des deux suit la structure du récit narratif (il décrit factuellement une séquence d'actions d'un personnage) tandis que l'autre adopte la structure conventionnelle d'un texte informatif (ce dernier traite d'un phénomène physique et de sa cause). Les deux textes sont liés au domaine de l'univers social (la fondation de la Nouvelle-France). Le caractère novateur de notre procédure émane du fait que nous invoquons principalement, pour chacun des deux textes, des arguments liés à ce que nous appelons la validité d'utilisateur, c'est-à-dire que nous posons l'hypothèse que des enseignants seront en mesure de s'entendre sur le degré de compréhension démontré par les jeunes participants et que leur avis sera corrélé de

manière substantielle avec le score formel (tel qu'attribué par les chercheurs), démontrant ainsi la validité de ce dernier. Les enseignants donneront leurs impressions de manière indépendante et à l'aveugle (sans disposer d'information sur les élèves) en utilisant une échelle générale de cotation. En complément, nous invoquons également un argument lié à la validité de construit. Selon un modèle théorique généralement accepté (Hoover et Tunmer, 2018), un score valide de compréhension devrait être positivement corrélé avec l'exactitude et la rapidité (c'est-à-dire la fluidité) de la lecture, notamment chez les élèves de 3^e année (Language and Reading Research Consortium, 2015).

CHAPITRE III

MÉTHODOLOGIE

3.1 Participants

L'échantillon est constitué de 306 élèves (52 % de filles) de 3^e année du primaire, âgés de 8,27 ans en moyenne. Les élèves proviennent de 34 classes relevant de 14 écoles du Centre de services scolaire de Montréal (CSSDM). Ces écoles sont elles-mêmes situées dans des quartiers très défavorisés (*M* indice de défavorisation = 8,46; MEES, 2018). La répartition des origines ethniques des élèves est la suivante : caucasienne (43,8 %), maghrébine (21,9 %), africaine et caribéenne (13,7 %), asiatique (12,1) et sud-américaine (8,5 %).

L'échantillon a participé à une étude d'intervention au cours de l'année scolaire 2018-2019 (voir Wagener et Dion, 2018). Le consentement parental a été demandé pour l'ensemble des élèves des classes participantes. Parmi les élèves pour lesquels ce consentement a été obtenu (82,7 %), neuf ont été sélectionnés dans chaque classe à partir de leur score à un bref test de lecture de mots (Desrochers, 2008). Le score correspondant au nombre de mots lus correctement en 45 secondes a été utilisé pour sélectionner les trois élèves les plus faibles en lecture, trois élèves moyens (par rapport à leur classe) ainsi que les trois élèves les plus forts. Cette procédure de sélection visait à constituer un échantillon représentant la diversité des habiletés en lecture dans les classes participantes. Un prétest (voir ci-dessous) a été réalisé, en décembre, pour

chaque élève ainsi sélectionné avant le début de l'intervention (cette dernière n'est pas abordée dans le présent mémoire).

3.2 Instruments

La compréhension en lecture a été évaluée en individuel en demandant aux élèves de lire à voix haute deux textes informatifs. Le premier texte *Les dangers de l'océan* (173 mots) est la version résumée et réécrite dans un français moderne et simplifié d'un extrait des récits autobiographiques de Samuel de Champlain (Marrache-Gouraud, 2010). Cet extrait a été choisi pour l'intérêt des événements relatés (une quasi-collision avec un iceberg au milieu de l'océan Atlantique). Deux membres de l'équipe de recherche spécialistes de l'enseignement de l'histoire (Marc-André Éthier et David Lefrançois) ont été consultés pour créer cette version du texte (*voir Annexe A* pour la version originale et *Annexe B* pour la version adaptée). Selon l'indice de Daoust *et al.* (1996) exprimé en niveau scolaire, la lisibilité du texte *Les dangers de l'océan* est de 3,8, soit un niveau correspondant à la 3^e année du primaire. Le deuxième texte *L'hiver* (142 mots) est une synthèse de deux courts textes tirés du matériel d'intervention rédigés par l'équipe de recherche (Dion *et al.*, 2018). *L'hiver* décrit la cause d'un élément important de la colonisation, la rigueur inattendue des hivers en Nouvelle-France (*voir Annexe C*). Au moment de la rédaction, deux experts de la modélisation climatique (Patrick Grenier et René Laprise) ont été consultés et ont indiqué une source pertinente (Riser et Lozier, 2013), en plus de réviser la version finale du texte. Selon l'indice de Daoust et collègues, *L'hiver* est également de lecture facile sur le plan de la syntaxe et du lexique avec un indice de 3,3 (en termes de complément, voir *Annexes D et E* pour une analyse en propositions des deux textes).

Les deux textes ont été lus au cours d'une même séance en débutant par *Les dangers de l'océan* (l'ordre de lecture n'a pas été contre-balancé). Avant la lecture du premier texte, l'assistant de recherche a offert une mise en contexte en s'assurant que l'élève

savait ce qu'était un iceberg (« un très gros morceau de glace qui flotte sur l'océan [...] qui peut être aussi gros que ton école ») et en lui expliquant que le texte avait été écrit « il y a 400 ans par une personne qui s'appelle Samuel de Champlain ». L'élève disposait d'un maximum de quatre minutes pour lire le texte à voix haute. Afin de mesurer la fluidité de lecture de l'élève, l'assistant a chronométré sa lecture puis a compté le nombre de mots lus sans erreurs (sans pénaliser l'élève lorsqu'il s'autocorrigeait). Un score de fluidité correspondant au nombre de mots lus correctement à la minute a été calculé (voir Fuchs *et al.*, 2001).

À la fin de la lecture du texte ou lorsque la période allouée à la lecture était terminée, l'assistant a demandé à l'élève d'essayer « d'expliquer ce que Champlain raconte dans le texte comme si je ne connaissais pas le texte ». Un maximum de deux minutes et de deux encouragements (ex. : « Quelque chose d'autre? » ou « Et après? ») a été donné à l'élève. Le rappel a été transcrit verbatim et enregistré en audionumérique à des fins de vérifications.

Les conditions de passation étaient similaires pour le deuxième texte *L'hiver*. L'élève disposait d'un maximum de trois minutes pour lire ce texte à voix haute. Pour la mise en contexte avant la lecture, l'assistant demandait à l'élève s'il savait ce qu'était la Nouvelle-France (« Est-ce que tu sais ce qu'est la Nouvelle-France? ») et précisait que c'est l'endroit qui s'appelle maintenant le Québec (« là, où nous vivons »). Il rappelait aussi brièvement qui était Samuel de Champlain (« un des premiers Français venus s'installer au Québec ») qui est mentionné dans *L'hiver*. La fluidité de lecture a été calculée et l'élève a disposé de deux minutes pour faire un rappel du texte. Soulignons que les consignes du rappel étaient exactement les mêmes que pour le premier texte.

Le rappel de chaque texte a été codifié en utilisant une grille de correction répertoriant tous les éléments importants de chacun des textes. Un élément a été considéré important si sa présence dans le texte est essentielle à la cohérence de ce dernier. La grille de

correction pour *Les dangers de l'océan* est présentée au tableau 3.1, celle pour *L'hiver* au tableau 3.2. Plusieurs versions préliminaires des grilles ont été mises à l'essai avant d'aboutir aux présentes versions.

En termes de cotation, un élément a été considéré comme présent dans le rappel seulement lorsqu'il a été énoncé en l'absence d'intrusion, c'est-à-dire sans contradiction ou confusion évidentes (Williams *et al.*, 2007). Les mentions (sans intrusion) des relations de cause à effet présentées dans chaque texte (deux pour *Les dangers de l'océan* et trois pour *L'hiver*) ont également été considérées. Tous les rappels ont été codifiés de manière indépendante par la candidate et par une seconde correctrice.

Calculé pour l'ensemble des items du rappel du texte *Les dangers de l'océan*, le pourcentage d'accord interjuges est de 90,6 et le Kappa de 0,76 (cet indice introduit une correction pour les accords potentiellement obtenus au hasard). Pour *L'hiver*, le pourcentage d'accord est de 92,4 et le Kappa est de 0,79. Tous les désaccords ont été résolus par consensus dans le cadre de rencontres impliquant la candidate, la deuxième correctrice et le directeur de recherche.

Tableau 3.1 Grille de correction pour *Les dangers de l'océan*

Idée	Intrusion
1 : Départ de la France/vont en Nouvelle-France	Départ de la Nouvelle-France (point de départ plutôt que de destination)
2 : (Personnages) naviguent/voyagent (se déplacent) en bateau (canots)/sont sur la mer/l'océan	Le but du voyage est d'aller voir des icebergs
3 : Vents font avancer le bateau (personnages) ou sont dans la bonne direction/bons vents	
4 : Iceberg (distinct de celui en 6/7) aperçu/signe qu'il va y en avoir (beaucoup) d'autres	
5 : Difficile de voir (nuit ou brume)	
6 : Iceberg (distinct de celui en 4) brise presque/touche le bateau/aperçu dernière minute	L'iceberg poursuit le bateau, se dirige sur le bateau ou le pousse vers le nord
7 : Iceberg (en 6) s'éloigne/laisse passer le bateau	
8 : (Personnages) ont eu peur	
Explicitation de la relation de cause à effet	
9 : Hiver pas terminé/partis trop tôt ⇔(présence) icebergs	
10 : Peur ⇔long pour se calmer	

Tableau 3.2 Grille de correction pour *L'hiver*

Idée	Intrusion
1 : Hiver froid en N.-F./hiver ennemi/N.-F. devient un bloc de glace/France plus chaud	France plus froide. (Simplement) Champlain n'aime pas l'hiver (sans mention que l'hiver en N.-F. est froid)
2 : (Personnage) ne comprend pas (se demande) pourquoi hivers N.-F. plus froids (qu'en France)	
3 : Canots ne peuvent pas avancer (sont gelés)	
4 : Aliments (fleurs) ne poussent pas (ont besoin de chaleur)	
5 : N.-F. pas plus près du Pôle Nord/France pas plus au sud	N.-F. proche du Pôle Nord
6 : Vents froids sur la N.-F.	Vents chauds sur la N.-F./vents réchauffent la N.-F.
7 : Vents chauds (réchauffent) la France	Vents froids sur la France
Explicitation de la relation de cause à effet	
8 : Froid/glace ⇔ canots ne peuvent pas avancer (difficile de se déplacer)	
9 : Froid/glace ⇔ aliments ne poussent pas	
10 : Hivers N.-F. plus froids ⇔ vents	N.-F. proche du Pôle Nord.

Afin d'établir la validité d'utilisateur du score de rappel et de tester les hypothèses formulées dans le cadre du présent projet, quatre enseignantes ont lu les transcriptions d'un échantillon aléatoire de 25 % des rappels (sélectionné séparément pour chacun des textes). Les enseignantes ont été recrutées par la candidate par le biais de communautés enseignantes (ex. : Groupe Facebook – Partage 2^e cycle). Elles

travaillent toutes au secteur régulier avec un minimum de cinq années d'expérience au 2^e cycle du primaire (3^e ou 4^e année). Elles ont donné leurs impressions sur chacun des rappels, de manière indépendante et à l'aveugle, sans avoir reçu aucune information sur les grilles de correction formelles. Elles ont plutôt uniquement utilisé une échelle d'impression globale en cinq points (de 0 à 4), dont trois des points d'ancrage sont définis en termes généraux (0 = n'a pas du tout compris le texte; 2 = a compris la moitié des éléments importants; 4 = a parfaitement compris le texte; les points intermédiaires 1 et 3 ne sont pas définis). Elles ont été informées « qu'afin de vous permettre d'ajuster vos évaluations, les cinq premiers rappels de chaque texte ont été sélectionnés de manière à représenter l'ensemble de la diversité au sein de l'échantillon ». En fait, ce sous-échantillon de calibration incluait un rappel très faible, un rappel très fort ainsi que trois rappels de qualité intermédiaire qui n'ont pas été identifiés comme tels ou présentés aux enseignantes dans un ordre particulier (une procédure similaire à celle utilisée par Arcand *et al.*, 2014). Les enseignantes ont disposé d'une période de deux semaines pour réaliser la tâche, durant leurs temps libres, en échange d'une compensation équivalente à leur salaire (leur consentement a été recueilli, voir *Annexe F*). L'échelle qu'elles ont utilisée a été prétestée dans le cadre d'un pilote impliquant trois enseignantes dont les impressions n'ont pas été considérées dans ce qui suit.

3.3 Procédure pour l'évaluation des élèves

Les évaluations des élèves ont été réalisées en individuel, en décembre 2018, le plus souvent dans le corridor à proximité de leur classe, isolés par un paravent. Ces évaluations ont été menées par la candidate et six assistants de recherche, des étudiants de baccalauréat, de maîtrise ou de doctorat dûment formés et étroitement encadrés. Les assistants ont réalisé les évaluations en suivant un protocole précis, et ce, sous la supervision de la candidate (qui était aussi coordonnatrice du projet) et du directeur de recherche. Les évaluations des élèves utilisées dans le cadre du présent projet

constituent une partie du prétest de l'étude d'intervention menée par le directeur de recherche.

3.4 Analyses

Le score (cotation formelle) pour chacune des évaluations a été calculé en utilisant l'analyse factorielle confirmatoire pour variables dichotomiques telle qu'implantée dans la version 8.4 du logiciel Mplus (Muthén et Muthén, 2019) et en considérant la structure nichée des données (élèves dans des classes)¹. En accord avec les recommandations méthodologiques récentes (ex. : McNeish, 2018; Graham, 2006), nous avons d'abord inspecté les cotations aux items de rappel pour chacun des textes, écarté les items inappropriés et établi la dimensionnalité (structure factorielle) des items conservés ainsi que la pondération accordée à ces derniers dans le calcul des scores de compréhension. Toutes ces opérations ont permis d'estimer la cohérence interne des scores en utilisant l'indice approprié. Après avoir calculé les scores de compréhension pour chacun des textes, nous avons utilisé à nouveau l'analyse factorielle confirmatoire pour mettre en lien ces scores avec la fluidité de lecture et les impressions des enseignantes. Des analyses descriptives ainsi que le coefficient de corrélation intraclasse ont été utilisés afin d'explorer les propriétés de ces impressions.

¹ Nous avons considéré d'utiliser la théorie de la réponse aux items, mais n'avons pas retenu cette option puisqu'elle présume généralement une structure unidimensionnelle (Ayala, 2009), un postulat qui n'est pas rencontré dans le cas d'un de nos textes.

CHAPITRE IV

RÉSULTATS

4.1 Analyses préliminaires

Les tableaux 4.1 à 4.4 présentent, séparément pour chacun des deux textes, les statistiques descriptives pour la cotation formelle aux items de rappel, incluant un indice de difficulté de l'item correspondant à la proportion d'élèves n'ayant pas mentionné l'élément correspondant du texte dans leur rappel, une grande proportion indiquant un niveau de difficulté élevé (ex. : Franzen, 2011) ou de prégnance plus faible. Des corrélations tétrachoriques pour variables dichotomiques sont également rapportées. La corrélation tétrachorique s'interprète essentiellement comme la corrélation de Pearson pour variables continues (Lorenzo-Seva et Ferrando, 2012).

Selon ces statistiques descriptives, les items se comportent de manière similaire, à deux égards, pour *Les dangers de l'océan* et pour *L'hiver*. Premièrement, les niveaux de difficulté sont très variables (Tableaux 4.1 et 4.3). Deuxièmement, les corrélations entre les cotations aux items sont, à quelques exceptions près, relativement faibles (Tableaux 4.2 et 4.4).

Pour ce qui est spécifiquement des items du rappel du texte *Les dangers de l'océan*, leurs niveaux de difficulté varient entre ,31 et ,91 (Tableau 4.1) et leurs corrélations

entre $-,01$ et $,56$ (Tableau 4.2). Pour ce texte, nous avons décidé de retirer l'item 10, ce dernier étant très difficile tout en étant essentiellement redondant avec l'item 8.

En ce qui concerne les items du rappel de *L'hiver*, leurs niveaux de difficulté varient entre $,17$ et $,89$ (Tableau 4.3) et leurs corrélations entre $-,06$ et $,93$ (Tableau 4.4). Étant donné le niveau de difficulté particulièrement élevé de l'item 5, nous avons décidé de le retirer. Bien que les items redondants puissent poser problème dans certaines analyses factorielles (Bollen et Lennox, 1991), cela n'a pas été le cas en ce qui nous concerne. Nous avons par conséquent décidé de conserver certaines paires d'items redondants (les items 3 et 8 et 4 et 9 du rappel de *L'hiver*).

Tableau 4.1 Niveau de difficulté (et erreur standard) des items du rappel du texte *Les Dangers de l'océan*

Item	Niveau de difficulté ^a	Erreur standard ^b
1 : départ de France	,75	,02
2 : ils sont sur la mer	,55	,03
3 : bons vents	,71	,03
4 : iceberg aperçu	,57	,03
5 : difficile de voir	,68	,03
6 : iceberg brise presque	,31	,03
7 : iceberg s'éloigne	,60	,03
8 : ils ont eu peur	,72	,03
9 : hiver non terminé/icebergs	,91	,02
10 : peur/long pour se calmer	,89	,02

Note. ^a Niveau de difficulté = proportion d'élèves n'ayant pas mentionné l'élément dans le rappel. ^b Estimation du degré d'erreur de la proportion échantillonnale relativement à la valeur du paramètre pour la population (Fleiss *et al.*, 2003).

Tableau 4.2 Corrélations tétrachoriques pour *Les Dangers de l'océan*

Item	1	2	3	4	5	6	7	8	9	10
1 : départ de France	—									
2 : ils sont sur la mer	,07	—								
3 : bons vents	,04	,15**	—							
4 : iceberg aperçu	,12*	,25**	,20**	—						
5 : difficile de voir	,06	,14*	,08	,28**	—					
6 : iceberg brise presque	,10	,13*	,06	-,01	,13*	—				
7 : iceberg s'éloigne	,03	,13*	,04	,02	,12*	,33**	—			
8 : ils ont eu peur	,10	,05	,04	,05	,18**	,17**	,06	—		
9 : hiver non terminé/icebergs	,13*	,12*	,02	,19**	,02	,09	,04	,03	—	
10 : peur/long pour se calmer	,06	,07	,07	,03	,18**	,15**	,10	,56**	,07	—

Note. * $p < ,05$. ** $p < ,01$.

Tableau 4.3 Niveau de difficulté (et erreur standard) des items du rappel du texte
L'hiver

Item	Niveau de difficulté ^a	Erreur standard ^b
1 : hiver froid en N.-F.	,17	,02
2 : se demande pourquoi	,63	,03
3 : canots ne peuvent avancer	,73	,03
4 : plantes ne peuvent pousser	,61	,03
5 : même latitude	,89	,02
6 : vents froids sur la N.-F.	,67	,03
7 : vents chauds sur la France	,71	,03
8 : froid/canots immobilisés	,76	,02
9 : froid/pas de plantes	,64	,03
10 : hiver plus froid/vents	,70	,03

Note. ^a Niveau de difficulté = proportion d'élèves n'ayant pas mentionné l'élément dans le rappel. ^b Estimation du degré d'erreur de la proportion échantillonnale relativement à la valeur du paramètre pour la population (Fleiss *et al.*, 2003).

Tableau 4.4 Corrélations tétrachoriques pour *L'hiver*

Item	1	2	3	4	5	6	7	8	9	10
1 : hiver froid en N.-F.	—									
2 : se demande pourquoi	,25**	—								
3 : canots ne peuvent avancer	,09	-,04	—							
4 : plantes ne peuvent pousser	,11	,02	,33**	—						
5 : même latitude	,07	,12*	,01	-,01	—					
6 : vents froids sur la N.-F.	,21**	,35**	-,02	,04	-,04	—				
7 : vents chauds sur la France	,19**	,34**	-,01	,02	-,01	,75**	—			
8 : froid/canots immobilisés	,11	-,06	,91**	,34**	-,02	-,02	-,03	—		
9 : froid/pas de plantes	,15**	,03	,34**	,93**	,01	,07	,04	,38**	—	
10 : hiver plus froid/vents	,24**	,41**	,05	,09	,03	,78**	,59**	,03	,09	—

Note. * $p < ,05$. ** $p < ,01$.

4.2 Dimensionnalité et pondération des items conservés

Calculer pour chaque texte un score correspondant simplement au nombre (non pondéré) d'éléments mentionnés dans le rappel n'est pas optimal, et ce, pour deux raisons. Premièrement, calculer un tel score revient à considérer que tous les items sont essentiellement interchangeables (voir McNeish, 2018), ce qui est discutable étant donné leur degré très variable de difficulté. Deuxièmement, comme les corrélations entre items sont généralement faibles ou modérées, tous les items n'évaluent pas nécessairement un seul et même aspect de la compréhension qui serait reflété par un seul score.

L'analyse factorielle confirmatoire (ex. : Kline, 2011) a été utilisée afin d'établir la dimensionnalité des rappels, c'est-à-dire le nombre de facteurs (scores) à dériver, ainsi que la pondération des items dans le calcul de ces scores. De manière plus spécifique, ces analyses permettent de tester dans quelle mesure les données (cotations aux items) coïncident avec les hypothèses spécifiques proposées (modèles). Suivant les recommandations habituelles (Hu et Bentler, 1999; Marsh *et al.*, 2005; Yu 2002), le degré d'ajustement (*fit*) de chaque modèle aux données a été estimé à l'aide de trois indices : le Confirmatory Fit Index (CFI > ,90), le Tucker Lewis Index (TLI > ,90) et le Root Mean Square Error of Approximation (RMSEA < ,08). Un ajustement inadéquat indique que certaines hypothèses du modèle sont trop restrictives et doivent être modifiées.

En ce qui concerne le rappel du texte *Les dangers de l'océan*, nous avons commencé par tester un modèle unidimensionnel (un facteur) avec des erreurs ou unicités (*uniqueness*) non corrélées, sans imposer de contraintes aux poids de saturation (*loadings*) des items. En vertu d'un tel modèle, tous les items contribuent, possiblement à divers degrés, à un facteur unique qui rend compte de l'ensemble des liens entre les

items (unicités non corrélées). Ce modèle ne décrit pas particulièrement bien les données (CFI = ,774; TLI = ,698; RMSEA = ,062), possiblement parce que le texte aborde différents sous-thèmes qui ne sont pas représentés de la même manière dans le rappel. Afin de tenir compte de la similarité entre certains items et d'améliorer l'ajustement (Morin *et al.*, 2016), le modèle a été modifié en permettant aux unicités des items 6, 7, et 8 (qui décrivent la quasi-collision avec l'iceberg) d'être corrélées (c.-à-d. de ne pas être entièrement expliquées par le facteur principal). L'ajustement du modèle modifié est excellent (CFI = ,968; TLI = ,951; RMSEA = ,034) et l'écart entre les données et les prédictions du modèle n'est pas significatif, χ^2 (dl = 24) = 28,48, *n.s.* La figure 4.1 présente le modèle. Un modèle plus restrictif dans lequel tous les poids de saturation ont été contraints à être égaux a également été testé. L'ajout de cette contrainte apparemment trop restrictive a fait diminuer de manière substantielle l'ajustement du modèle (CFI = ,891; TLI = ,877; RMSEA = ,039; pour les critères de changement d'ajustements, voir Chen, 2007; Cheung et Rensvold, 2002). Cette dernière contrainte n'a donc pas été retenue.

Le rappel du texte *Les dangers de l'océan* peut donc être représenté à l'aide d'un seul facteur (score) auquel les items retenus contribuent à divers degrés. Comme le soulignent les méthodologues (ex. : Graham, 2006), le degré de cohérence interne de ce facteur n'est pas représenté adéquatement par le coefficient alpha ($\alpha = ,52$). Le coefficient H est préférable dans la mesure où il n'est pas biaisé à la baisse par des poids de saturation inégaux ou à la hausse par des unicités corrélées (McNeish, 2018). Ce coefficient plus libéral suggère que la cohérence interne du score est acceptable ($H = ,72$).

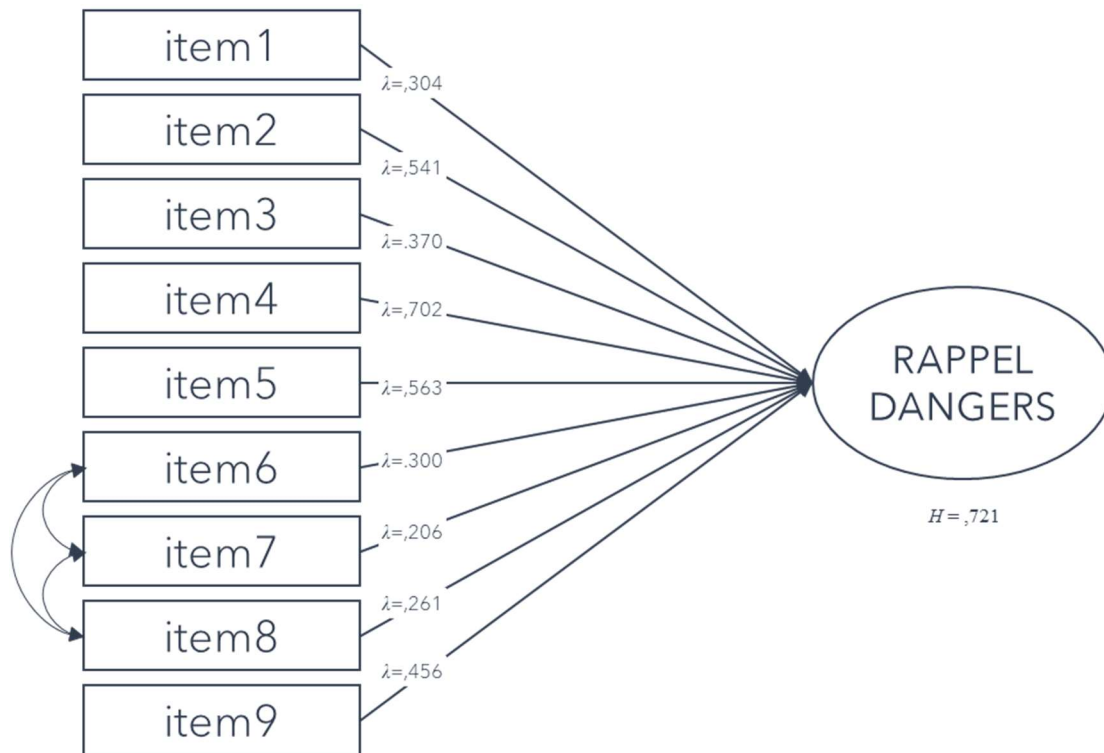


Figure 4.1 Modèle de mesure finale pour *Les dangers de l'océan*

Note : λ : poids de saturation standardisé. Tous les poids sont significatifs à $p < ,05$.

Pour ce qui du rappel du texte *L'hiver*, nous avons également considéré comme point de départ un modèle à un facteur avec unicités non corrélées et sans contrainte imposée aux poids de saturation. Les analyses pour ce modèle n'ont pas convergé, ce qui indique un ajustement problématique. De plus, le fait de permettre à un nombre parcimonieux d'unicités de corrélérer entre elles n'a pas mené à l'identification d'un modèle unidimensionnel adéquat. Un examen des sous-thèmes abordés dans ce texte a permis d'identifier un modèle à deux facteurs facilement interprétable. Un des facteurs regroupe les cinq items concernant la température froide en Nouvelle-France alors que l'autre facteur regroupe plutôt les quatre items décrivant les conséquences du froid. L'ajustement de ce modèle à deux facteurs, avec trois unicités corrélées, et sans

contrainte liée aux poids, est adéquat (CFI = ,997; TLI = ,994; RMSEA = ,036), χ^2 (dl = 24) = 25,09, *n.s.* Il est représenté à la figure 4.2. Contraindre les poids de saturation à être égaux au sein de chaque facteur a rendu l'ajustement du modèle problématique (CFI = ,984; TLI = ,982; RMSEA = ,064). Cette version modifiée du modèle n'a donc pas été retenue.

Le rappel du texte *L'hiver* peut donc être représenté à l'aide de deux scores (facteurs) auxquels les items contribuent à divers degrés. Encore une fois, le coefficient alpha sous-estime la cohérence interne (facteur froid $\alpha = ,78$; facteur conséquences $\alpha = ,82$). Les valeurs pour le coefficient *H* sont de ,86 pour le facteur froid et de ,89 pour le facteur conséquences.

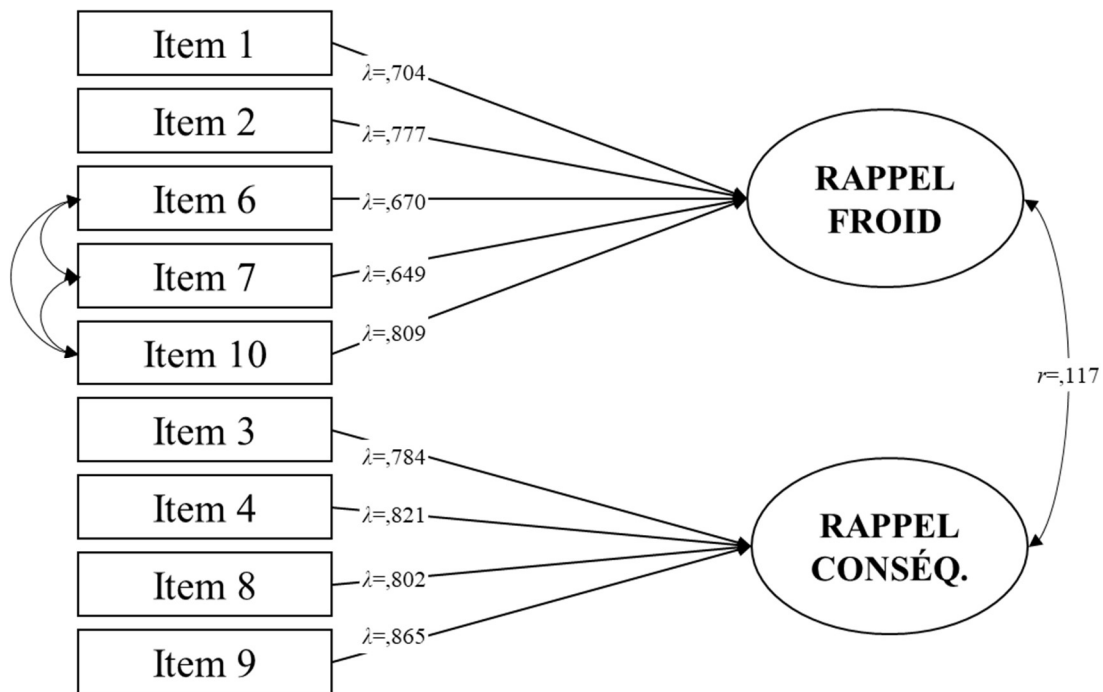


Figure 4.2 Modèle de mesure finale pour *L'hiver*

Note : λ : poids de saturation standardisé. Tous les poids sont significatifs à $p < ,05$.

4.3 Statistiques descriptives sur les impressions des enseignantes

Les moyennes et les écarts-types des impressions des quatre enseignantes sur le sous-échantillon aléatoire de rappels sont présentés au tableau 4.5. Pour les deux textes, les impressions sont en moyenne similaires pour les quatre enseignantes et elles varient manifestement d'un élève à l'autre (écarts-types). À noter que toutes les enseignantes ont utilisé l'ensemble des points de l'échelle (de 0 à 4) pour les deux textes.

Tableau 4.5 Impression moyenne des quatre enseignantes sur la qualité du rappel des deux textes

Enseignante ^a	<i>Les dangers de l'océan</i> (n = 76)	<i>L'hiver</i> (n = 78)
1	1,7 (1,2)	1,8 (1,2)
2	1,6 (1,2)	2,0 (1,3)
3	1,6 (1,1)	1,7 (1,3)
4	1,3 (1,2)	1,6 (1,3)

Note. Échelle en cinq points : 0 = n'a pas du tout compris le texte; 2 = a compris la moitié des éléments importants; 4 = a parfaitement compris le texte. L'écart-type est entre parenthèses. ^a Numéro d'identification de l'enseignante.

L'hypothèse selon laquelle les enseignantes seraient en mesure de s'entendre sur la qualité des rappels a été testée à l'aide du coefficient de corrélation intraclasse. Ce coefficient peut être calculé de plusieurs façons, mais varie, dans tous les cas, entre 0 et 1 (accord parfait; McGraw et Wong, 1996). Comme nous souhaitons déterminer dans quelle mesure l'impression moyenne des quatre enseignantes permettait de distinguer les rappels des élèves, nous avons utilisé la formule pour évaluation moyenne et accord absolu (plutôt que relatif) entre évaluateurs. Dans ces analyses réalisées séparément pour les deux textes, les élèves et les enseignantes sont considérés comme des facteurs aléatoires (c.-à-d. échantillonnés au sein d'une population à laquelle nous souhaitons généraliser). Pour le rappel du texte *Les dangers de l'océan*, le coefficient est de ,93, alors qu'il est de ,94 pour le rappel du texte *L'hiver*, ce qui indique la présence d'un degré élevé d'accord entre les enseignantes.

4.4 Lien entre la fluidité, la cotation du rappel et l'impression des enseignantes

L'analyse factorielle confirmatoire a été à nouveau utilisée pour établir la validité des scores dérivés de notre cotation des rappels en examinant l'association entre ces derniers, l'impression moyenne des enseignantes ainsi que la fluidité. Comme les impressions des enseignantes étaient disponibles seulement pour un sous-échantillon aléatoire d'approximativement 1 élève sur 4, l'imputation multiple (10 bases de données) pour données manquantes complètement au hasard (Enders, 2010) a été utilisée pour estimer les valeurs manquantes pour l'impression moyenne, ce qui a permis de réaliser les analyses sur l'échantillon complet.

Les résultats pour le texte *Les dangers de l'océan* sont présentés à la figure 4.3. Tel qu'attendu, la fluidité est significativement associée au facteur dérivé de notre cotation du rappel, alors que ce dernier est associé à l'impression moyenne des enseignantes. En fait, le facteur dérivé de notre cotation du rappel explique plus de 50% de la variance des impressions moyennes.

Pour ce qui est du texte *L'hiver* (Figure 4.4), la fluidité est significativement associée aux deux facteurs dérivés de notre cotation du rappel, alors que ces deux facteurs sont en lien avec l'impression moyenne des enseignantes. Conjointement, les deux facteurs expliquent approximativement 78% de la variance des impressions moyennes des enseignantes.

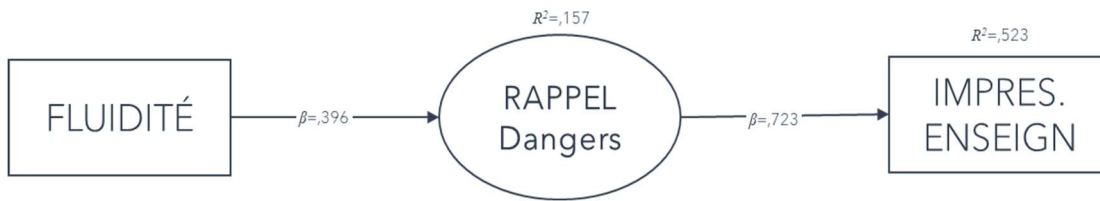


Figure 4.3 Lien entre la fluidité, le score au rappel et l'impression moyenne des enseignantes pour *Les dangers de l'océan*

Note. Tous les paramètres sont significatifs à $p < ,05$.

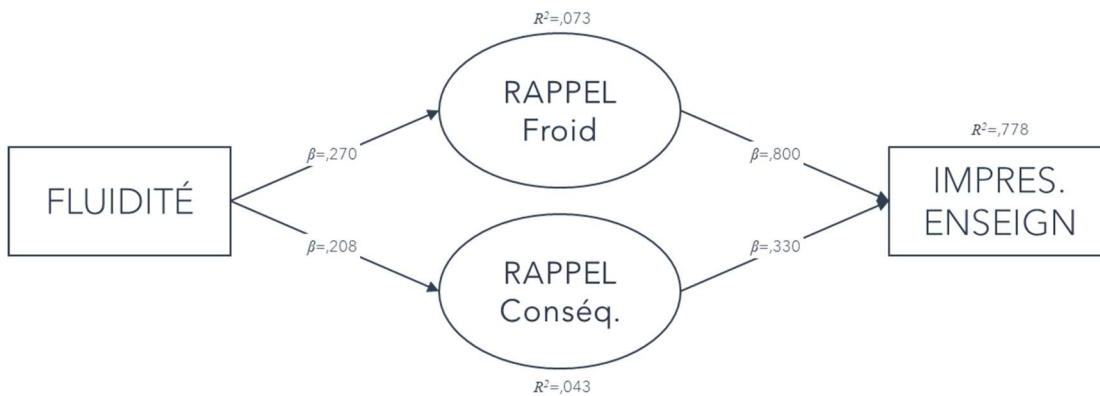


Figure 4.4 Lien entre la fluidité, la cotation des rappels et les impressions des enseignantes pour *L'hiver*

Note. Tous les paramètres sont significatifs à $p < ,05$.

4.5 Calcul des scores de compréhension

Pour que des évaluations de la compréhension comme les nôtres puissent être utilisées dans un contexte de pratique, il faut minimalement que les enseignants puissent calculer les scores sans avoir à utiliser un logiciel dispendieux et peu convivial d'analyses factorielles confirmatoires (ex. : Mplus). Heureusement, il est possible de

calculer assez simplement le score d'un élève à l'aide d'un chiffrier Excel (Figure 4.5) dans lesquels les éléments mentionnés dans le rappel sont pondérés par les poids de saturation, additionnés, puis convertis en un score d'une valeur interprétable (ex. : sur 9, le nombre d'éléments du rappel considérés pour le texte *Les dangers de l'océan*).

Item	Élément présent	Pondération
1: départ de France	1	0,30
2: ils sont en mer	1	0,54
3: bons vents	0	0,37
4: iceberg aperçu	0	0,70
5: difficile de voir	0	0,56
6: iceberg brise presque	1	0,30
7: iceberg s'éloigne	1	0,21
8: ils ont peur	1	0,26
9: hiver non terminé/iceberg	0	0,46
Score de l'élève:		3,9/9

Figure 4.5 Chiffrier permettant de calculer le score au rappel du texte *Les dangers de l'océan*

Le score calculé de cette façon est distribué de manière raisonnablement normale avec un effet de plancher minimal pour le rappel unidimensionnel du texte *Les dangers de l'océan*. Ce score est distribué normalement selon les indices d'aplatissement (-,74) et de symétrie (,31) avec un effet de plancher (score = 0) minimal de 7,2%. Étant donné la nature bifactorielle de la cotation de son rappel, deux scores sont calculés pour le texte *L'hiver*, un pour les éléments liés au froid et un autre pour les éléments liés aux conséquences du froid. Ces deux scores sont à la limite de la normalité statistique selon les indices d'aplatissement (froid = -,94; conséquences du froid = -,83) et de symétrie (froid = ,56 ; conséquences du froid = ,75). Cependant, l'effet de plancher est

substantiel pour le score lié au facteur froid (14% des scores = 0) et, en particulier, pour le facteur lié au facteur conséquences du froid (51% des scores = 0).

CHAPITRE V

DISCUSSION

L'objectif de ce mémoire était de mettre à l'essai une approche novatrice de validation de tests en compréhension de lecture. L'approche proposée invoque principalement des arguments dits de validité d'utilisateur, c'est-à-dire le degré selon lequel le score formel à l'évaluation reflète adéquatement la compétence de l'élève selon les normes informelles du milieu. Pour qu'une telle approche soit applicable, il est essentiel que les experts œuvrant dans le milieu soient en mesure d'appliquer ces normes de manière cohérente. En lien avec nos hypothèses, nous avons démontré que des enseignantes exprimaient des impressions similaires sur la qualité du rappel des élèves et que leurs impressions correspondaient plutôt étroitement avec la cotation formelle de la qualité de ce rappel. Nos résultats illustrent également comment il est possible d'invoquer à la fois des arguments liés à la validité d'utilisateur et à la validité de construit. Concernant ce dernier point, en accord avec une théorie généralement acceptée (Hoover et Tunmer, 2018; Language and Reading Research Consortium, 2015), nous avons démontré que les élèves avec une meilleure fluidité en lecture obtenaient généralement des scores de rappel plus élevés. S'il semble pertinent, d'après nos résultats, de considérer des arguments liés à la validité d'utilisateur, il faut garder en tête que valider un test représente une opération complexe et qu'une multitude d'arguments peuvent être invoqués (Markus et Borsboom, 2013). Il est important, dans ce contexte, d'établir l'utilité et le caractère unique du nouveau type d'argument que nous avons élaboré dans le cadre du présent mémoire.

Avant d'établir la validité de nos scores de rappel, nous avons dû les calculer, ce qui s'est avéré exigeant. Il est bien établi qu'il est possible de coter, avec un bon accord interjuges, des aspects relativement subtils de la qualité du rappel (pour une recension, voir Reed et Vaughn, 2012). Ce qui est moins clair, c'est comment cette cotation peut être convertie en score reflétant la compréhension. van den Broek et ses collègues (ex. : Trabasso *et al.*, 1989) considèrent seulement les éléments (ou propositions) les plus centraux du texte et dérivent un score qui accorde le même poids à tous ces éléments (ex. : Kendeou *et al.*, 2009). Les propriétés psychométriques de ce type de scores, notamment leur cohérence interne, ne sont pas rapportées (ex. : Beker *et al.*, 2019; Clinton et van den Broek, 2012; Kendeou *et al.*, 2009). Dans le cadre du présent mémoire, nous avons innové en dérivant nos scores à l'aide de l'analyse factorielle confirmatoire, et en estimant leur cohérence interne à l'aide de l'indice le plus approprié selon les développements méthodologiques récents (McNeish, 2018). La procédure de dérivation des scores que nous avons utilisée a le mérite d'être entièrement transparente et aisément reproductible.

L'examen détaillé de la cotation aux items a révélé que le niveau de difficulté de ces derniers était très variable, ce qui suggère qu'il n'aurait pas été optimal de leur accorder un poids uniforme lors du calcul du score (c.-à-d. utiliser un score simplement égal au nombre d'éléments cotés comme présents dans le rappel) ou de présumer qu'ils reflétaient tous le même construit. Notons que nous avons envisagé la possibilité de traiter les items comme des indicateurs causaux (Bollen et Lennox, 1991). En psychométrie, les items d'un même instrument sont habituellement considérés comme des indicateurs *d'effet*, c'est-à-dire que les réponses à ces items seraient toutes, à divers degrés, influencées par le construit mesuré. Par exemple, les réponses aux items d'un inventaire d'humeurs dépressives seraient le reflet du degré de dépression (le construit) ressenti par le répondant. De ce point de vue, les réponses aux items d'effet devraient démontrer une cohérence interne (c.-à-d. être toutes similaires pour un répondant en particulier) puisqu'elles sont toutes influencées (possiblement à divers degrés) par le

même construit. Dans le cas par contre des indicateurs *causaux*, c'est la réponse à chacun des items qui détermine le construit (plutôt que l'inverse). C'est le cas, par exemple, lorsque des items sur l'éducation, le revenu et le prestige de la profession, en principe indépendamment les uns des autres, causent le statut socio-économique (le construit) d'un adulte. De ce point de vue, la rétention des différents éléments du texte (les items) pourrait être la cause du degré de compréhension (le construit), une idée qui n'est pas incompatible avec les théories sur le développement de la compréhension (ex. : van den Broek *et al.*, 1999, Chap. 3). Toutefois, considérant la vive controverse entourant les indicateurs causaux et les difficultés non résolues liées à leur estimation (Bollen et Diamantopoulos, 2017; Edwards, 2011; Howell, 2014), cette option n'a pas été retenue dans le cadre du présent mémoire.

En comparaison, l'analyse des impressions des enseignantes s'est avérée relativement simple étant donné leur convergence évidente. Il est clair, en d'autres termes, que les quatre enseignantes ayant évalué à l'aveugle et de manière indépendante le sous-échantillon de rappels pour les deux textes exprimaient un point de vue similaire sur la qualité de ces rappels. Afin de refléter ce degré de consensus, nous avons utilisé la corrélation intraclasse, un indice qui peut être calculé de différentes façons selon la nature des données et des inférences que le chercheur souhaite réaliser (McGraw et Wong, 1996). Selon la façon retenue ici, les corrélations intraclasse élevées que nous avons observées suggèrent, en particulier, qu'une impression moyenne basée sur l'opinion de quatre enseignantes est hautement généralisable. En d'autres termes, étant donné le degré de consensus observé, il est fort probable qu'il existe dans le milieu scolaire une norme relativement consensuelle et que des résultats similaires auraient été obtenus en recourant aux impressions moyennes d'un autre échantillon (comparable) d'enseignantes. Il faut souligner que les impressions des enseignantes ont été recueillies en utilisant une procédure précise. En particulier, les enseignantes connaissaient toutes très bien la population cible d'élèves (elles enseignaient toutes depuis au moins cinq ans au 2^e cycle du primaire). De plus, elles ont eu l'occasion de

calibrer leurs impressions en cotant d'abord des rappels représentant la diversité des habiletés au sein du sous-échantillon. Finalement, elles ont pu réaliser la tâche dans des conditions relativement favorables, c'est-à-dire pendant leurs temps libres en recevant une compensation équivalant à leur salaire. Il n'est pas évident que des résultats similaires seraient obtenus en utilisant une procédure différente (ex. : en recueillant les impressions d'enseignantes en formation).

Dans un même ordre d'idées, le fait que les quatre enseignantes aient pu s'entendre sur la qualité des rappels des élèves découle probablement en partie des caractéristiques des évaluations. L'auteure de ce mémoire a eu l'occasion d'échanger avec elles après que leurs impressions ont été transmises. Ces échanges informels ont révélé que les enseignantes considéraient que les deux textes étaient pertinents et d'un niveau de difficulté approprié pour des élèves de 3^e année et que le rappel (tel que nous l'avons sollicité) était une bonne façon de déterminer ce que chaque élève « avait vraiment compris ». Il est probable que leurs impressions n'auraient pas convergé de manière aussi claire si les enseignantes avaient douté de la pertinence des évaluations et, par extension, du sens à accorder aux réponses des élèves (c.-à-d. à leur rappel). En d'autres termes, la validité de contenu (un endossement de l'évaluation par les experts) pourrait être une condition nécessaire (mais possiblement insuffisante) à la validité d'utilisateur.

Au-delà du constat général selon lequel il semble utile de considérer des arguments liés à la validité d'utilisateur, il est pertinent d'examiner nos résultats séparément pour les deux textes. Les résultats sont essentiellement sans équivoque pour l'un d'entre eux, *Les dangers de l'océan*. L'analyse factorielle confirmatoire a effectivement permis d'identifier un facteur qui représentait bien la cotation à tous les items retenus et le score dérivé de ce facteur était corrélé avec la fluidité en lecture en plus d'expliquer approximativement 50% de la variance des impressions moyennes des enseignantes. De plus, ce score était distribué de manière relativement normale avec un effet de plancher minimal. Les résultats pour ce texte confirment que la qualité du rappel peut

parfois, comme c'est apparemment le cas ici, refléter de manière adéquate la compréhension globale (c.-à-d. plutôt que d'un aspect en particulier) d'un texte informatif par de jeunes lecteurs.

Le portrait est, par contre, plus complexe pour l'autre texte, *L'hiver*. Dans ce cas, l'analyse factorielle confirmatoire n'a pas permis d'identifier un facteur unique qui aurait résumé de manière adéquate les réponses à l'ensemble des items retenus. L'analyse a plutôt identifié deux facteurs distincts. Même si les scores dérivés de ces deux facteurs sont en lien avec la fluidité en lecture et expliquent, conjointement, un pourcentage substantiel (78%) des impressions moyennes des enseignantes, leurs distributions sont caractérisées par de forts effets de plancher, en particulier en ce qui concerne le score de rappel des conséquences du froid (de l'hiver en Nouvelle-France). Cela signifie que certains aspects de ce texte se sont avérés apparemment trop difficiles pour les élèves de 3^e année de notre échantillon. Les résultats pour ce dernier texte illustrent, par la négative, l'importance d'examiner attentivement les propriétés psychométriques de scores dérivés du rappel plutôt que de simplement se fier sur leur validité apparente. En d'autres termes, même si le rappel est la façon la plus transparente d'évaluer la compréhension selon les modèles théoriques généralement acceptés (ex. : Language and Reading Research Consortium, 2016), et possiblement de l'avis de plusieurs enseignants, il n'est pas garanti qu'un score dérivé du rappel possède des propriétés psychométriques adéquates.

Cette divergence des résultats pour les deux évaluations pourrait être attribuable à une différence subtile dans la structure des textes. Bien qu'il s'agisse de deux textes d'un niveau de lisibilité similaire qui présentent des informations liées au même domaine (l'univers social), un d'entre eux (*Les dangers de l'océan*) décrit les péripéties de personnages historiques en utilisant une trame narrative, alors que l'autre décrit les déterminants et conséquences d'un phénomène non social (climatique, *L'hiver*). Il est possible que les élèves de 3^e année, qui possèdent une connaissance limitée du texte

informatif (ex. : Duke, 2004), aient plus de facilité avec des textes qui ressemblent à certains égards aux textes narratifs qu'ils connaissent mieux, c'est-à-dire des textes qui décrivent les actions et motivations de personnages. Étant donné l'attention portée aux motifs et actions des personnages historiques, l'univers social apparaît, en ce sens, une excellente porte d'entrée pour initier les élèves au texte informatif.

En termes de retombées immédiates, est-ce que les évaluations élaborées dans le cadre de ce mémoire, en particulier celle du texte *Les dangers de l'océan*, sont prêtes à être utilisées? Il serait raisonnable de faire valoir que nous avons démontré la fidélité (cohérence interne) et la validité (d'utilisateur et de construit) de cette dernière évaluation et qu'elle pourrait conséquemment être utilisée dans le cadre d'un projet recherche-intervention. Dans le cadre de ce type de projet, il est important de notamment recourir à des évaluations dites de « transfert proximal » qui mesurent spécifiquement si les élèves ont appris ce qui leur a été enseigné lors de l'intervention expérimentale (Fuchs, Hendricks *et al.*, 2018). De telles évaluations sont le plus souvent élaborées par les chercheurs qui ont créé l'intervention, ce qui soulève des doutes quant à la rigueur ou à l'indépendance du processus d'évaluation. Ces doutes sont légitimes étant donné que la validité (et la fidélité) des évaluations de transfert est rarement démontrée (ex. : Dion *et al.*, 2011; pour une exception, voir Williams *et al.*, 2016). Même si la validité d'une évaluation n'est jamais établie de manière définitive ou absolue (Markus et Borsboom, 2013), nous considérons que la procédure de validation proposée ici représente une option pertinente pour démontrer de manière raisonnablement convaincante la validité d'une évaluation de compréhension dans le cadre d'une recherche-intervention menée auprès d'une population particulière.

Il reste à démontrer, par ailleurs, qu'une évaluation comme celle recourant au texte *Les dangers de l'océan* est prête à être utilisée dans les écoles, au même titre par exemple que les évaluations de suivi des gains de fluidité réalisées par des élèves en difficulté (voir Dion, Roux et Dupéré, 2011, Chap. 5). Dans ce dernier cas, il a été démontré que

les évaluations de gains de fluidité fournissent une information rigoureuse (fidèle et valide), que les praticiens sont en mesure de les utiliser au quotidien (ex. : en disposant de peu de temps pour la correction) et que la cotation est suffisamment transparente pour que les groupes concernés (enseignants et élèves) puissent interpréter les scores (ex. : lorsque ces derniers sont représentés sur un graphique). Dans la mesure où de telles conditions doivent effectivement être rencontrées, il reste à démontrer que le rappel du texte *Les dangers de l'océan* peut être corrigé en temps réel (pendant que l'élève effectue le rappel de texte) dans des conditions normales de pratique et qu'il est possible d'offrir aux enseignants et aux élèves un rationnel convaincant pour la cotation (c.-à-d. le pointage variable d'un item à l'autre). De plus, l'absence de normes fait en sorte que l'évaluation serait probablement d'une utilité limitée pour les psychologues scolaires (qui doivent habituellement déterminer si l'élève démontre des habiletés comparables à celles de ses pairs).

Notons également que nous n'avons pas considéré les arguments de validité de conséquences, c'est-à-dire dans quelle mesure l'utilisation de nos évaluations, et plus spécifiquement, l'interprétation de leurs scores (AERA, APA et NCME, 2014), sert l'intérêt des élèves, par exemple en permettant à l'enseignant d'adapter les activités de lecture offertes en classe. De telles considérations sont évidemment importantes même lorsque le score est susceptible d'être considéré comme pertinent par le milieu comme c'est le cas ici. Dumas et McNeish (2018) avancent, par exemple, que des scores faibles à une évaluation complétée à un seul moment pourraient malheureusement être utilisés par les enseignants pour justifier des attentes limitées par rapport aux capacités d'apprentissage du bassin d'élèves desservi par l'école. Notons que, comme c'est souvent le cas (ex. : Kendeou *et al.*, 2009), les scores à nos évaluations de rappel étaient en moyenne relativement peu élevés.

L'étude présentée dans ce mémoire comporte à la fois des forces et des limites. Parmi ses forces, notons la taille de l'échantillon ($N = 306$) et les particularités de ce dernier,

notamment le fait qu'il a été recruté en milieu défavorisé et qu'il est constitué d'élèves avec des niveaux très diversifiés d'habiletés en lecture (ce qui permet d'éviter les problèmes liés à la restriction de variance). L'équipe de recherche a pu obtenir la collaboration des écoles pour évaluer cet échantillon parce que la cueillette de données s'inscrit dans le cadre d'un projet d'intervention initié en partie à la demande du milieu scolaire (voir Wagener et Dion, 2018). Le fait que les évaluations aient été réalisées en individuel en suivant un protocole détaillé représente également une force de l'étude. Quant aux limites, les conditions d'évaluation n'étaient par ailleurs pas idéales étant donné la situation de surpopulation qui prévaut actuellement dans les écoles, notamment à Montréal. Dans la plupart des cas, les assistants de recherche ont réalisé les évaluations dans le corridor à l'extérieur de la classe, dans un environnement passablement bruyant où un certain nombre de personnes se déplaçaient, ce qui aurait pu distraire les élèves. L'équipe de recherche a tenté de limiter les distractions, avec un certain succès, en utilisant un paravent. Par souci d'uniformité et de simplicité, les assistants ont également offert un soutien minimal lors du rappel (c.-à-d. un maximum de deux encouragements génériques, ex. : « Quelque chose d'autre? ». Avec ce type de soutien, il est presque inévitable que la qualité (et la longueur) du rappel ait été influencée, au moins en partie, par les habiletés verbales des élèves indépendamment, dans une certaine mesure, de leur compréhension du texte. Kendeou et collègues (2009) ont offert un soutien beaucoup plus direct aux élèves lors du rappel (ex. : « Tu m'as parlé de [événement décrit dans le texte mentionné par l'élève]. Est-ce que tu peux me dire ce qui est arrivé avant? »). Il serait utile d'explorer l'effet d'un tel protocole sur la qualité du rappel après avoir démontré que les assistants de recherche étaient effectivement capables de l'implanter de manière fidèle (ce que Kendeou *et al.*, 2009, n'ont pas fait). Finalement, notons que nous n'avons pas démontré que nos évaluations étaient sensibles aux apprentissages (ex. : suite à une intervention), une limite partagée par plusieurs mesures de compréhension reposant sur le rappel (Reed et Vaughn, 2012). Il est envisageable qu'une évaluation possède une bonne validité d'utilisateur tout en

étant insensible, par exemple, aux progrès réalisés par les élèves en cours d'année scolaire.

Malgré ces limites, nous pensons avoir démontré que l'examen de la validité d'utilisateur est une approche utile et peu exigeante en termes logistiques pour explorer la pertinence d'évaluations de la compréhension en lecture. Il est possible, en particulier, de recueillir les impressions d'enseignants en peu de temps et à relativement peu de frais et, ce faisant, de montrer que les chercheurs et les praticiens interprètent généralement les scores de la même façon. Comme nous l'avons noté en introduction, une interprétation commune des scores est susceptible de faciliter le transfert de connaissances. En fait, l'approche de validation que nous proposons pourrait répondre à des besoins importants en éducation et nous considérons que son potentiel doit être exploré davantage. Il serait possible, par exemple, d'examiner la validité d'utilisateur d'évaluations reposant sur une cotation relativement complexe de la performance des élèves, par exemple de la qualité de leurs compositions écrites (ex. : De Smedt *et al.*, 2020) ou de leurs présentations orales (enregistrées audionumériquement).

ANNEXE A

LES DANGERS DE L'OCÉAN, VERSION ORIGINALE

Nous partîmes d'Honfleur le premier jour de mars, avec un vent favorable jusqu'au 8 du mois, et ensuite nous fûmes contrariés par le vent de sud-sud-ouest et ouest-nord-ouest qui nous fit aller jusqu'à la hauteur de 42 degrés de latitude, sans pouvoir nous orienter vers le sud, pour nous mettre sur le droit chemin de notre route. Après avoir donc subi plusieurs coups de vent et avoir été contrariés par le mauvais temps, avoir néanmoins encore connu de grandes peines et difficultés à force de tenir d'un bord à l'autre, nous fîmes en sorte d'arriver à environ quatre-vingts lieues du grand banc où se fait la pêche du poisson vert. Là, nous rencontrâmes des glaces de plus de trente à quarante brasses de haut, ce qui nous fit bien réfléchir à ce que nous devons faire : nous craignons d'en rencontrer d'autres la nuit et d'être poussés contre elles si le vent venait à changer, et nous jugions bien aussi que ce ne serait pas les derniers, car nous étions partis de trop bonne heure de France.

Alors que nous avions donc navigué tout au long de ce même jour à basse voile-le plus près du vent que nous le pouvions-, la nuit venue, il se leva une brume si épaisse et si obscure qu'à peine pouvions-nous voir la longueur du vaisseau. Environ sur les onze heures de la nuit, les matelots avisèrent d'autres glaces qui nous donnèrent de l'appréhension, mais enfin la diligence des mariniers fut telle que nous les évitâmes. Alors que nous pensions avoir passé les dangers, nous vîmes à rencontrer une autre glace devant notre vaisseau, que les matelots aperçurent si tard que nous étions presque dessus. Et comme chacun se recommandait à Dieu, ne pensant jamais pouvoir éviter cette glace qui était sous notre beaupré, l'on criait à celui qui tenait le gouvernail qu'il fit porter. En effet, cette glace, qui était fort grande, dérivait au vent de telle façon qu'elle passa contre le bord de notre vaisseau, qui s'arrêta net, comme s'il n'avait pas bougé pour la laisser passer sans oser l'endommager. Et bien que nous fussions hors

de danger, le sang de chacun n'en fut pas aisément apaisé, vu la frayeur que nous avions eue. Nous louâmes Dieu de nous avoir délivrés de ce péril.

ANNEXE B

LES DANGERS DE L'OCÉAN, VERSION ADAPTÉE

Les dangers de l'océan

Nous avons quitté la France le 1er mars. Pendant huit jours, le vent a soufflé dans la bonne direction. Ensuite, des mauvais vents ont poussé notre bateau trop au sud.

Nous avons réussi à remonter vers le nord. Nous avons commencé à apercevoir d'immenses icebergs. Ça nous a fait réfléchir.

Il y aurait sûrement d'autres icebergs sur l'océan parce que nous n'avions pas attendu la fin de l'hiver avant de partir. Pendant la nuit, nous ne pourrions pas voir les icebergs à cause de la noirceur.

La nuit venue, une brume se leva. La brume était si épaisse qu'à un bout du bateau, nous pouvions à peine voir l'autre bout.

À la dernière minute, nous avons aperçu un iceberg. Le vent l'a poussé et il est venu se frotter sur notre bateau. Notre bateau aurait pu se briser.

Heureusement, l'iceberg s'est un peu éloigné, comme pour laisser passer notre bateau. Nous avons eu besoin d'un long moment pour réussir à nous calmer parce que nous avons eu très peur.

ANNEXE C

L'HIVER

L'hiver

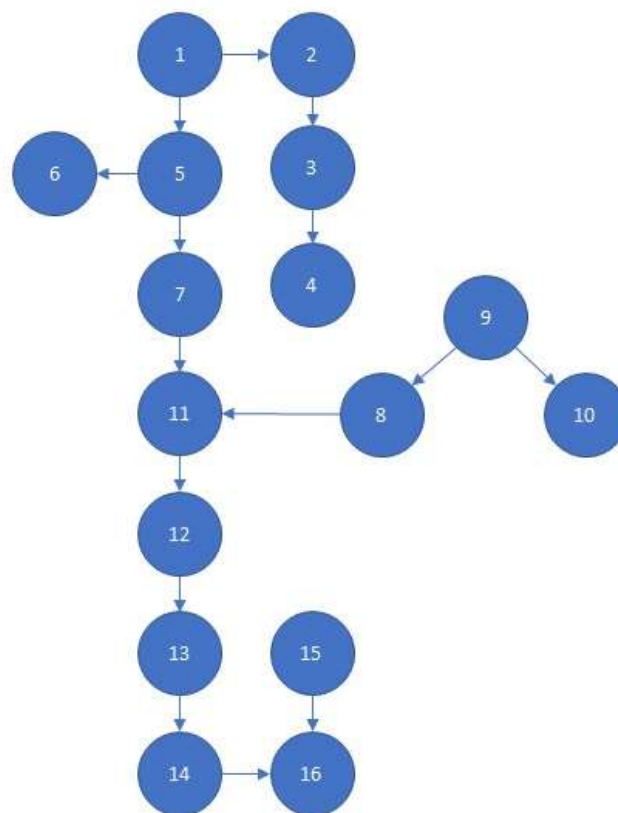
Samuel de Champlain a déjà passé des hivers en Nouvelle-France. Il sait que l'hiver est son plus terrible ennemi. En hiver, tout est gelé pendant plusieurs mois.

La Nouvelle-France devient un immense bloc de glace. À cause du froid, il est difficile de se déplacer. Les canots ne peuvent plus avancer sur le fleuve. À cause du froid, les aliments ne peuvent pas pousser. Les plantes ont besoin de chaleur.

Champlain ne comprend pas pourquoi les hivers de la Nouvelle-France sont beaucoup plus froids que ceux de la France. La Nouvelle-France n'est pas plus près du Pôle Nord que la France.

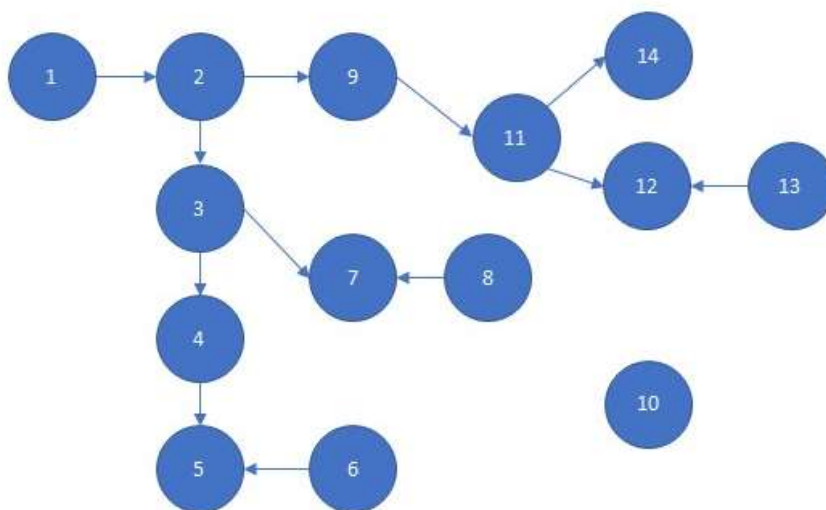
Nous avons maintenant réussi à comprendre ce qui arrive. L'hiver de la Nouvelle-France est très froid à cause de la direction des vents. Ces vents froids arrivent souvent du nord. La France est réchauffée par des vents chauds du sud.

ANNEXE D

ANALYSE EN PROPOSITIONS DES *DANGERS DE L'OCÉAN*

Analyse en propositions du texte *Les dangers de l'océan* selon les critères proposés par Trabasso *et al.* (1989). La présence d'un lien (flèche) indique que l'action décrite dans la proposition subséquente ne se serait pas produite en l'absence de l'action décrite dans la proposition précédente. 1. « Départ de la France », 2. « Le vent souffle dans la bonne direction », 3. « De mauvais vents poussent le bateau au sud. », 4. « Réussissent à remonter vers le nord », 5. « Aperçoivent des icebergs », 6. « Ça les fait réfléchir », 7. « D'autres icebergs sont attendus », 8. « La nuit empêche de voir les icebergs », 9. « La nuit, une brume se lève », 10. « La brume empêche de voir l'autre bout du bateau », 11. « À la dernière minute, un iceberg est aperçu », 12. « Le vent pousse l'iceberg sur le bateau », 13. « Le bateau manque de se briser », 14. « L'iceberg s'éloigne du bateau. », 15. « Ils ont besoin d'un long moment pour se calmer », 16. « parce qu'ils ont eu très peur »

ANNEXE E

ANALYSE EN PROPOSITIONS DE *L'HIVER*

Analyse en propositions du texte *L'hiver* selon les critères proposés par Trabasso *et al.* (1989). La présence d'un lien (flèche) indique que l'action décrite dans la proposition subséquente ne se serait pas produite en l'absence de l'action décrite dans la proposition précédente. 1. « Champlain a déjà passé des hivers en Nouvelle-France », 2. « L'hiver est son plus terrible ennemi », 3. « En hiver, tout est gelé », 4. « La Nouvelle-France est un immense bloc de glace », 5. « À cause du froid, il est difficile de se déplacer », 6. « Les canots ne peuvent plus avancer », 7. « Les aliments ne peuvent pas pousser », 8. « Les plantes ont besoin de chaleur », 9. « Champlain ne comprend pas pourquoi les hivers de la Nouvelle-France sont beaucoup plus froids que ceux de la France », 10. « La Nouvelle-France n'est pas plus près du Pôle Nord que la France », 11. « Réussissent à comprendre ce qui arrive », 12. « L'hiver de la Nouvelle-France est très froid à cause des vents », 13. « Les vents froids arrivent souvent du nord », 14. « La France est réchauffée par des vents chauds du sud »

ANNEXE F

FORMULAIRE DE CONSENTEMENT DES ENSEIGNANTS EXPERTS



Enseignant

FORMULAIRE D'INFORMATION ET DE CONSENTEMENT

Titre du projet de recherche : *Lire à deux pour mieux comprendre les textes informatifs*

Chercheur responsable : Eric Dion, Ph.D., Université du Québec à Montréal

Membres de l'équipe : Véronique Dupéré, Ph.D., Université de Montréal

Marc-André Éthier, Ph.D., Université de Montréal

David Lefrançois, Ph.D., Université du Québec en Outaouais

Douglas Fuchs, Ph.D., Université Vanderbilt

Coordonnatrice : Audrey Wagener

Organisme de financement : Conseil de recherche en sciences humaines du Canada

Préambule

Nous vous invitons, cher enseignant, à participer à un projet de recherche.

Il est important de bien lire ce qui suit. Si ce n'est pas clair ou si vous avez des questions, n'hésitez pas à nous les poser (voir coordonnées d'Eric Dion à la fin du formulaire).

Objectifs du projet

Nous avons évalué la compréhension du texte informatif d'élèves de 3^e année en utilisant le rappel. Nous voudrions nous assurer que nous avons corrigé ces évaluations comme le feraient des enseignants.

Nature de la participation

Nous aimerions donc que vous lisiez la transcription de 150 courts rappels d'élèves (d'habituellement une vingtaine de mots) et donniez un score à chacun de ces rappels

sur une échelle globale. Vous pouvez faire ce travail à domicile, au rythme qui vous convient. Normalement, ce travail de correction devrait vous prendre environ quatre heures.

Avantages et inconvénients.

Participer au projet ne présente pas vraiment d'avantages pour vous. Il est possible par ailleurs que notre façon d'évaluer la compréhension vous donne des idées pour évaluer vos propres élèves. À part le temps que vous aurez à consacrer à la correction, la participation ne présente par ailleurs pas d'inconvénients.

Compensation

Vous serez compensé l'équivalent de 30\$ de l'heure pour réaliser le travail de correction, c'est-à-dire 120\$ au total si la correction prend 4 heures comme prévu.

Confidentialité

Les rappels que nous allons vous faire parvenir auront été anonymisés, c'est-à-dire que le nom des élèves n'apparaîtra nulle part. De plus, votre nom ne sera jamais divulgué lors de la présentation des résultats de la recherche.

Les données recueillies dans le cadre du projet seront détruites cinq ans après la publication des derniers articles scientifiques, thèse ou mémoire.

Participation volontaire et droit de retrait

Votre participation est entièrement volontaire. Vous pouvez vous retirer du projet (cesser les corrections) en tout temps. Même si vous ne terminez pas les corrections, nous allons vous offrir la compensation pour les heures travaillées.

Personnes-ressources :

Vous pouvez contacter le responsable du projet, Eric Dion, au numéro (514) 528-9358 ou à l'adresse courriel dion.e@uqam.ca, pour des questions additionnelles sur le projet.

Le Comité institutionnel d'éthique de la recherche avec des êtres humains (CIEREH) a approuvé ce projet et en assure le suivi. Pour toute information vous pouvez communiquer avec la coordonnatrice du Comité au numéro 987-3000 poste 7753 ou par courriel à l'adresse : ciereh@uqam.ca. Pour toute question concernant vos droits en tant que participant à ce projet de recherche ou si vous avez des plaintes à formuler, vous pouvez communiquer avec le bureau de l'ombudsman de l'UQAM (Courriel: ombudsman@uqam.ca; Téléphone: (514) 987-3151.

Consentement du participant : Par la présente, je reconnais avoir lu le présent formulaire d'information et de consentement. Je comprends les objectifs du projet et ce que ma participation implique. Je confirme avoir disposé du temps nécessaire pour réfléchir à ma décision de participer. Je reconnais avoir eu la possibilité de contacter le

responsable du projet afin de poser toutes les questions concernant ma participation et que l'on m'a répondu de manière satisfaisante. Je comprends que je peux me retirer du projet en tout temps, sans pénalité d'aucune forme, ni justification à donner. Je consens volontairement à participer à ce projet de recherche.

Déclaration du chercheur principal :

Je, soussigné, déclare avoir expliqué les objectifs, la nature, les avantages, les risques du projet et autres dispositions du formulaire d'information et de consentement et avoir répondu au meilleur de ma connaissance aux questions posées.



Eric Dion, Ph.D.
Département d'éducation et formation spécialisées
Université du Québec à Montréal

Votre signature :

Date :

Votre nom (lettres moulées) :

APPENDICE A

LETTRE DE CONSENTEMENT PARENTAL

Parent



FORMULAIRE D'INFORMATION ET DE CONSENTEMENT

- Titre du projet:** *Mieux comprendre les textes informatifs*
- Chercheur responsable :** *Eric Dion, Université du Québec à Montréal*
- Membres de l'équipe :** *Véronique Dupéré, et Marc-André Éthier, Université de Montréal, David Lefrançois, Université du Québec en Outaouais, Douglas Fuchs, Université Vanderbilt*
- Coordonnatrice :** Audrey Wagener
- Financement :** Conseil de recherche en sciences humaines du Canada

Nous invitons votre enfant à participer à un projet qui va se dérouler dans sa classe.
Si vous avez des questions après avoir lu ce qui suit, n'hésitez pas à communiquer avec Eric Dion, le responsable du projet.

Notre objectif

Nous voulons trouver la meilleure façon d'aider les élèves à comprendre les textes informatifs (ex. : qui expliquent pourquoi il fait froid en hiver).

Cette année, tous les élèves de la classe de (nom du professeur) vont lire ensemble des textes clairs et intéressants sur les débuts de la Nouvelle-France (Canada).

Plusieurs classes participent et la façon de lire les textes va être différente d'une classe à l'autre. Dans certaines, les élèves vont discuter des textes. Dans d'autres, ils vont faire des liens entre les textes ou examiner attentivement les explications dans les textes.

La participation de votre enfant

Pour trouver la meilleure façon d'aider les élèves à comprendre, nous devons comparer leurs progrès dans les différentes classes. C'est pour cette raison que nous aimerions que votre enfant participe.

Si vous acceptez qu'il participe, une assistante, (nom de l'assistant), va évaluer rapidement (5 minutes) sa lecture. Elle va ensuite choisir dans la classe 9 élèves avec des habiletés de lecture différentes.

Si votre enfant fait partie de ses 9 élèves, (nom de l'assistant) va faire des évaluations plus complètes de sa lecture, au début de l'année (20 minutes), à huit reprises pendant l'année (5 minutes) et une dernière fois à la fin de l'année (20 minutes).

Nous allons aussi demander à (nom du professeur) de remplir un bref questionnaire sur l'attention de votre enfant en classe.

Avantages, inconvénients et compensation

(nom de l'assistant), l'assistante, va souvent venir en classe. Votre enfant devrait être à l'aise avec elle. Si jamais il semble stressé ou s'il demande d'arrêter les évaluations, elle va le faire.

Pour souligner ses efforts, (nom de l'assistant) va lui donner quelques autocollants.

Confidentialité

Les résultats de votre enfant vont servir seulement à comparer les façons de lire en classe. Son nom ne sera pas mentionné nulle part.

Nous allons conserver ses évaluations en sécurité et les détruire 5 ans après la fin du projet. Elles n'influenceront pas ses notes au bulletin.

Vos droits et ceux de votre enfant

Si vous acceptez que votre enfant participe, vous ne renoncez à aucun de ses droits et ne libérez les chercheurs et leurs institutions d'aucune obligation civile et professionnelle.

La participation de votre enfant est complètement volontaire et même si vous acceptez aujourd'hui qu'il participe, vous pouvez changer d'idée n'importe quand, sans avoir à vous justifier.

Peu importe votre décision, votre enfant ne sera pas pénalisé. Si vous décidez de le retirer, nous allons détruire tous les renseignements le concernant.

Si vous avez des questions :

Si certains aspects du projet ne vous semblent pas clairs, vous pouvez contacter le responsable du projet, Eric Dion. Courriel : dion.e@uqam.ca; Téléphone : (514) 987-3000 poste 4970.

Vous pouvez aussi contacter le comité d'éthique de la recherche qui a approuvé ce projet. Courriel : ciereh@uqam.ca; Téléphone : (514) 987-3000 poste 7753.

Pour des questions sur vos droits ou pour porter plainte, vous pouvez finalement communiquer avec l'ombudsman de l'UQAM. Courriel : ombudsman@uqam.ca; Téléphone: (514) 987-3151.

Signature parent/tuteur légal : Je comprends que la participation de mon enfant est volontaire et que je peux ou qu'il peut y mettre fin en tout temps, sans avoir à me justifier et sans être pénalisé. Je consens à ce qu'il participe à ce projet.

Nom de mon enfant : _____


Ma signature :

Date :

Déclaration du chercheur principal :

Je, soussigné, déclare avoir expliqué les objectifs, la nature, les avantages, les risques du projet et autres dispositions du formulaire d'information et de consentement et avoir répondu au meilleur de ma connaissance aux questions posées.

Signature :



Date : 16 octobre 2018

Nom et coordonnées : Eric Dion, (514) 987-3000, poste 4970, dion.e@uqam.ca

Un exemplaire de ce document signé doit être remis au représentant légal de l'enfant

RÉFÉRENCES

- Afia, K., Dion, E., Dupéré, V., Archambault, I., et Toste, J. (2019). Parenting practices during middle adolescence and high school dropout. *Journal of Adolescence*, 76, 55-64.
- Akin-Little, K. A., Eckert, T. L., Lovett, B. J., et Little, S. G. (2004). Extrinsic reinforcement in the classroom: Bribery or best practice. *School Psychology Review*, 33(3), 344-362.
- American Educational Research Association [AERA], American Psychological Association [APA] et National Council on Measurement in Education [NCME]. (2014). *Standards for educational and psychological testing*. Washington, DC : American Educational Research Association.
- Arcand, M.S., Dion, E., Lemire-Théberge, L., Guay, M.H., Barrette, A., Gagnon, V., Caron, P.-O. et Fuchs, D. (2014). Segmenting texts into meaningful word groups : Beginning readers' prosody and comprehension. *Scientific Studies of Reading*, 18(3), 208-223.
- Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY : Guilford Press.
- Bagnato, S. J., Goins, D. D., Pretti-Frontczak, K., et Neisworth, J. T. (2014). Authentic assessment as “best practice” for early childhood intervention: National consumer social validity research. *Topics in Early Childhood Special Education*, 34(2), 116-127.
- Beker, K., van den Broek, P. et Jolles, D. (2019). Children's integration of information across texts: reading processes and knowledge representations. *Reading and Writing*, 32(3), 663–687.

- Berger, M. J., et Desrochers, A. (2011). *L'évaluation de la littératie*. Ottawa, ON : Les Presses de l'Université d'Ottawa.
- Bollen, K. A. et Diamantopoulos, A. (2017). In defence of causal-formative indicators : A minority report. *Psychological Methods*, 22, 581-596.
- Bollen, K.A., et Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, 110(2), 305-314.
- Carpenter R. G., et Paris, S. G. (2005). Issues of Validity and Reliability in Early Reading Assessments. Dans S. G. Paris et S. A. Stahl (dir.), *Children's reading comprehension and assessment* (p. 279-304). Mahwah, NJ : Lawrence Erlbaum Associates Publishers.
- Chall, J.S. et Jacobs, V.A. (2003). The classic study on poor children's fourth-grade slump. *American Educator*, 27(1), 14-15.
- Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., et Delbridge, K. (1997). Reactions to cognitive ability tests: the relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology*, 82(2), 300-310.
- Chen, F.F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, 14, 464-504.
- Cheung, G. W. et Rensvold, R. B. (2002). Evaluating goodness-of fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 233-255.
- Clinton, V., et van den Broek, P. (2012). Interest, inferences, and learning from texts. *Learning and Individual Differences*, 22, 650-663.
- Coan, J. A., et Gottman, J. M. (2007). The Specific Affect (SPAFF) coding system. Dans J. A. Coan et J. J. B. Allen (dir.), *Handbook of Emotion Elicitation and Assessment* (p.106-123). New York, NY: Oxford University Press.
- Cormier, P., Desrochers, A., et Sénéchal, M. (2006). Validation et consistance interne d'une batterie de tests pour l'évaluation multidimensionnelle de la lecture en français. *Revue des sciences de l'éducation*, 32(1), 205-225.

- Coté, N., Goldman, S. R., et Saul, E. U. (1998). Students making sense of informational text: Relations between processing and representation. *Discourse Processes*, 25(1), 1-53.
- Crocker, L. et Algina, J. (2008). *Introduction to classical & modern test theory*. Mason, OH : Cenpage Learning.
- Cronbach, L. J., et Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281-302.
- Daoust, F., Laroche, L. et Ouellet, L. (1996). SATO-Calibrage: Présentation d'un outil d'assistance au choix et à la rédaction de textes pour l'enseignement. *Revue québécoise de linguistique*, 25(1), 205-234.
- De Smedt, F. Graham, S. et Van Keer, H. (2020). "It takes two": The added value of structured peer-assisted writing in explicit writing instruction. *Contemporary Educational Psychology*, 60. Prépublication disponible en ligne.
- Desrochers, A. (2008). *The assessment of reading skills among French speaking children: Test construction procedure and psychometric properties*. Ottawa, ON : Cognitive Psychology of Language Laboratory, Université d'Ottawa.
- Desrochers, A. et Saint-Aubin, J. (2008). Sources de matériel en français pour l'élaboration d'épreuves de compétences en lecture et en écriture. *Canadian Journal of Education/revue canadienne de l'éducation*, 31(2), 305-326.
- Desrochers, A., Simon, M., Thompson, G. L. (2011). Formes et fonctions de l'évaluation de la littératie. Dans M. J. Berger et A. Desrochers (dir.), *L'évaluation de la littératie* (p. 29-62). Ottawa, ON : Presses de l'Université d'Ottawa.
- Dion, E., Cloutier, S., Éthier, M.-A., et Lefrançois, D. (2018). *Apprendre à lire à deux. Activité lecture et écriture de 3e année*. Document non-publié. Université du Québec à Montréal. Montréal, Qc.
- Dion, E., Roux, C., et Dupéré, V. (2011). Utilisation et développement des mesures de progrès en lecture. Dans M. J. Berger et A. Desrochers (dir.), *L'évaluation de la littératie* (p. 117-137). Ottawa, ON : Presses de l'Université d'Ottawa.

- Dion, E., Roux, C., Landry, D., Fuchs, D., Wehby, J., et Dupéré, V. (2011). Improving attention and preventing reading difficulties among low-income first-graders: A randomized study. *Prevention Science, 12*(1), 70-79.
- Duke, N. K. (2004). The case for informational text. *Educational Leadership, 61*(6), 40-45.
- Dumas, D. G. et McNeish, D. M. (2018). Increasing the consequential validity of reading assessment using dynamic measurement modeling: A comment on Dumas and McNeish (2017). *Educational Researcher, 47*(9), 612-614.
- Edwards, J. R. (2011). The fallacy of formative measurement. *Organizational Research Methods, 14*, 370-388.
- Enders, C. K. (2010). *Applied missing data analysis*. New York, NY : Guilford Press.
- Fleiss, J. L., Levin, B. et Cho Paik, M. (2003). *Statistical methods for rates and proportions* (3e éd.). Hoboken, NJ: Wiley.
- Fletcher, J. M. (2006). Measuring reading comprehension. *Scientific Studies of Reading, 10*(3), 323-330.
- Franzen, M.D. (2011). Item analysis. Dans J.S. Kreutzer, J. DeLuca et B. Caplan (dir.), *Encyclopedia of Clinical Neuropsychology*. New York, NY : Springer.
- Froese-Germain, B. (1999). *Les tests standardisés: atteinte à l'équité en éducation*. [Rapport préparé dans le cadre de la campagne des dossiers nationaux en éducation]. Ottawa, ON : Canadian Teachers Federation.
- Fuchs, D. et Fuchs, L. S. (1986). Test procedure bias: a meta-analysis of examiner familiarity effects. *Review of Educational Research, 56*(2), 243-262.
- Fuchs, L. S., Fuchs, D., Hosp, M. K., et Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*(3), 239-256.

- Fuchs, L. S., Gilbert, J. K., Fuchs, D., Seethaler, P. M., et N. Martin, B. (2018). Text comprehension and oral language as predictors of word-problem solving: Insights into word-problem solving as a form of text comprehension. *Scientific Studies of Reading*, 22(2), 152-166.
- Fuchs, D., Hendricks, E., Walsh, M.E, Fuchs, L.S., Gilbert, J. K., Zhang Tracy, W.,... et Peng, P. (2018). Evaluating a Multidimensional Reading Comprehension Program and Reconsidering the Lowly Reputation of tests of Near-Transfer. *Learning Disabilities Research & Practice*, 33(1), 11-23.
- Galton, F. (1890). Exhibition of instruments (1) for testing perception of differences of tint, and (2) for determining reaction-time. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 19, 27-29.
- Gough, P. B., et Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and special education*, 7(1), 6-10.
- Gould, S. J. (1981). *The mismeasure of man*. New York, NY : Norton & Company.
- Graesser, A. C., Singer, M., et Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, 101(3), 371-395.
- Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability. *Educational and Psychological Measurement*, 66(6), 930-944.
- Hirsch, E. D., Jr. (2003). Reading comprehension requires knowledge—of words and the world: Scientific insights into the fourth-grade slump and the nation's stagnant comprehension scores. *American Educator*, 27(1), 10–29.
- Hoover, W. A., et Tunmer, W. E. (2018). The simple view of reading: Three assessments of its adequacy. *Remedial and Special Education*, 39(5), 304-312.
- Hosp, M. K., et Fuchs, L.S. (2005). Using CBM as an indicator of decoding, word reading, and comprehension: Do the relations change with grade? *School Psychology Review*, 34(1), 9-26.
- Howell, R. D. (2014). What is the latent variable in causal indicators models? *Measurement: Interdisciplinary Research and Perspectives*, 12(4), 141-145.

- Hu, L.T., et Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1-55.
- Irwin, J. (2007). *Teaching reading comprehension processes* (3^e éd.). Boston, MA : Allyn and Bacon.
- Johnston, P. (1984). Prior knowledge and reading comprehension test bias. *Reading Research Quarterly*, 19(2), 219-239.
- Keenan, J. M., Betjemann, R. S., et Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, 12(3), 281-300.
- Kendeou, P., Van den Broek, P., White, M. J., et Lynch, J. S. (2009). Predicting reading comprehension in early elementary school: The independent contributions of oral language and decoding skills. *Journal of Educational Psychology*, 101(4), 765-778.
- Kintsch, W. (1986). Learning from text. *Cognition and instruction*, 3(2), 87-108.
- Kintsch, W. et Kintsch, E. (2005). Comprehension. Dans S. G. Paris et S. A. Stahl (dir.), *Children's reading comprehension and assessment* (p. 71-92). Mahwah, NJ : Lawrence Erlbaum Associates Publishers.
- Kintsch, W., et van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological review*, 85(5), 363-394.
- Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3^e éd.). New York: Guilford Press.
- Klingner, J. K. (2004). Assessing reading comprehension. *Assessment for effective Intervention*, 29(4), 59-70.
- Krokoff, L. J., Gottman, J. M., et Hass, S. D. (1989). Validation of a global rapid couples interaction scoring system. *Behavioral Assessment*, 11(1), 65-79.
- Language and Reading Research Consortium (2015). Learning to read: Should we keep things simple? *Reading Research Quarterly*, 50(2), 151-169.

- Language and Reading Research Consortium (2016). Use of the Curriculum Research Framework (CRF) for developing a reading-comprehension curricular supplement for the primary grades. *Elementary School Journal*, 116(3), 459-486.
- Laurier, M., Tousignant, R., et Morissette, D. (2005). *Les principes de la mesure et de l'évaluation des apprentissages*. Montréal, QC : Gaëtan Morin.
- Laveault, D., et Grégoire, J. (2014). *Introduction aux théories des tests en éducation et en psychologie* (3^e éd.). Bruxelles : DeBoeck-Université.
- Linderholm, T., et van den Broek, P. (2002). The effects of reading purpose and working memory capacity on the processing of expository text. *Journal of Educational Psychology*, 94(4), 778-784.
- Lorch Jr, R. F., et van den Broek, P. (1997). Understanding reading comprehension: Current and future contributions of cognitive science. *Contemporary Educational Psychology*, 22(2), 213-246.
- Lorenzo-Seva, U., et Ferrando, P. J. (2012). TETRA-COM: A comprehensive SPSS program for estimating the tetrachoric correlation. *Behavior Research Methods*, 44(4), 1191-1196.
- Maeder, C. (2010). *La forme noire : Test de compréhension écrite de récits 9-12 ans*. Ortho édition.
- Mandler, J. M. et Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology*, 9(1), 111-151.
- Markus, K. A. et Borsboom, D. (2013). *Frontiers of test validity tests theory: measurement, causation, and meaning*. New York: Routledge.
- Marrache-Gouraud, M. (2010). *Samuel de Champlain, Voyages au Canada*. Paris : Gallimard.
- Marsh, H.W., Hau, K.-T., et Grayson, D. (2005). *Goodness of fit in structural equation models*. Dans A. Maydeu-Olivares et J. McArdle (dir.), *Contemporary*

Psychometric (p.275-340). Mahwah, NJ : Laurence Erlbaum Associates Publishers.

McAuley, E., Duncan, T., et Tammen, V. V. (1989). Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for Exercise and Sport*, 60(1), 48-58.

McGraw, K. O. et Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30-46.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from there. *Psychological Methods*, 23(3), 412-433.

Meyer, B. J. et Ray, M. N. (2011). Structure strategy interventions: Increasing reading comprehension of expository text. *International Electronic Journal of Elementary Education*, 4(1), 127-152.

Michaud, M., Dion, E., Barrette, A., Dupéré, V., et Toste, J. (2017). Does knowing what a word means influence how easily its decoding is learned? *Reading & Writing Quarterly*, 33(1), 82-96.

Ministère de l'Éducation et de l'Enseignement supérieur du Québec (2011). *Cadre d'évaluation des apprentissages Français, langue d'enseignement, Enseignement primaire 1^{er}, 2^e et 3^e cycle du primaire*.

Ministère de l'Éducation et de l'Enseignement supérieur du Québec (2018). *Indice de défavorisation par école 2018-2019*. Québec : Gouvernement du Québec.

Monetta, L. (2015). Fiches descriptives des outils validés et/ou normés en franco-québécois pour l'évaluation du langage et de la parole, de 1980 à 2014. *REPAR*. Récupéré de <http://www.repar.veille.qc.ca/fichier.php/92/Fiches%20descriptives%20orthophonie%20.pdf>

Morgan, P. L., Fuchs, D., Compton, D. L., Cordray, D. S., et Fuchs, L. S. (2008). Does early reading failure decrease children's reading motivation? *Journal of Learning Disabilities*, 41(5), 387-404.

- Morin, A.J.S., Arens, A.K., et Marsh, H.W. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling*, 23(1), 116-139.
- Morrow, L.M. (1988). Retelling stories as a diagnostic tool. Dans S.M. Glazer, L. W. Searfoss et L.M. Gentile (dir.), *Reexamining reading diagnosis: New trends and procedures* (p. 128-149). Newark, NJ : International Reading Association.
- Muthén, L.K. et Muthén, B.O. (2019). *Mplus User's Guide*. Muthén & Muthén.
- Nelligan, E. (1998). *Poésies complètes : Nelligan*. Montréal : Typo.
- Parmar, R. S., Frazita, R., et Cawley, J. F. (1996). Mathematics assessment for students with mild disabilities: An exploration of content validity. *Learning Disability Quarterly*, 19(2), 127-136.
- Pearson, P. D., et Hamm, D. N. (2005). The assessment of reading comprehension: A review of practices: past, present, and future. Dans S. G. Paris et S. A. Stahl (dir.), *Children's reading comprehension and assessment* (p. 13-69). Mahwah, NJ : Lawrence Erlbaum Associates Publishers.
- Perfetti, C. A., Landi, N. et Oakhill, J. (2005). The acquisition of reading comprehension skill. Dans M.J. Snowling et C. Hulme (dir.), *The science of reading: A handbook* (p. 227-247). Hoboken, NJ : Blackwell Publishing.
- PIRLS. Progress in International Reading Literacy Study (2017). *PIRLS 2016 International Results in Reading*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.
- RAND. Reading Study Group (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: Rand.
- Rawson, K. A., et Kintsch, W. (2005). Rereading effects depend on time of test. *Journal of Educational Psychology*, 97(1), 70-80.
- Reed, D. K., et Vaughn, S. (2012). Retell as an indicator of reading comprehension. *Scientific Studies of Reading*, 16(3), 187-217.

- Reiss, S. (2005). Extrinsic and intrinsic motivation at 30: Unresolved scientific issues. *The Behavior Analyst*, 28(1), 1-14.
- Riser S. C., et Lozier M. S. (2013). Rethinking the Gulf Stream. *Scientific American*, 308(2), 50-55.
- Roux, C., Dion, E., Barrette, A., Dupéré, V., et Fuchs, D. (2015). Efficacy of an intervention to enhance reading comprehension of students with high-functioning autism spectrum disorder. *Remedial and Special Education*, 36(3), 131-142.
- Rowling, J.K. (2016). *Harry Potter: La Collection Complète (1-7)*. Londres : Pottermore.
- Rubow, C. C., Vollmer, T. R., et Joslyn, P. R. (2018). Effects of the good behavior game on student and teacher behavior in an alternative school. *Journal of Applied Behavior Analysis*, 51(2), 382-392.
- Sarrazin, G. (1995). Test de rendement pour francophones, adaptation canadienne. Toronto: The Psychological Corporation.
- Sartori, R. (2010). Face validity in personality tests: psychometric instruments and projective techniques in comparison. *Quality & Quantity*, 44, 749-759.
- Shields, M., Connor Gorber, S., et Tremblay, M. S. (2008). Effets des mesures sur l'obésité et la morbidité, *Rapports sur la santé*, 19(2), 87-95.
- Simon, M. (2011). Les qualités d'un instrument d'évaluation de la littératie. Dans M. J. Berger et A. Desrochers (dir.), *L'évaluation de la littératie* (p. 287-314). Ottawa, ON : Presses de l'Université d'Ottawa.
- Snodgrass, M. R., Chung, M. Y., Meadan, H., et Halle, J. W. (2018). Social validity in single-case research: A systematic literature review of prevalence and application. *Research in Developmental Disabilities*, 74, 160-173.
- Solórzano, R. W. (2008). High stakes testing: Issues, implications, and remedies for English language learners. *Review of Educational Research*, 78(2), 260-329.

- Statistique Canada. 2009. Enquête canadienne sur les mesures de la santé (ECMS), cycle 1, 2007 à 2009 [Questionnaire auprès des ménages]. Récupéré de http://www.statcan.gc.ca/imdb-bmdi/instrument/5071_Q1_V1-fra.pdf
- Taleb, N. N. (2009). *Le hasard sauvage: comment la chance nous trompe*. Les Belles Lettres. Paris : Les Belles Lettres.
- Tolar, T. D., Barth, A. E., Francis, D. J., Fletcher, J. M., Stuebing, K. K., et Vaughn, S. (2012). Psychometric properties of maze tasks in middle school students. *Assessment for Effective Intervention*, 37(3), 131-146.
- Trabasso, T., Van den Broek, P., et Suh, S. Y. (1989). Logical necessity and transitivity of causal relations in stories. *Discourse Processes*, 12(1), 1-25.
- van den Broek, P. W., et Kendeou, P. (2017). Development of reading comprehension: Change and continuity in the ability to construct coherent representations. Dans K. Cain, D. L. Compton et R. K. Parrila (dir.), *Theories of reading development (Vol. 15)* (p.283-308). Philadelphia, PA : John Benjamins Publishing Company.
- van den Broek, P., Kendeou, P., Lousberg, S., et Visser, G. (2011). Preparing for reading comprehension: Fostering text comprehension skills in preschool and early elementary school children. *International Electronic Journal of Elementary Education*, 4(1), 259-268.
- van den Broek, P. W., White, M. J., Kendeou, P., et Carlson, S. (2009). Reading between the lines: Developmental and individual differences in cognitive processes in reading comprehension. Dans R. K. Wagner, C. Schatschneider et C. Phythian-Sence (dir.), *Beyond decoding: The behavioral and biological foundations of reading comprehension* (p. 107-123). New York, NY : Guilford Press.
- van den Broek, P., Young, M., Tzeng, Y., et Linderholm, T. (1999). The Landscape model of reading: Inferences and the online construction of memory representation. Dans H. van Oostendorp et S. R. Goldman (dir.), *The construction of mental representations during reading* (p. 71-98). Mahwah, NJ : Lawrence Erlbaum Associates Publishers.
- Wagener, A. et Dion, E. (2018, novembre). *S'assurer que l'enseignement du décodage a du sens pour les jeunes élèves à risque de difficulté*

d'apprentissage. Communication présentée au 29^e Colloque de l'ADOQ (Association des orthopédagogues du Québec). Laval, Québec.

Wechsler, D. (2008). WIAT-II Test de rendement individuel de Wechsler, deuxième édition, version pour francophones.

Williams, J. P. (1991). Comprehension by learning-disabled and nondisabled adolescents of personal/social problems presented in text. *The American Journal of Psychology*, 104(4), 563-586.

Williams, J. P. (1993). Comprehension of students with and without learning disabilities: Identification of narrative themes and idiosyncratic text representations. *Journal of Educational Psychology*, 85(4), 631-641.

Williams, J. P., Hall, K. M., et Lauer, K. D. (2004). Teaching expository text structure to young at-risk learners: Building the basics of comprehension instruction. *Exceptionality*, 12(3), 129-144.

Williams, J. P., Kao, J. C., Pao, L. S., Ordynans, J. G., Atkins, J. G., Cheng, R., et DeBonis, D. (2016). Close analysis of texts with structure (CATS): An intervention to teach reading comprehension to at-risk second graders. *Journal of Educational Psychology*, 108(8), 1061-1077.

Williams, J. P., Nubla-Kung, A. M., Pollini, S., Stafford, K. B., Garcia, A. et Snyder, A. E. (2007). Teaching cause-effect text structure through social studies content to at-risk second graders. *Journal of Learning Disabilities*, 40(2), 111-120.

Williams, J. P., Pollini, S., Nubla-Kung, A. M., Snyder, A. E., Garcia, A., Ordynans, J. G. et Atkins, J. G. (2014). An intervention to improve comprehension of cause/effect through expository text structure instruction. *Journal of Educational Psychology*, 106(1), 1-17.

Wolf, M. M. (1978). Social validity: The case for subjective measurement or How applied behavior analysis is finding its heart 1. *Journal of Applied Behavior Analysis*, 11(2), 203-214.

Yopp, R. H., et Yopp, H. K. (2012). Young children's limited and narrow exposure to informational text. *The Reading teacher*, 65(7), 480-490.

Yu, C.Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. Los Angeles, CA : University of California.