UNIVERSITÉ DU QUÉBEC À MONTRÉAL

L'EFFET DU STATUT LINGUISTIQUE SUR LA RÉUSSITE DES ÉLÈVES CANADIENS AUX ITEMS DE MATHÉMATIQUES DU PISA

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

MAÎTRISE EN ÉDUCATION

PAR

RACHA KHODR

SEPTEMBRE 2025

REMERCIEMENTS

Ce mémoire n'aurait jamais vu le jour sans le soutien, la générosité et la patience de plusieurs personnes à qui je tiens à exprimer toute ma reconnaissance.

Je remercie tout d'abord mon directeur de recherche, professeur Guillaume Loignon, pour sa bienveillance, sa rigueur intellectuelle, sa disponibilité et sa grande humanité. J'ai commencé ce parcours avec lui dès mes premiers jours au Canada. Dès notre toute première rencontre, je lui ai dit que j'étais « chanceuse » de l'avoir comme directeur. Avec le recul, je sais maintenant que j'étais bien plus que chanceuse : j'étais entre de très bonnes mains. J'ai été guidée avec une bienveillance, une rigueur et une confiance rare par quelqu'un d'exceptionnel. Son accompagnement, toujours respectueux de mon rythme, et sa confiance sans limites ont fait toute la différence. Il a cru en moi-même dans les moments où je doutais. Son soutien a été une boussole précieuse, et je lui en serai toujours reconnaissante.

Je tiens également à exprimer toute ma gratitude à ma codirectrice de recherche, professeure Dan Thanh Duong Thi, dont les conseils avisés et le regard attentif ont représenté un appui essentiel à ma persévérance. Merci pour votre soutien constant et votre confiance.

Un immense merci à mes parents, mon mari, mes enfants, ainsi qu'à mes proches, ma famille, mes cousins et cousines, en particulier Nada et Walid, qui m'ont portée dans les moments difficiles. Leur présence, leurs encouragements et leur amour ont été un véritable moteur pour continuer, même quand l'énergie venait à manquer. Enfin, je remercie toutes les personnes, connues ou anonymes, qui ont croisé ma route et m'ont, à leur manière, aidée à avancer.

À vous tous : merci du fond du cœur.

DÉDICACE

À mes parents, pour leur amour, leur soutien et leurs sacrifices,

À mon mari Wassim El Hassan, pour sa patience et sa compréhension,

À mes enfants, Hana, Adnan et Karim El Hassan, qui sont ma plus grande source de motivation et de bonheur,

À Nabiha Wafai et Hayat Iskandarani, pour leur support infini et leur bienveillance tout au long de ce parcours,

À mes professeurs, pour la qualité de leur enseignement et leur engagement,

À ma codirectrice Dan Thanh Duong Thi, pour tous les efforts et le soutien qu'elle m'a dédiés,

Et surtout, à mon directeur de recherche, Guillaume Loignon, pour sa patience, sa générosité et le partage de son savoir, qui ont été essentiels à la réalisation de ce travail.

Je vous exprime à toutes et tous ma profonde gratitude.

AVANT-PROPOS

Ce mémoire, c'est bien plus qu'un projet de recherche. C'est un bout de mon histoire, un reflet de mes convictions, et le résultat d'un parcours rempli de défis, de patience et de persévérance. J'ai choisi d'explorer le thème de l'équité en éducation, avec un regard particulier sur les élèves qui, comme mes enfants ou certains de mes élèves, doivent apprendre et être évalués dans une langue qui n'est pas la leur. Ce sujet me touche profondément, parce qu'il me ressemble, et parce qu'il parle de réalités que je côtoie au quotidien.

Cette recherche a vu le jour dans un contexte personnel difficile, entre les rendez-vous médicaux, les traitements et les journées où avancer un paragraphe était une petite victoire. Mais chaque étape franchie, aussi minuscule soit-elle, m'a rapprochée un peu plus de ce but que je m'étais fixé : faire entendre une voix, celle d'une éducation plus juste, plus inclusive.

Je tiens à remercier de tout cœur ceux qui m'ont accompagnée pendant ce parcours. À mes proches, qui m'ont portée quand j'étais fatiguée. À mon directeur de recherche, professeur Guillaume Loignon, pour sa bienveillance, sa rigueur et ses précieux conseils. Et à mon équipe médicale, pour son soutien aussi professionnel qu'humain.

J'espère que ce mémoire pourra, à sa manière, contribuer à faire évoluer les pratiques éducatives vers plus d'équité. Si ce travail peut aider à mieux comprendre les obstacles que rencontrent certains élèves et inspirer des changements concrets, alors tout cela aura vraiment valu la peine.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS, DES SIGLES ET DES ACRONYMES

- CMEC: Conseil des ministres de l'Éducation (Canada)
- EQAO: Office of Education Quality and Accountability (Ontario) (Office de la qualité et de la responsabilité en éducation)
- MEQ: Ministère de l'Éducation du Québec
- OCDE: Organisation de Coopération et de Développement Économiques
- PIRLS: Progress in International Reading Literacy Study (Étude internationale sur les progrès de la lecture)
- PISA: Programme for International Student Assessment (Programme international pour le suivi des acquis des élèves)
- TIMSS: Trends in International Mathematics and Science Study (Étude internationale des mathématiques et des sciences)
- FDI: Fonctionnement Différentiel des Items
- SIBTEST: Simultaneous Item Bias TEST (Test de biais simultané des items)
- MH: Procédure Mantel-Haenszel
- GLMM: Generalized Linear Mixed Model (Modèle linéaire hiérarchique généralisé)
- HGLM: Hierarchical Generalized Linear Model (Modèle linéaire généralisé hiérarchique)
- TRI: Théorie de Réponse aux Items
- 2P L: Two-Parameter Logistic Model (Modèle à deux paramètres logistiques)
- IEA: International Association for the Evaluation of Educational Achievement (Association internationale pour l'évaluation du rendement scolaire)
- MAR: Missing At Random (Données manquantes conditionnelles)
- MCAR: Missing Completely At Random (Données manquantes complètement aléatoires)
- MNAR: Missing Not At Random (Données manquantes non aléatoires)
- DRN: Differential Response Non-uniform (Fonctionnement différentiel des items non uniforme)
- DRU: Differential Response Uniform (Fonctionnement différentiel des items uniforme)
- ALSI: Analyseur Lexico-syntaxique Intégré

LISTE DES SYMBOLES ET DES UNITÉS

- Γ (gamma): Coefficient d'association utilisé dans les analyses de FDI
- h: Taille d'effet de Cohen (indice de différence de proportions)
- $p < 0,05$: Niveau de signification statistique
- X^2 (1): Valeur du test du chi-carré avec 1 degré de liberté
- M: Moyenne
- ET: Écart-type
- t(df): Statistique t avec degrés de liberté spécifiés
- d: Taille d'effet de Cohen (pour les tests t)
- $R^2\Delta$: Changement dans le coefficient de détermination (R^2)
- distance glm: Score de propension basé sur un modèle logistique généralisé

RÉSUMÉ

Ce projet de recherche s'intéresse à l'influence des facteurs linguistiques sur la performance des élèves en mathématiques, en mettant l'accent sur les enjeux d'équité dans les évaluations à grande échelle. À partir des données de l'enquête PISA 2012, l'étude vise à détecter la présence de fonctionnement différentiel des items (FDI) entre les élèves canadiens et canadiennes dont la langue du test est leur langue maternelle (L1) et ceux pour qui elle constitue une langue seconde (L2). Pour ce faire, une méthode de pairage par score de propension a été utilisée afin de contrôler les variables confondantes et d'isoler l'effet du statut linguistique, dans un contexte méthodologique complexe lié à la conception par livret des tests PISA.

Les résultats révèlent la présence de plusieurs items présentant un FDI en défaveur des élèves L2, suggérant que certains obstacles linguistiques peuvent affecter injustement la performance dans des épreuves censées évaluer uniquement des habiletés mathématiques. L'étude met en valeur l'utilité du pairage comme stratégie méthodologique pour garantir des comparaisons valides entre groupes. Elle contribue ainsi à une meilleure compréhension des biais potentiels dans les évaluations internationales standardisées.

Cette recherche propose des pistes concrètes pour améliorer la conception des tests, favoriser l'équité en éducation et soutenir des pratiques d'évaluation plus inclusives, en particulier pour les élèves dont la langue d'apprentissage diffère de celle de l'évaluation.

Mots-clés : fonctionnement différentiel des items (FDI), équité en évaluation, PISA 2012, pairage par score de propension, biais linguistique, évaluation des habiletés mathématiques.

ABSTRACT

This research project focuses on the influence of linguistic factors on student performance in mathematics, with an emphasis on equity issues in large-scale assessments. Using data from the 2012 PISA survey, the study aims to detect the presence of Differential Item Functioning (DIF) between Canadian students whose test language is their mother tongue (L1) and those for whom it is a second language (L2). To achieve this, a propensity score matching method was used to control for confounding variables and isolate the effect of linguistic status within the complex methodological framework of PISA's booklet-based test design.

The results reveal the presence of several items displaying DIF to the disadvantage of L2 students, suggesting that certain linguistic barriers may unfairly affect performance on assessments intended to evaluate only mathematical skills. The study highlights the usefulness of matching as a methodological strategy to ensure valid group comparisons. It thereby contributes to a better understanding of potential biases in international standardized assessments.

This research offers concrete recommendations to improve test design, promote equity in education, and support more inclusive assessment practices, particularly for students whose language of learning differs from that of the assessment.

Keywords: Differential Item Functioning (DIF), equity in assessment, PISA 2012, propensity score matching, linguistic bias, assessment of mathematical skills.

INTRODUCTION

Ce projet de recherche vise à approfondir le domaine de l'évaluation des habiletés mathématiques des élèves. Il souligne la nécessité de prendre en compte les aspects linguistiques dans la conception des tests d'évaluation des aptitudes en mathématiques afin d'éviter tout biais potentiel. L'objectif est d'aider les évaluateurs à mieux spécifier ces aspects linguistiques pour comprendre plus précisément leur influence sur la réussite et la performance des élèves en mathématiques. En examinant attentivement ces aspects, ce projet cherche à mettre en évidence leur importance dans le processus de l'élaboration des tests, en vue de favoriser la réussite des élèves dans cette matière.

Dans le cadre de cette étude, qui repose sur l'utilisation secondaire de données, nous nous intéressons à l'impact des caractéristiques linguistiques sur la performance en mathématiques des élèves. Nous analysons les résultats aux tests de mathématiques de l'enquête PISA 2012, en mettant l'accent sur les différences entre les groupes linguistiques. L'édition 2012 du PISA plaçait en effet l'emphase sur les mathématiques, et contient donc davantage d'items dans cette discipline. Pour ce faire, nous avons recours à l'analyse de Fonctionnement Différentiel des Items (FDI), qui examine les réponses aux questions de mathématiques, ou « items », afin de comparer les performances des différents groupes linguistiques. Le programme PISA, développé par l'Organisation de Coopération et de Développement Économiques (OCDE), évalue la capacité des jeunes de 15 ans à travers le monde à mettre en pratique leurs connaissances en lecture, mathématiques et sciences dans des contextes réels. L'approche FDI permet de détecter d'éventuelles différences dans les réponses des groupes linguistiques, ce qui peut indiquer la présence de potentiels obstacles linguistiques dans les évaluations de mathématiques (Zumbo, 2007). Bien que ce phénomène ne constitue pas un biais en tant que tel, son analyse reste essentielle pour identifier des enjeux susceptibles d'être liés à la langue dans les tests mathématiques et garantir l'équité du dispositif d'évaluation. Ces recherches sont importantes pour les concepteurs d'évaluations, car elles visent à détecter et corriger le FDI lié à la langue et à promouvoir l'équité dans l'évaluation des habiletés mathématiques. Notre recherche s'inscrit dans un contexte où nous cherchons à favoriser l'équité dans les évaluations en mathématiques.

Cette recherche aspire à fournir des données probantes pouvant contribuer à l'amélioration des pratiques d'évaluation et à la conception de tests plus équitables. En mettant en lumière l'effet des facteurs linguistiques dans les évaluations de mathématiques, elle vise à sensibiliser les enseignants, concepteurs de tests et décideurs éducatifs aux défis rencontrés par les élèves dont la langue d'apprentissage est une langue seconde. L'objectif est de promouvoir des pratiques d'évaluation plus inclusives et de garantir des conditions d'évaluation équitables pour tous les élèves, quels que soient leur langue d'apprentissage ou leur profil linguistique.

Ainsi, cette étude s'inscrit dans une démarche visant à améliorer l'équité des évaluations éducatives, en veillant à ce que les différences de performance des élèves reflètent véritablement leurs habiletés en mathématiques, et non des barrières linguistiques externes aux habiletés évaluées.

Le premier chapitre de cette étude expose la problématique ainsi que les objectifs et les questions de la recherche, en soulignant l'importance sociale et scientifique de celle-ci. Ce chapitre présentera le problème de recherche spécifique découlant des écrits scientifiques sélectionnés. C'est ce problème de recherche qui sera au cœur de l'étude présentée à travers ce mémoire. Il aborde de façon générale un sujet relevant du domaine de la mesure et de l'évaluation en éducation, en se concentrant spécifiquement sur l'impact des caractéristiques linguistiques sur la performance des élèves lors des évaluations en mathématiques.

Dans le deuxième chapitre, nous présentons le cadre conceptuel ainsi qu'une revue de la littérature sur la validité en évaluation, en abordant sa dimension d'équité. Nous examinons également le concept de Fonctionnement Différentiel des Items (FDI) et le rôle important de cette méthode dans la détection des biais linguistiques potentiels présents dans les questions des évaluations de mathématiques de l'enquête PISA 2012. Parallèlement, nous explorons les facteurs qui influencent la performance en mathématiques, en nous appuyant sur les recherches scientifiques existantes. Nous analysons en particulier l'influence des caractéristiques linguistiques sur la performance des élèves. Cette revue de littérature nous permet de mieux cerner les contours de notre propre analyse et de situer notre contribution dans le champ de recherche existant.

Dans le troisième chapitre, nous exposons la méthodologie adoptée pour cette étude, en mettant l'accent sur l'utilisation d'une méthode de pairage pour identifier le FDI. Nous décrivons les données utilisées, ainsi que les outils employés et étapes suivies pour appliquer une méthode de pairage, puis analyser la présence potentielle de FDI.

Dans le quatrième chapitre, nous présentons les résultats des analyses menées, en mettant en évidence les effets du statut linguistique sur la performance aux items de mathématiques, et en identifiant les items présentant un FDI entre les groupes linguistiques. Les résultats sont interprétés à la lumière de la qualité du pairage réalisé et des spécificités liées à la conception par livret des tests PISA.

Dans le cinquième chapitre, nous proposons une discussion approfondie des résultats. Nous y analysons leur portée, les mettons en relation avec la littérature existante, et discutons les effets du pairage sur la détection du FDI. Cette section inclut également une comparaison avec d'autres études similaires, ainsi qu'une réflexion sur l'impact de la langue sur la performance en mathématiques, les limites de l'étude et les perspectives de recherche futures.

Enfin, la conclusion de l'étude résume les principaux résultats, souligne la contribution originale de cette recherche et propose des recommandations concrètes, tant sur les plans pédagogiques, méthodologiques et politiques, afin de renforcer l'équité en éducation.

CHAPITRE 1

PROBLÉMATIQUE

1.1 Contexte général

L'augmentation de la diversité ethnique et linguistique au Canada peut entraîner des défis dans le domaine de l'éducation. Le Canada connaît déjà une diversité linguistique notable avec une minorité anglophone vivant au Québec, une province francophone, et des minorités francophones vivant en contexte anglophone hors Québec (Gouvernement du Canada, 2021). De plus, il est prévu que d'ici 2036, les enfants issus de l'immigration représenteront entre 39 % et 49 % de l'ensemble des enfants, ce qui accentuera la diversité ethnique, linguistique et culturelle au sein de la population (Lory, 2023). Dans ce contexte, il demeure important d'assurer une évaluation équitable des aptitudes des élèves. Le ministère de l'Éducation du Québec (MEQ) souligne justement l'importance de l'équité dans l'évaluation et préconise des pratiques qui tiennent compte des caractéristiques des individus et des groupes afin d'éviter l'introduction de biais (MEQ, 2003).

Pour qu'un examen soit équitable, les élèves devraient avoir une « opportunité sans entrave » (*unobstructed opportunity*) à démontrer leur maîtrise de la compétence évaluée, indépendamment de facteurs tels que leur bagage linguistique, ainsi que leur statut socio-économique, ou leur familiarité avec les différents formats d'évaluations utilisés (AERA, APA et NCME, 2014). Dans cette recherche, nous désignons par « biais linguistique » l'influence indésirable et imprévue des caractéristiques linguistiques du test ou de la langue de l'élève sur sa capacité à démontrer sa compétence (Loignon, 2021b).

Les mathématiques, en particulier, représentent un cas intéressant pour l'étude des biais linguistiques. D'une part, les mathématiques sont une discipline dont la maîtrise importe pour la réussite scolaire autant que la participation citoyenne (Pichette et al., 2023; Dinglasan et Patena, 2013). D'autre part, en raison de leur symbolisme en apparence universel, les mathématiques pourraient sembler à l'abri des biais linguistiques (Chronaki et Planas, 2018). Durand-Guerrier et Chellougui (2015) soulignent toutefois que, contrairement à la croyance répandue selon laquelle les mathématiques sont universelles et intemporelles, de nombreux travaux en didactique des mathématiques montrent l'importance de prendre en compte les aspects culturels et linguistiques dans l'enseignement de cette matière. En effet, comme l'ont montré plusieurs études, dont Buono et Jang (2021) ont fait la synthèse, la compréhension des

relations mathématiques est étroitement liée à la maîtrise linguistique et peut difficilement être conçue comme une compétence isolée de la langue.

1.1.1 Sources de biais linguistiques

Les facteurs linguistiques qui influencent la réussite d'items de mathématiques peuvent se diviser en deux catégories interreliées : les facteurs associés à l'élève, et ceux découlant de l'épreuve ou du test en tant que tel (Loignon, 2021b). Les facteurs associés à l'élève incluent, par exemple, la langue parlée à la maison et le niveau de compétence dans la langue d'instruction. Même si les apprenants dont la langue maternelle diffère de celle de l'évaluation (désignés dans cette recherche comme élèves L2, pour langue seconde) possèdent les connaissances mathématiques requises, leur bagage linguistique peut constituer un désavantage (Abedi et Lord, 2001). Par exemple, un élève L2 peut rencontrer des difficultés supplémentaires qui ne sont pas liées à ses habiletés mathématiques réelles, mais plutôt à des barrières linguistiques (Martiniello, 2009).

Du côté des facteurs propres au test, mentionnons le vocabulaire utilisé, de même que la structure des phrases, qui peuvent influencer la capacité à répondre correctement aux questions (Abedi et Lord, 2001). Dempster et Reddy (2007) ont ainsi montré que la lisibilité des items influençait la performance des élèves à des épreuves de sciences. En outre, Persson (2016) soulève que, lorsque le langage employé est flou et décontextualisé, les élèves peuvent rencontrer des difficultés à reconnaître la terminologie propre au domaine et, par conséquent, ne parviennent pas à démontrer pleinement leurs compétences réelles.

La complexité linguistique des épreuves de mathématiques se manifeste différemment pour les élèves L2, il s'agit en fait d'un obstacle supplémentaire (Abedi et Gándara, 2006; Riccardi et al., 2020). Martiniello (2009) souligne que la complexité linguistique des écrits académiques, caractérisée par des termes techniques peu fréquents et des constructions syntaxiques complexes, creuse l'écart de performance entre les élèves L2 et leurs pairs dont la langue d'apprentissage est leur langue maternelle (élèves L1) (Abedi et Gándara, 2006; Abedi et Lord, 2001).

1.1.2 Analyse des biais linguistiques par analyse de FDI

L'analyse FDI (Fonctionnement Différentiel des Items), connue en anglais sous le terme Differential Item Functioning (DIF), est une approche statistique utilisée pour identifier les items de test qui fonctionnent

différemment pour différents groupes d'élèves ayant le même niveau de compétence. Un item présente un FDI lorsque les élèves de différents groupes, mais ayant le même niveau de compétence, ont des probabilités différentes de répondre correctement à cet item. Les items avec FDI peuvent entraîner une mesure biaisée des aptitudes, car celle-ci est influencée par des facteurs dits « nuisibles » (Magis et al., 2010). La présence de FDI contrevient donc avec l'objectif d'évaluer les aptitudes de manière équitable.

Suivant Zumbo (2007), nous distinguons l'impact de l'item, qui désigne une situation où un FDI apparent résulte de réelles différences entre les groupes dans l'habileté mesurée par l'item, et le biais de l'item, qui survient lorsque le FDI est attribuable à une caractéristique de l'item non pertinente pour l'habileté évaluée, constituant ainsi une source d'injustice en évaluation. Comme le souligne Zumbo (2007), le biais d'item constitue une condition suffisante, mais non nécessaire, pour qu'un impact apparaisse, ce qui distingue ces deux concepts. Lee et Randall (2011) définissent un item comme biaisé s'il y a une différence dans la probabilité de bonne réponse entre des groupes autrement équivalents à l'égard des habiletés requises pour réussir l'item. Cette distinction est importante dans le cadre du présent mémoire, car elle permet de mieux expliquer les différences de performance entre les élèves. En comprenant si une différence de performance est due à une véritable différence de compétence (impact de l'item) ou à un biais injuste (biais de l'item), les éducateurs et les chercheurs peuvent mieux cibler les interventions nécessaires pour rendre les évaluations plus équitables.

Plusieurs études ont porté spécifiquement sur le FDI associé à la langue. Lee et Randall (2011) ont réalisé une revue de la littérature scientifique menant à des résultats mitigés : quatre des huit études sélectionnées n'ont pas identifié d'items présentant un biais linguistique, possiblement en raison d'échantillons de taille insuffisante ou d'autres défis méthodologiques. Cette étude soulignait donc déjà l'importance de faire des choix méthodologiques adéquats dans les études de FDI, sans quoi l'effet de l'item peut difficilement être différencié du biais de l'item.

D'autres études ont depuis montré la présence de FDI associés à la langue, en employant notamment les items ou données tirées d'épreuves à grande échelle. Par exemple, Buono et Jang (2021) ont utilisé une méthode d'analyse de FDI pour analyser les items d'un test de mathématiques administré obligatoirement à des élèves de 6^{e} année de l'Ontario, afin d'étudier l'impact de la difficulté linguistique des items sur la réussite d'apprenants L1 et L2. Leurs résultats ont montré que certaines caractéristiques linguistiques,

notamment le vocabulaire abstrait ou peu fréquent, avaient un impact plus important sur les performances des apprenants L2 que sur les performances des apprenants L1.

Cependant, l'application de l'analyse FDI dans le cadre des évaluations à grande échelle peut poser des défis méthodologiques notamment en ce qui a trait à la gestion des données manquantes. Comme nous le verrons dans ce qui suit, certaines de ces évaluations peuvent constituer une source de données pertinente pour l'étude du FDI associé à la langue. Toutefois, les données tirées de ces enquêtes présentent des défis d'analyse. Dans le cadre de l'enquête PISA, la passation des tests s'effectue selon un dispositif par « livrets » (booklet design), où les items sont répartis en différents cahiers administrés aléatoirement aux élèves (OCDE, 2014). Cette approche introduit une complexité méthodologique dans l'analyse du FDI, nécessitant des ajustements statistiques spécifiques.

1.1.3 Difficultés dans l'application de l'analyse de FDI à des données d'enquêtes à grande échelle

L'enquête PISA, ou Programme International pour le Suivi des Acquis des élèves (PISA), constitue une source de données populaires pour les études s'intéressant à l'effet de divers facteurs sur la performance en mathématiques. Il s'agit d'une évaluation mondiale portant sur les compétences des élèves en mathématiques, en lecture et en sciences. PISA définit les compétences globales comme la capacité à analyser des enjeux complexes, à comprendre diverses perspectives et à agir pour le bien commun (OCDE, 2020). Plutôt que d'évaluer la mémorisation, PISA mesure la capacité des élèves à mobiliser et appliquer leurs connaissances dans des contextes réels (CMEC, 2019). Cette conception rejoint le cadre du CMEC (2020), qui présente les compétences globales comme un ensemble intégré de savoirs, d'habiletés et d'attitudes applicables à divers contextes. Contrairement aux approches purement cognitives, PISA et le CMEC privilégient une vision globale intégrant les dimensions intellectuelles, sociales et éthiques. Notre recherche s'inscrit dans cette perspective pragmatique, centrée sur les aptitudes et les capacités d'analyse, plutôt que sur des savoirs purement académiques.

L'enquête PISA cible l'ensemble des jeunes de quinze ans dans les pays participants. Au-delà de la composante formelle de l'évaluation, le PISA vise à collecter des données sur les attitudes et les croyances des élèves, leur accès et utilisation des technologies de l'information et de la communication, ainsi que des informations démographiques, telles que le sexe, l'année d'études, la langue maternelle et le statut socio-économique et culturel (Cheema, 2019).

Dans PISA, la passation des tests s'effectue selon un dispositif par livrets (« booklet design »), où l'on soumet aux élèves seulement une fraction des items disponibles (Frey et Bernhardt, 2012), ce qui complique l'application d'analyses de FDI (Akcan et Kabasakal, 2023). Les items sont répartis dans différents livrets distribués aléatoirement aux élèves (Bourny et al., 2001). Cette méthode vise à assurer une estimation efficace des aptitudes des élèves, en distribuant les items parmi les participants de manière équilibrée (Frey et Bernhardt, 2012). Chaque livret contient un certain nombre de clusters (groupes d'items), et chaque item apparaît dans différentes positions à travers les livrets.

Le *booklet design* offre plusieurs avantages, notamment une grande réduction du nombre d'items nécessaires pour évaluer précisément les élèves, ainsi qu'un contrôle des effets de position, lesquelles peuvent survenir lorsque la fatigue entraine une diminution de la probabilité de répondre correctement à un item apparaissant plus tard dans le livret (Frey et Bernhardt, 2012). Cependant, ce choix méthodologique entraîne un grand nombre de données manquantes, puisque les élèves ne sont exposés qu'à une fraction des items. En somme, ces données résultent d'un choix méthodologique assumé et doivent être distinguées des données manquantes qui surviennent lorsque l'élève ne répond pas à un item.

Les données manquantes représentent un obstacle majeur pour l'analyse du FDI, car elles peuvent diminuer l'efficacité des méthodes de détection de FDI (Akcan et Kabasakal, 2023). La répartition des items en livrets accentue ce problème en limitant la comparabilité directe entre élèves et en réduisant la puissance statistique, compliquant ainsi l'interprétation des résultats. L'impact de ces données varie selon leur mécanisme : pour les données manquantes complètement aléatoires (MCAR), comme lorsque la non-réponse est attribuable à l'élève (codée « 0 »), par exemple lorsqu'un élève oublie simplement de répondre à une question ou saute accidentellement un item sans raison particulière, l'effet est principalement une réduction de la puissance statistique sans biais majeur (Little et Rubin, 2019). En revanche, pour les données manquantes conditionnelles (MAR), qui représentent environ 85% des manques dans PISA en raison de la répartition des items entre livrets, les méthodes de détection de FDI peuvent voir leurs taux de faux positifs augmenter jusqu'à 30% (Sulis et Porcu, 2017). Ces absences MAR sont systématiques, mais explicables par des variables observées, notamment la structure du *booklet design*, dans laquelle un élève ne répond pas à un item simplement parce qu'il ne figurait pas dans le livret qui lui a été attribué aléatoirement. Les données manquantes non aléatoires (MNAR) posent les problèmes les plus graves, car liées à des facteurs inobservés, comme lorsqu'un élève évite délibérément de répondre aux questions qu'il juge trop difficiles, introduisant ainsi des biais potentiellement importants dans l'estimation (Enders, 2022).

Une stratégie simple de gestion des données manquantes est la suppression de cas (*listwise deletion*), qui consiste à effacer les données des individus ayant des données manquantes. Cette stratégie, déjà critiquée (Aydin, 2024; Wang et Aronow, 2023), n'est pas appropriée pour les données issues d'un *booklet design*, car elle entraînerait la perte de l'ensemble des données, chaque élève ne répondant qu'à une fraction des items. Cela illustre la nécessité d'une gestion des données manquantes adaptée aux données provenant d'un *booklet design*. Parmi les méthodes d'appariement utilisées pour pallier l'effet des données manquantes, on peut citer les approches fondées sur la moyenne (*mean-based matching),* qui permettent de combler partiellement les lacunes tout en conservant un certain équilibre entre les groupes comparés.

1.1.4 Stratégies pour l'analyse de FDI sur des données issues d'un booklet design

Les études s'intéressant à l'identification de biais linguistiques dans les enquêtes PISA ont employé une variété de stratégies face aux difficultés méthodologiques découlant du *booklet design.* Ajello et al. (2018) ont relevé que la compétence en lecture influençait la réussite des items de mathématiques de l'enquête PISA 2012, en utilisant une méthode d'imputation et des modèles multiniveaux pour prendre en compte les erreurs de mesure dues aux données manquantes. Liu et Bradley (2021), quant à eux, ont comparé l'application de trois approches de détection du FDI, soit la procédure Mantel-Haenszel, le modèle de Rasch et le modèle linéaire généralisé hiérarchique (HGLM), afin d'analyser les biais potentiels dans les items mathématiques de l'enquête PISA 2012 et d'identifier un FDI lié au statut linguistique (locuteur natif de l'anglais ou élèves apprenant l'anglais).

D'autres études portant sur le même thème ont plutôt contourné les défis méthodologiques du *booklet design*. Par exemple l'étude d'Akcan et Kabasakal (2023) a évalué deux méthodes de détection du FDI : la procédure classique de Mantel-Haenszel et le modèle MIMIC (Multiple Indicators Multiple Causes). Les chercheurs se sont intéressés aux données manquantes complètement aléatoires (*missing completely at random*, ou *MCAR*) dans les données PISA et ont, pour cette raison, limité leurs analyses à un seul livret, sélectionnant ainsi des élèves ayant tous répondu aux mêmes items. Cette approche méthodologique a permis de s'assurer que tous les élèves sélectionnés avaient répondu aux mêmes items, créant ainsi un cadre standardisé pour comparer les performances des deux méthodes de détection du FDI avec des taux de données manquantes fixés à 5% et 15%, et en testant différentes techniques d'imputation.

En contexte canadien, Roth et al. (2015) ont mené des analyses de FDI montrant que le bagage linguistique des élèves pouvait exercer une influence indésirée sur leur réussite à des problèmes de mathématiques.

Les auteurs ont combiné une analyse selon la théorie de la réponse aux items (TRI) pour identifier statistiquement les items biaisés entre groupes linguistiques à des évaluations qualitatives visant à déterminer les sources linguistiques des écarts. Cette étude n'a cependant pas directement employé les données PISA, mais plutôt réutilisé une sélection d'items PISA; de nouveaux participants ont été recrutés pour les fins de l'étude et ont répondu aux mêmes items. Des recherches supplémentaires seraient donc nécessaires afin de vérifier si leurs résultats peuvent se généraliser au contexte plus large des véritables données PISA pancanadiennes.

Le pairage (*matching*) est une approche prometteuse qui a été proposée pour l'identification de FDI (Chen et al., 2020), y compris pour des données d'enquête PISA recueillies au moyen de la méthode de livrets (Arikan et al., 2018). Il s'agit de pairer chaque élève du groupe focal (dans notre cas, les apprenants L2) à un élève ayant des caractéristiques similaires, mais provenant du groupe de référence (apprenants L1). On établit ainsi deux groupes équivalents, ce qui aide à déterminer si la différence observée dans la probabilité d'obtenir la bonne réponse est due à un FDI de l'item. Le pairage demeure toutefois une approche méconnue dans le domaine de l'éducation (Lane et al., 2012). Un travail important reste à faire pour explorer l'application du pairage à l'identification du FDI d'origine linguistique. Plus spécifiquement, à notre connaissance, cette approche n'a pas encore été employée en contexte canadien pour investiguer le FDI lié à la langue dans les évaluations mathématiques des enquêtes à grande échelle. La présente étude place donc l'accent sur la méthode de pairage (*matching*) comme procédure pour identifier le FDI. Cette approche est intéressante, car, en sélectionnant des élèves ayant des caractéristiques similaires, elle permet d'isoler de manière intuitive l'effet de la variable d'intérêt sur la réponse aux items, suivant le principe populaire voulant que l'on compare « des pommes avec des pommes ». Plutôt que de recourir à des méthodes d'analyse du FDI plus complexes, nous privilégions une démarche à la fois simple, méthodologiquement rigoureuse et adaptée au contexte particulier des données PISA, permettant de dégager plus clairement les différences observées entre les groupes linguistiques. L'objectif est de mieux cerner dans quelle mesure les écarts constatés dans les réponses aux items peuvent être attribués à des FDI linguistiques, et non à d'autres variables confondantes, ce qui est important dans un contexte comme celui de PISA, où la structure en livrets impose des contraintes analytiques supplémentaires.

1.2 Objectifs de l'étude

Le problème visé par cette étude est la présence potentielle de FDI lié à la langue dans les évaluations de mathématiques de l'enquête PISA au Canada, ce qui pourrait désavantager les élèves apprenant dans une

langue seconde (L2) par rapport à leurs pairs dont la langue première est la langue d'enseignement (L1). Cette recherche vise à analyser l'impact des facteurs linguistiques sur la performance en mathématiques des élèves, en utilisant les données de l'évaluation PISA 2012 comme outil d'investigation, et non comme objet central de l'étude. L'objectif de cette étude est d'examiner de quelle manière le statut linguistique des élèves (L1 ou L2) influence leurs résultats aux items de mathématiques de l'enquête PISA. Cette situation soulève des questions sur l'équité des évaluations standardisées dans un contexte de diversité linguistique croissante. Ce mémoire se penche également sur les défis méthodologiques posés par la conception en livrets (*booklet design*) de l'enquête PISA et explore l'efficacité des approches de pairage pour identifier le fonctionnement différentiel d'items, contribuant ainsi à mieux comprendre les enjeux d'équité et à informer les pratiques pédagogiques et les politiques éducatives pour les élèves multilingues.

1.2.1 Portée et importance de l'étude

Le sujet de cette recherche est important puisque la diversité linguistique dans les salles de classe canadiennes ne cesse d'augmenter, soulevant des questions quant à son impact sur l'éducation et l'évaluation des habiletés mathématiques des élèves. Les résultats de cette recherche ont donc le potentiel de profiter aux enseignants, en leur permettant de mieux comprendre l'influence de la langue sur les performances de leurs élèves en mathématiques et d'adapter leur pédagogie en conséquence. Ils bénéficieront également aux concepteurs d'examens standardisés, qui pourront s'appuyer sur ces résultats pour développer des tests minimisant les obstacles linguistiques. Sur le plan scientifique, cette recherche comble un manque dans la littérature actuelle en analysant spécifiquement les FDI liés à la langue dans les données canadiennes de PISA 2012 et en relevant le défi des données manquantes associé à la structure complexe des données PISA, qui complique l'application d'analyses de FDI. Elle contribue ainsi à la popularisation des méthodes de pairage pour l'analyse de DIF dans des contextes de données complexes.

Au-delà des enjeux scientifiques et sociaux évoqués ci-dessus, cette recherche enrichit les sciences de l'éducation en adaptant des méthodes éprouvées d'autres disciplines, telles que l'appariement par score de propension et l'analyse du FDI. Ces méthodes, largement validées en biomédecine pour comparer des traitements (Gayat & Porcher, 2012) ou optimiser des études cliniques (Maekawa et al., 2024), offrent des outils pertinents pour aborder des défis éducatifs complexes. Leur adaptation à l'étude des FDI liés à la langue dans les évaluations à grande échelle montre comment le transfert interdisciplinaire peut renouveler nos analyses et éclairer des enjeux essentiels liés à l'équité éducative.

L'utilisation des données canadiennes de l'enquête PISA 2012 s'avère pertinente pour cette recherche, tant pour des raisons méthodologiques que contextuelles. Tout d'abord, le cycle PISA 2012 présente des avantages spécifiques pour l'étude des habiletés de mathématiques. Cette édition avait en effet les mathématiques comme domaine principal d'évaluation, ce qui se traduit par un nombre plus élevé d'items en mathématiques par rapport aux autres cycles. Cela permet d'accroître la précision et la portée des analyses de FDI. De plus, tous les élèves participants ont répondu à des items de mathématiques dans leur cahier de test, ce qui assure une certaine uniformité dans les données recueillies. Ce point est important pour comparer les performances entre élèves de langue première (L1) et de langue seconde (L2). Par ailleurs, ces données comprennent déjà les variables clés nécessaires à cette étude : les performances en mathématiques, la langue maternelle, le statut d'immigration et diverses informations démographiques. Avec un échantillon représentatif de plus de 21 000 élèves issus des dix provinces canadiennes, ces données offrent une couverture nationale robuste (Mou et Atkinson, 2020). La participation importante du Canada à cette édition garantit une représentativité suffisante pour mener des analyses approfondies du contexte canadien.

Le choix des données canadiennes repose également sur la diversité linguistique croissante du pays, liée à une immigration importante et à la structure bilingue officielle (français et anglais). Le contexte linguistique canadien est unique, avec ses minorités francophones vivant dans des provinces à majorité anglophone comme l'Ontario, et ses minorités anglophones résidant au Québec, province majoritairement francophone. Cette situation engendre des défis particuliers quant à l'évaluation équitable des habiletés mathématiques. L'étude de Roth et al. (2015) a déjà mis en évidence des différences de performance liées à ces réalités linguistiques, soulevant des interrogations sur la validité des tests standardisés pour ces groupes minoritaires. L'analyse des données PISA 2012 permettra donc d'approfondir la compréhension de ces enjeux d'équité dans un contexte marqué par une grande diversité linguistique.

Sur le plan pratique, au moment de débuter les analyses, les microdonnées de PISA 2012 étaient entièrement disponibles, complètes, bien structurées et accompagnées de la documentation nécessaire, ce qui constitue une base solide pour des analyses fiables. À l'inverse, les données du cycle plus récent, PISA 2022, également centré sur les mathématiques, n'étaient pas encore entièrement disponibles ni pleinement accessibles au moment du lancement du projet, ce qui présentait certaines limitations pour leur utilisation. Ces éléments combinés justifient pleinement le choix de cette cohorte pour répondre aux objectifs de la recherche.

1.2.2 Aperçu et structure du mémoire

Dans la suite, nous présentons le contexte général de notre étude, en mettant l'accent sur l'importance de la performance en mathématiques dans le parcours scolaire et professionnel des élèves. Nous donnons un aperçu des évaluations en mathématiques à grande échelle, des facteurs influençant la performance en mathématiques, ainsi que de l'influence du biais linguistique sur cette dernière. Un accent particulier est mis sur les travaux portant sur l'approche du fonctionnement différentiel des items (FDI), que nous jugeons les plus pertinents. Cette approche examine comment les éléments d'un test, appelés dans ce contexte « items », fonctionnent différemment selon les groupes de personnes. Un aperçu des approches les plus courantes d'analyse du FDI sera d'abord présenté, puis une attention particulière sera accordée aux méthodes de pairage, centrales dans notre démarche analytique.

CHAPITRE 2

CADRE THÉORIQUE

Le chapitre précédent a exposé l'objectif principal de cette étude, qui est d'examiner les FDI potentiels liés à la langue dans les évaluations mathématiques du PISA au Canada. Cette recherche analyse l'influence du statut linguistique des élèves (L1 ou L2) sur leur performance et explore des méthodes pour identifier le fonctionnement différentiel des items (FDI) dans le contexte particulier de la diversité linguistique canadienne. Le présent chapitre a pour objectif d'établir les fondements théoriques nécessaires à la compréhension des concepts clés abordés dans cette recherche. Il permettra d'approfondir la notion de validité en évaluation, en mettant l'accent sur le phénomène du FDI et sur l'importance de le prendre en compte. Nous explorerons également les différents facteurs pouvant influencer la performance en mathématiques, en nous appuyant sur les données de l'enquête PISA. Nous examinerons par ailleurs l'impact des caractéristiques linguistiques des items sur les résultats des élèves, en particulier pour ceux dont la langue du test est une langue seconde. En abordant ces différents aspects théoriques, nous établirons les bases nécessaires pour mieux comprendre les enjeux liés à la validité des évaluations en mathématiques, ainsi que l'influence des facteurs linguistiques sur la performance des élèves.

Cette étude analyse l'effet du statut linguistique sur les performances en mathématiques des élèves de PISA en comparant le fonctionnement différentiel des items (FDI) entre les élèves dont la langue maternelle correspond à la langue du test (L1) et ceux de langue seconde (L2). Nous souhaitons déterminer si les items favorisent ou défavorisent les élèves en fonction de leur groupe linguistique, indépendamment de leurs habiletés réelles en mathématiques. Ce projet aborde la validité dans les évaluations ainsi que les enjeux liés à l'élaboration de tests justes et équitables, en particulier en ce qui a trait à la dimension linguistique. Pour ce faire, nous examinerons différentes notions, telles que l'évaluation des apprentissages, la validité des scores, le fonctionnement différentiel des items (FDI) et ses méthodes d'analyse courantes, ainsi que le pairage des données, en les situant dans le contexte de la présente étude.

2.1 Facteur influençant la performance en mathématiques

La performance en mathématiques résulte de multiples influences, allant des facteurs cognitifs aux dimensions affectives, sociales et contextuelles. L'examen de ces déterminants permet de mieux comprendre les écarts entre élèves et d'orienter les pratiques éducatives. Les sous-sections suivantes

mettent en lumière l'importance des mathématiques dans le parcours scolaire et professionnel, les perceptions et résultats des élèves, ainsi que d'autres facteurs liés à leur réussite en mathématiques.

2.1.1 Importance des mathématiques dans le cheminement scolaire et professionnel futur des élèves

Les aptitudes en mathématiques des élèves sont fondamentales pour leur réussite future, que ce soit dans leur parcours scolaire ou dans leur carrière. Les recherches ont établi un lien étroit entre la performance des élèves en mathématiques et leurs choix académiques et professionnels ultérieurs. Pichette et al. (2023) ont souligné que les élèves qui obtiennent de meilleurs résultats en mathématiques au secondaire tendent à présenter un taux de diplomation plus élevé, à accéder plus facilement aux études postsecondaires, à obtenir de meilleures notes dans ces études et à afficher un taux de diplomation postsecondaire plus élevé. D'autres chercheurs, comme Dinglasan et Patena (2013), ont mis en évidence l'importance des habiletés en mathématiques pour comprendre diverses disciplines, notamment l'ingénierie, les sciences, les sciences sociales et les arts. De plus, ils ont mis en évidence le rôle de l'auto-efficacité et de l'anxiété des élèves envers les mathématiques, une plus grande auto-efficacité et une moindre anxiété conduisant à de meilleurs résultats scolaires. Malabayabas (2022) a également souligné l'importance du bien-être scolaire, en particulier du concept de soi en mathématiques et de l'engagement dans le travail scolaire, en relation avec les aspirations éducatives et les résultats en mathématiques.

Par ailleurs, des études ont montré que la performance des élèves en mathématiques peut influencer leurs choix professionnels ultérieurs. Sells (1973) a qualifié les mathématiques de « filtre critique » permettant de sélectionner efficacement les élèves pour des carrières prestigieuses. Samo (2021) a renforcé cette idée en affirmant que les mathématiques permettent de prédire les futures contributions créatives et les rôles de leadership dans des fonctions importantes sur le marché du travail. Kogu (2017) a également noté que les résultats en mathématiques ont un impact important sur les aspirations professionnelles des élèves, soulignant qu'un niveau élevé d'habiletés mathématiques est requis pour accéder à des carrières exigeantes des habiletés logiques, d'investigation, critique et analytique, telles que la médecine, l'ingénierie, l'informatique, les sciences naturelles, les sciences sociales, les sciences physiques, les affaires et le commerce. Cependant, Stone et al. (2006) ont nuancé cette perspective en soulignant que les mathématiques ne sont plus seulement une exigence pour les carrières scientifiques et d'ingénierie prestigieuses, mais qu'un certain degré de culture mathématique est désormais requis pour toute personne entrant sur le marché du travail ou cherchant à progresser dans sa carrière. Ainsi, l'enseignement des mathématiques au secondaire joue un rôle important dans la préparation des jeunes à leur vie

d'adulte, impliquant l'acquisition de connaissances et de compétences essentielles pour devenir des citoyens et des travailleurs compétents à l'avenir. À cet âge, les élèves font souvent leur premier choix professionnel dans de nombreux pays, en raison des options et des filières proposées par les systèmes éducatifs qui peuvent les aider à déterminer leur orientation future (Demonty et al., 2013). En général, les élèves qui obtiennent de bons résultats scolaires à l'âge de 15 ans ont plus de chances de terminer des études postsecondaires et d'exercer une profession qualifiée à l'âge de 25 ans. Mais ceux qui n'obtiennent pas de bons résultats risquent davantage d'abandonner l'école (OCDE, 2023a).

2.1.2 Enjeux des perceptions et des résultats des élèves en mathématiques

Bien que les mathématiques soient perçues comme une matière d'une grande importance pour la réussite scolaire et qu'elles soient considérées comme un savoir de base pour d'autres disciplines, de nombreux élèves rencontrent des difficultés dans cette matière, pouvant mener à de l'anxiété à son égard (Mahmud et al., 2024). Cette situation est inquiétante, car les données de PISA 2009 montrent que 22 % des jeunes de 15 ans dans les pays de l'OCDE rencontrent de graves difficultés en mathématiques (OCDE, 2010). De plus, selon le MEQ (2019), de nombreux élèves éprouvent des difficultés à transférer les concepts mathématiques qu'ils ont appris à de nouveaux problèmes. L'OCDE souligne également qu'en 2010, plus de la moitié des élèves âgés de 15 ans des pays membres éprouvaient régulièrement de l'anxiété dans ce domaine. Une étude de Demonty et al. (2013) vient appuyer ces constats en révélant que plus de 40 % des élèves de l'OCDE ne reconnaissent pas l'importance des mathématiques et que plus de la moitié ont une attitude négative envers cette matière. Face à ces défis, il est important de rechercher et de comprendre les obstacles qui peuvent perturber la réussite des élèves et réduire leurs performances en mathématiques, afin de les aider à surmonter ces difficultés. Inci et al. (2023) indiquent que de nombreux élèves ont des perceptions négatives des mathématiques, car ils les considèrent comme un domaine rempli de formules complexes et de concepts abstraits, sans en percevoir la pertinence dans la vie quotidienne. De plus, les enjeux liés aux défis de résolution de problèmes soulignent l'importance de la prise en compte des barrières linguistiques pour assurer un apprentissage inclusif. Plusieurs recherches montrent que les élèves éprouvent souvent des difficultés à résoudre les problèmes de mots en mathématiques (Parmar et al., 1996; Wang et al., 2016). La résolution de ces problèmes comporte des défis linguistiques importants, tels que la présence de termes ambigus et la complexité des expressions linguistiques, qui peuvent nuire à la compréhension des élèves (Helwig et al., 1999; Wang et al., 2016). Il est donc essentiel de prendre en compte ces barrières linguistiques dans le processus d'enseignement et d'évaluation des mathématiques.

Dans le contexte québécois, une inquiétude est apparue concernant la baisse des taux de réussite aux épreuves ministérielles de mathématiques de 4e secondaire et à l'augmentation des échecs en mathématiques au 3e et 4e secondaire. L'article de Dion-Viens (2023) intitulé « Taux d'échec trop élevé en maths : Québec a fait passer des élèves qui ont eu 55 % » met en lumière une baisse importante du taux de réussite aux épreuves ministérielles de mathématiques de 4e secondaire dans la séquence « Culture, société et technique », qui ont été réintroduites en juin 2022 après une suspension de deux ans due à la pandémie. Cette baisse affecte 80 % des centres de services scolaires, où seulement un élève sur deux a réussi l'examen. Ces situations soulèvent la question du développement de nouvelles approches d'enseignement et d'évaluation, nécessaires pour favoriser la réussite des élèves en mathématiques, ainsi que dans la société et sur le marché du travail de demain.

2.1.3 Facteurs affectant la réussite en mathématiques

Les mathématiques sont souvent perçues comme une langue internationale (Inci et al., 2023), transcendant les barrières linguistiques et culturelles. La littératie mathématique, telle que définie par PISA, englobe la capacité d'utiliser les mathématiques dans divers contextes pour raisonner, appliquer des concepts et prendre des décisions éclairées (OCDE, 2023b). Cependant, la réussite des élèves en mathématiques est influencée par une multitude de facteurs complexes et interdépendants (Akyüz, 2014).

Wang et al. (2023a) distinguent ces facteurs à plusieurs niveaux : individuel, familial, scolaire et contextuel. Au niveau individuel, des caractéristiques comme l'âge, l'auto-efficacité en mathématiques et les résultats d'apprentissage sont liées positivement à la réussite, tandis que l'anxiété mathématique et le redoublement sont associés négativement. Le statut linguistique et le statut d'immigration jouent aussi un rôle important, illustrant l'interaction complexe entre ces variables. De plus, des études récentes ont également montré que la capacité visuo-spatiale joue un rôle important dans la performance en mathématiques chez les adolescents, en influençant la compréhension de concepts géométriques et la résolution de problèmes. Delage et al. (2024) ont observé que les élèves qui réussissent bien les tâches mathématiques tendent également à bien performer dans des tâches spatiales, suggérant une interconnexion entre ces deux domaines cognitifs.

Les différences entre les sexes sont également significatives. Perez Mejias et al. (2021) montrent que les garçons réussissent mieux les tâches de rotation mentale, tandis que les filles excellent dans des tâches verbales comme la lecture et la mémoire verbale. Ces différences sont amplifiées par les stéréotypes

sociaux (CMEC, 2020). L'âge des élèves constitue un autre facteur : Mourji et Abbaia (2013) observent une baisse du rendement avec l'accumulation de retard scolaire, alors que Wang et al. (2023a) suggèrent que des expériences éducatives cumulatives peuvent, au contraire, améliorer la performance des plus âgés.

Le statut linguistique influence directement la réussite. Buono et Jang (2021) montrent que les élèves dont la langue d'apprentissage est une langue seconde (L2) rencontrent davantage de difficultés face à des énoncés complexes que leurs pairs L1, même à habiletés égales. Martiniello (2009) et Lee et Randall (2011) confirment que la complexité lexicale et syntaxique des items favorise les L1, particulièrement dans les questions d'analyse de données. Les compétences en lecture ont également une incidence majeure. Ajello et al. (2018) démontrent que de bonnes habiletés en lecture facilitent la compréhension des énoncés mathématiques. Grimm (2008) insiste sur l'importance de ces compétences dès le primaire, et Larwin (2010) estime qu'elles expliquent jusqu'à 56 % de la variance des performances en mathématiques.

La confiance en soi et la motivation sont déterminantes. Lee et Stankov (2018) ainsi que Larwin (2010) montrent que la confiance dans ses habiletés améliore la réussite. Baten et al. (2019), Michaelides et al. (2019) et Arthur et al. (2022) ajoutent que l'intérêt, l'engagement et la motivation sont essentiels. À l'inverse, l'anxiété mathématique nuit fortement à la performance, comme l'illustre Chinn (2009).

Au niveau familial, le statut socio-économique (SES) est central. L'OCDE (2016, 2019) indique que les élèves issus de milieux défavorisés obtiennent en moyenne de moins bons résultats. Das et Sinha (2017) expliquent que les enfants de familles à SES élevé bénéficient de meilleures conditions de développement intellectuel, physique et émotionnel. Le rapport PISA 2012 précise qu'une hausse de l'indice de statut économique, social et culturel (ESCS) entraîne une augmentation plus marquée des scores en France que dans la moyenne de l'OCDE. Le capital social familial contribue également à la réussite. Zhang (2021) souligne l'effet des attitudes et attentes parentales dans le domaine scientifique, tandis que Le Mener et al. (2017) mettent en avant l'influence directe du niveau d'éducation des parents. Le statut d'immigration est aussi déterminant. L'OCDE (2016) montre que les élèves de première et deuxième génération obtiennent souvent des scores inférieurs. Cependant, Cheng et al. (2014) indiquent que certains groupes, notamment de deuxième génération, peuvent améliorer leur performance.

Au niveau scolaire, la perception des enseignants est décisive. Jaegers et Lafontaine (2020) révèlent que les attentes et le soutien des enseignants influencent directement la motivation et la réussite des élèves. Enfin, le système scolaire lui-même a un effet : selon l'OCDE (2023a), les élèves des écoles francophones

obtiennent en moyenne de meilleurs résultats en mathématiques que ceux des écoles anglophones. Pour conclure, la réussite en mathématiques résulte d'une interaction complexe entre facteurs individuels, familiaux et scolaires.

2.2 Utilisation des données PISA

2.2.1 Aperçu des évaluations à grande échelle en mathématiques (à l'échelle internationale et au Canada)

L'évaluation des habiletés mathématiques des élèves est au cœur de plusieurs programmes internationaux. Les principales évaluations internationales à grande échelle sont le Programme international pour le suivi des acquis des élèves (PISA) de l'OCDE, l'Étude internationale des mathématiques et des sciences (TIMSS) et l'Étude internationale sur les progrès de la lecture (PIRLS) de l'Association internationale pour l'évaluation du rendement scolaire (IEA) (Cordero et al., 2018). Ces évaluations recueillent également des données contextuelles via des questionnaires soumis aux élèves, enseignants, chefs d'établissement et parents, permettant des comparaisons internationales approfondies (Mullis et Martin, 2017). Comme le soulignent Kasimov (2019) et Rocher (2009), elles fournissent des informations précieuses sur les performances scolaires, établissent des normes internationales et contribuent à l'amélioration des résultats éducatifs, tout en permettant le positionnement des pays à l'échelle mondiale. Nous présentons ci-dessous les principales enquêtes internationales comportant des items de mathématiques.

PISA, initié par l'OCDE, se déroule tous les trois ans depuis 2000 et évalue les habiletés en lecture, mathématiques et sciences des élèves de 15 ans dans le but de déterminer s'ils possèdent les connaissances essentielles pour participer pleinement à la société moderne, en mettant l'accent sur l'un des trois domaines à chaque cycle (Cordero et al., 2018). PISA vise à évaluer non seulement les résultats scolaires, mais aussi la capacité des élèves à mobiliser leurs connaissances dans la pratique, ce qui est essentiel pour leur réussite future (Kasimov, 2019).

En ce qui concerne TIMSS, organisée par l'IEA tous les quatre ans depuis 1995, cette enquête mesure les performances en mathématiques et sciences des élèves de 4e et 8e année, sauf en Norvège où ce sont les 5^{e} et 9^{e} année pour permettre une meilleure comparaison avec les autres pays nordiques, compte tenu de la différence d'âge des élèves (Nilsen et Teig, 2024; Bergem et al., 2016). Elle utilise plus de 200 items en mathématiques et 250 en sciences, sous forme de QCM et de questions ouvertes (Mullis et Martin, 2017).

Quant à PIRLS, également menée par l'IEA, elle évalue spécifiquement les habiletés en lecture des élèves de 4e année, avec des enquêtes menées tous les cinq ans depuis 2001. Les élèves participants répondent à des questions conçues pour évaluer leur compréhension de textes écrits (Cordero et al., 2018).

Au Canada, il n'existe pas d'examens nationaux de mathématiques ni de programmes nationaux dans cette matière, l'éducation étant une responsabilité provinciale (Simmt, 2015; CMEC, 2024), ce qui signifie que chaque province et territoire a son propre ministère de l'Éducation et établit ses propres normes et directives pour les examens ministériels (CMEC, 2024). Cependant, de nombreux aspects de l'enseignement des mathématiques sont remarquablement similaires à travers le pays (Simmt, 2015). Cette cohérence résulte des efforts de collaboration entre les provinces et territoires, notamment par l'intermédiaire du Conseil des ministres de l'Éducation du Canada (CMEC, 2024). Ainsi, bien que chaque province gère ses propres examens ministériels selon ses normes, il existe une certaine harmonisation à l'échelle nationale (CMEC, 2024). Par exemple, en Ontario, l'Office de la qualité et de la responsabilité en éducation (EQAO) administre une évaluation standardisée des habiletés en mathématiques des élèves de 9e année, alignée sur le programme provincial et couvrant des domaines clés comme les nombres, l'algèbre et la géométrie (EQAO, 2020). Au Québec, le ministère de l'Éducation conçoit les épreuves ministérielles obligatoires, tandis que les commissions scolaires les administrent et en transmettent les résultats au ministère pour le calcul des taux de réussite et moyennes finales par programme (MEQ, 2024). Ces épreuves concernent notamment les mathématiques en 4e secondaire. Bien que la pandémie ait perturbé ces évaluations en 2019-2020 et 2020-2021, elles ont repris avec une pondération réduite en 2021-2022 et 2022-2023 (MEQ, 2024).

2.2.2 Schéma de collecte de données par livrets

Une caractéristique courante des enquêtes internationales en éducation est l'emploi d'une méthode d'échantillonnage sophistiquée appelée conception par livret ou « *booklet design* » (Frey et Bernhardt, 2012). Dans ce système, les élèves ne répondent qu'à une partie des questions disponibles, celles-ci étant réparties dans différents cahiers distribués de manière aléatoire (Bourny et al., 2001). Cette méthode permet d'optimiser l'évaluation en réduisant le nombre de questions par élève tout en maintenant une couverture large des aptitudes évaluées. Elle permet également de contrôler les effets de position des questions, compensant ainsi la fatigue potentielle des élèves au fur et à mesure du test (Frey et Bernhardt, 2012). Cependant, une collecte de données selon la conception par livrets génère inévitablement un grand nombre de données manquantes, puisque les élèves ne disposent pas tous de l'ensemble des items dans

leur cahier de test. Cela complique l'application des techniques traditionnelles d'analyse du FDI, lesquelles requièrent généralement des ensembles de données complètes (Akcan et Kabasakal, 2023).

2.2.3 Importance de l'enquête PISA dans le domaine de l'éducation.

Au fil des années, les évaluations à l'échelle mondiale telles que le PISA sont devenues un sujet central d'intérêt pour les chercheurs de divers domaines, y compris les économistes, les statisticiens, les ingénieurs et les éducateurs (Aricak et al., 2023). Le rapport PISA clarifie les différences de résultats scolaires entre divers groupes d'élèves, ce qui en fait un outil précieux pour reconnaître les inégalités en matière d'éducation. Les décideurs politiques du système éducatif peuvent utiliser ces informations pour entreprendre des mesures favorisant des chances égales à tous les élèves, quel que soit leur milieu socio-économique, culturel ou linguistique (OCDE, 2018). Au Canada, les résultats de PISA montrent que les élèves se classent généralement au-dessus de la moyenne de l'OCDE en lecture et en sciences, mais leurs performances en mathématiques demeurent plus variables selon les provinces et le statut linguistique. Par exemple, les élèves francophones obtiennent en moyenne de meilleurs résultats que leurs pairs anglophones, particulièrement au Québec, tandis qu'aucune différence significative n'est observée dans les autres provinces (OCDE, 2023a). Ces constats ont incité les décideurs canadiens à mettre en place des mesures ciblées afin de soutenir l'enseignement des mathématiques et de réduire les écarts entre groupes d'élèves, qu'ils soient liés à la langue d'enseignement ou au contexte socio-économique (OCDE, 2023a). Ainsi, comme dans d'autres pays, la validité et la fiabilité des tests PISA demeurent essentielles pour guider les décisions éducatives au Canada. Un autre exemple est fourni par l'étude de Kirwan (2015) met en lumière l'influence de PISA sur le développement des programmes de mathématiques en Irlande. La note moyenne obtenue par le pays lors de l'évaluation de PISA a provoqué des inquiétudes au sein du gouvernement irlandais, entraînant l'intégration des principes de PISA dans les politiques éducatives. Cela démontre que l'impact de cette évaluation internationale sur les choix politiques en matière d'éducation en Irlande était important (Kirwan, 2015). Par conséquent, l'invalidité des résultats de ces tests peut induire en erreur les décideurs politiques en éducation, ce qui peut avoir des implications nocives et dangereuses sur l'ensemble du système éducatif (Liu et al., 2019). En d'autres termes, il est indispensable de garantir la validité et la fiabilité de ces tests afin d'assurer des décisions éducatives justes et éclairées, tout en minimisant les biais possibles, qu'ils soient linguistiques, culturels, ethniques ou autres.

2.2.4 Choix des données canadiennes de PISA 2012

L'enquête PISA 2012, menée par l'OCDE, s'avère être un choix pertinent pour notre étude. Cette édition a mis l'accent sur les mathématiques, offrant ainsi un large éventail de variables détaillées associées à cette discipline. Parmi ces variables, on trouve des indicateurs tels que l'anxiété mathématique (ANXMAT), la motivation instrumentale pour les mathématiques (INSTMOT), l'intérêt pour les mathématiques (INTMAT), et l'auto-efficacité en mathématiques (MATHEFF) (OCDE, 2013). En plus de ces variables spécifiques aux mathématiques, PISA fournit également des renseignements utiles sur les caractéristiques individuelles des élèves. Ces données incluent des informations démographiques, le statut socio-économique, le statut linguistique et d'immigration, ainsi que des détails sur le niveau d'éducation des parents et les caractéristiques des établissements scolaires, notamment si le système scolaire est anglophone ou francophone (OCDE, 2013). En ciblant spécifiquement les élèves de 15 ans, l'âge typique du secondaire, ces données correspondent parfaitement à l'objet de notre recherche.

L'utilisation des données canadiennes du PISA 2012 se révèle particulièrement pertinente pour plusieurs autres raisons. Tout d'abord, elle inclut un vaste échantillon de 21 544 élèves issus de 885 écoles réparties dans les dix provinces, assurant ainsi une représentativité nationale adéquate (Mou et Atkinson, 2020). L'ampleur de cet échantillon constitue une base solide pour mener une analyse approfondie et des conclusions statistiquement significatives. De plus, le contexte canadien présente un intérêt particulier en raison d'une tendance préoccupante observée ces dernières années. Mou et Atkinson (2020) ont mis en évidence une baisse statistiquement significative des résultats moyens en mathématiques au Canada entre 2003 et 2015. Devant cette tendance préoccupante, Stokke (2015) souligne l'importance de cerner et d'analyser les facteurs qui y contribuent. Ainsi, une analyse approfondie des données canadiennes de PISA 2012 pourrait mettre en lumière ces facteurs, offrant des éclaircissements précieux pour guider les décisions politiques visant à améliorer les résultats en mathématiques dans le pays.

Le choix des données canadiennes se justifie également pour des raisons linguistiques. Le Canada présente un contexte bilingue et multiculturel complexe, ce qui en fait un terrain d'étude particulièrement pertinent. Comme l'expliquent Lauwerier et Akkari (2020), le Canada est caractérisé par une diversité linguistique où les élèves « autochtones » parlent généralement la langue du test à la maison, tandis que de nombreux élèves issus de l'immigration parlent une autre langue, avec des distinctions parfois entre les premières et deuxièmes générations. Cette diversité linguistique est au cœur de l'identité nationale et de la structure fédérale du Canada, comme le souligne Christofides (2010). Avec ses deux langues officielles, l'anglais et

le français, le Canada présente une situation unique où l'on trouve des minorités francophones vivant dans des provinces anglophones, comme l'Ontario, et des minorités anglophones vivant au Québec, province majoritairement francophone. Cette complexité linguistique présente des défis particuliers pour assurer l'équité dans l'évaluation des capacités des élèves lors de tests standardisés. Roth et al. (2015) ont mis en évidence dans leur étude l'influence des facteurs linguistiques et contextuels sur la performance des élèves. Dans ce contexte, Davis (2023) souligne l'importance d'une évaluation plus inclusive qui tienne compte des réalités linguistiques variées des élèves au Canada, particulièrement dans un contexte où l'immigration et la mondialisation ont accru la diversité culturelle et linguistique.

Ainsi, étudier le FDI lié à la langue dans les tests éducatifs du contexte canadien constitue une étape essentielle pour promouvoir l'équité. Cette analyse objective des différences de fonctionnement des items constitue une étape préalable à toute conclusion sur la présence possible de biais linguistique, permettant de garantir que chaque élève voit ses capacités reconnues et valorisées, indépendamment de sa langue maternelle. Dans cette étude, nous visons donc à examiner si le fonctionnement des items de PISA 2012 au sein de l'échantillon canadien varie selon le groupe linguistique et s'il favorise un groupe par rapport à un autre.

2.3 Influence des caractéristiques linguistiques sur la réponse aux items

De nos jours, l'importance de la langue dans l'enseignement des mathématiques est un sujet qui prend de l'ampleur. Les gouvernements et les systèmes éducatifs prennent de plus en plus conscience de la diversité linguistique croissante des élèves au sein des écoles, et s'efforcent de développer des méthodes d'enseignement, des politiques éducatives et des programmes adaptés à cette réalité, afin de répondre aux besoins variés des apprenants, tout en prenant en considération leurs origines linguistiques, culturelles et sociales (Barwell et al., 2019).

Les travaux de Moschkovich (2007), Barwell et al. (2007), ainsi que Njomgang-Ngansop et Durand-Guerrier (2011) mettent en évidence l'importance de considérer le multilinguisme dans la recherche en didactique des mathématiques. Durand-Guerrier et Chellougui (2015) soulignent la nécessité de s'éloigner de la conception des mathématiques comme étant universelles et de reconnaître l'impact des facteurs linguistiques et culturels sur l'apprentissage des mathématiques. Ils mettent également en lumière l'importance de la recherche dans ce domaine et appellent à une exploration plus approfondie susceptible de guider les futures orientations de l'enseignement des mathématiques, favorisant ainsi une pédagogie

adaptée à la diversité dans un contexte multilingue et multiculturel. Barwell et ses collègues (2017) ont identifié trois phases distinctes dans l'évolution de la recherche en éducation mathématique sur la diversité linguistique : de l'analyse de la cognition mathématique individuelle influencée par la langue, à l'étude des dynamiques collectives en classe, jusqu'à l'examen des contextes sociopolitiques pour promouvoir l'équité et l'intégration dans l'enseignement et l'évaluation des mathématiques. Comme pour l'historique que fait Zumbo (2007) concernant l'analyse du FDI, on remarque une progression où l'influence de facteurs périphériques est de plus en plus considérée en évaluation.

2.3.1 Impact des aspects linguistiques des élèves sur leur performance en mathématiques.

La langue constitue un élément essentiel du processus d'acquisition et d'enseignement des mathématiques, en permettant de décrire, d'expliquer et de communiquer des concepts abstraits et complexes (Schleppegrell, 2007; O'Halloran, 2005; Lemke, 2003). Peng et al. (2020) affirment, dans leur méta-analyse portant sur le sujet de l'analyse des relations mutuelles entre le langage et les mathématiques, qu'il existe une interaction entre la maîtrise de la langue et l'application des aptitudes mathématiques avancées. Une bonne maîtrise linguistique dans l'apprentissage des mathématiques permet à l'élève de consacrer davantage d'efforts cognitifs à une réflexion mathématique plus profonde, ce qui lui permet de maîtriser des habiletés mathématiques avancées. Par exemple, un élève qui maîtrise le vocabulaire mathématique et les normes d'écriture n'a plus besoin de consacrer autant de ressources cognitives à la compréhension des termes. Il peut plutôt se concentrer davantage sur la logique et les stratégies de résolution de problèmes. D'ailleurs, dans le cadre de l'apprentissage des mathématiques, deux perspectives s'affrontent quant au rôle de la langue. Autrefois, on croyait que les mathématiques étaient moins dépendantes du langage que la littérature, l'histoire ou d'autres matières scolaires. Cependant, cette vision a évolué. Aujourd'hui, on reconnaît l'importance du langage dans l'apprentissage des mathématiques comme dans d'autres disciplines (Schleppegrell, 2007). Pour comprendre les concepts mathématiques et communiquer les résultats après la résolution des problèmes, il est nécessaire que les élèves utilisent le langage de manière efficace. Cela suggère l'existence d'un lien étroit entre l'évaluation des compétences mathématiques des élèves et leurs habiletés linguistiques.

Dans le même contexte de la relation entre mathématiques et langue, Buono et Jang (2021) soulignent la richesse des débats entourant la question de savoir si le langage et les mathématiques peuvent être considérés comme des constructions distinctes ou interconnectées. Selon certains chercheurs, les capacités en langage et en mathématiques sont étroitement liées dès le plus jeune âge, car elles se

développent ensemble (Purpura et al., 2017; Purpura et Reid, 2016; Buono et Jang, 2021). Pour ces chercheurs, le traitement verbal et numérique est souvent nécessaire pour accomplir les tâches mathématiques, notamment les tâches d'arithmétique (LeFevre et al., 2010; Vukovic et Lesaux, 2013; Buono et Jang, 2021). D'autres chercheurs nuancent cette interconnexion entre ces deux domaines et soutiennent que certains éléments des mathématiques peuvent être compris indépendamment du langage. Par exemple, il existe des aspects de la cognition mathématique qui ne dépendent pas nécessairement de l'utilisation de mots numériques, tels que les nombres. Ils peuvent être traités par un système non verbal, tel que les symboles et les schémas (Castelli et al., 2006; Dehaene et al., 2004; Buono et Jang, 2021). Ces résultats complémentaires mettent en lumière le fait que ce sujet est un débat actif et qu'il importe de poursuivre les recherches concernant l'effet de la langue sur le processus d'enseignement des mathématiques, y compris ses évaluations.

2.3.2 Obstacles linguistiques des apprenants des mathématiques d'une langue seconde

Les recherches portant sur l'apprentissage des élèves de la langue seconde analysent comment la barrière linguistique peut faire obstacle à leur capacité à acquérir les concepts et le vocabulaire spécifiques aux mathématiques (Peng et al., 2020). Ces élèves sont ceux dont la langue d'apprentissage des mathématiques n'est pas leur langue maternelle. Ces recherches mettent en évidence les défis qu'une maîtrise insuffisante de la langue d'enseignement pose pour une bonne communication orale ou écrite et sur une participation active en classe de mathématiques (Moschkovich, 2002; Pimm, 1987; Schleppegrell, 2007; Peng et al., 2020). De plus, lorsque les apprenants des mathématiques de langue seconde (les locuteurs non natifs) transfèrent les sens des termes mathématiques de leur langue maternelle vers la langue seconde, ils ont tendance à les interpréter d'une manière différente de ce qui est attendu (Kazima, 2007; Peng et al., 2020). Ce transfert sémantique erroné du vocabulaire mathématique de la langue maternelle vers la langue seconde peut poser un problème lors des évaluations de mathématiques et affecter négativement les performances de ces apprenants. En fait, la mémoire des élèves bilingues pour les concepts mathématiques est liée à la langue dans laquelle ces concepts ont été initialement appris (Dehaene et al., 1999; Spelke et Tsivkin, 2001; Van Rinsveld et al., 2016; Buono et Jang, 2021), ce qui montre que la langue et les mathématiques sont deux systèmes interconnectés dans le cerveau, ce qui pourrait désavantager ces élèves lors des évaluations en langue seconde, car cela demande d'eux non seulement des capacités mathématiques, mais aussi des capacités linguistiques solides.

2.3.3 Impact des caractéristiques linguistiques des items des mathématiques sur la performance des élèves

Comme nous l'avons vu, il est important de tenir compte de la diversité linguistique croissante lors de l'élaboration et de la mise en œuvre des examens en mathématiques afin de garantir leur équité pour tous les groupes linguistiques (Liu et Bradley, 2021). De nombreuses études antérieures ont démontré que le facteur linguistique joue un rôle important dans la réussite et les performances des élèves en mathématiques (Abedi et Lord, 2001; Abedi, 2002; Haag et al., 2013; Loughran, 2014; Liu et Bradley, 2021). Ces études ont montré que les items d'évaluation des mathématiques présentent des aspects linguistiques qui défavorisent les élèves de langue seconde (Abedi et al., 2005; Abedi et Lord, 2001; Banks et al., 2016; Martiniello, 2008; Wolf et Leon, 2009; Liu et Bradley, 2021; Buono et Jang, 2021). Les différences dans les capacités linguistiques entre les élèves natifs de la langue du test et ceux de langue seconde peuvent affecter les résultats aux tests standardisés en mathématiques (Abedi, 2002; Solano-Flores et Li, 2008; Goodrich et al., 2021) étant donné que ces évaluations sont conçues pour des populations monolingues. Ceci pose des défis en termes de justice et soulève la possibilité de biais lors de l'évaluation des habiletés mathématiques des élèves bilingues (Goodrich et al., 2021). Il manque par ailleurs de données probantes quant à la validité des adaptations des tests de mathématiques dans une langue différente (Goodrich et al., 2021).

Plusieurs recherches ont été menées pour évaluer la présence de biais linguistiques dans les items des évaluations de mathématiques. Buono et Jang (2021) ont examiné la validité et l'équité des tests de rendement standardisés en mathématiques, en se concentrant particulièrement sur les élèves de langue anglaise (ELL), qui représentent les locuteurs non natifs de l'anglais. Cette recherche analyse la complexité linguistique dans une évaluation à grande échelle des mathématiques, mandatée par la province de l'Ontario, au Canada, ainsi que son impact sur la performance des ELL, à partir d'un échantillon de 28 items d'évaluation. Parmi ces items, onze ont été identifiés comme contenant un langage complexe, et sept d'entre eux ont présenté un fonctionnement différentiel significatif (FDI), favorisant les locuteurs natifs de l'anglais (EL1), par rapport aux ELL.

Banks et al. (2016) ont mené également une recherche qui visait à déterminer si, malgré des capacités mathématiques similaires, les élèves de 7e et 8e année en apprentissage d'une langue seconde (SLL) étaient désavantagés par rapport à leurs pairs non-SLL lors de la résolution de problèmes mathématiques à forte charge linguistique. En moyenne, les scores globaux des SLL se sont avérés inférieurs à ceux des

non-SLL pour chaque test. Cependant, seules deux catégories de problèmes présentaient un fonctionnement différentiel (FDI) statistiquement significatif, défavorable aux SLL. Si les évaluations de mathématiques comportent un niveau élevé d'exigences linguistiques, cela soulève des questions quant à la validité de ces évaluations pour mesurer précisément les habiletés mathématiques ciblées. Les résultats des tests deviennent alors difficiles à interpréter comme un reflet fidèle des capacités mathématiques réelles, car la compréhension de la langue introduit un facteur de confusion supplémentaire (Young, 2009; Buono et Jang, 2021).

Dans la même perspective, l'étude de Liu et Bradley (2021) examine, à partir de l'échantillon américain du PISA 2012, les écarts de réussite en mathématiques entre les apprenants de langue anglaise (ELL) et non-ELL. Elle identifie les items présentant des différences de fonctionnement significatives (FDI) défavorables aux ELL en utilisant trois méthodologies FDI (Différentiel d'Item Fonctionnement). Les résultats obtenus révèlent que ces items tendent généralement à avantager les non-ELL. Pour garantir des évaluations équitables auprès d'apprenants bilingues, il est essentiel d'analyser le FDI entre les groupes linguistiques, cependant, cette démarche demeure insuffisante en soi.

Il est tout aussi essentiel de contrôler rigoureusement les biais de sélection pour isoler les véritables différences liées à la langue, comme le souligne l'étude de Goodrich et al. (2021). Cette recherche a examiné les différences de fonctionnement d'items entre les versions espagnole et anglaise d'une évaluation de compétences mathématiques. Initialement, 11% des items semblaient présenter un FDI. Cependant, après avoir pris en compte le biais de sélection, c'est-à-dire les différences entre groupes sur d'autres caractéristiques que les capacités mathématiques et la langue, tels que le statut socio-économique ou d'autres variables démographiques, aucun item ne s'est révélé réellement biaisé sur le plan du contenu. Cette étude met en lumière la nécessité de contrôler rigoureusement le biais de sélection dans l'analyse du FDI. En effet, si les groupes comparés diffèrent sur d'autres variables que celles étudiées, cela peut fausser les résultats.

En somme, ces travaux montrent que les exigences linguistiques des items d'évaluation en mathématiques peuvent introduire des biais importants, désavantageant les élèves de langue seconde et remettant en question la validité et l'équité des tests standardisés. Il faut donc prendre en compte ces variables confondantes pour obtenir une image plus juste de la compétence mathématique et des différences réelles de fonctionnement des items entre les groupes linguistiques.

2.4 Validité des scores

L'évaluation des apprentissages est un processus qui implique la collecte et la discussion d'informations provenant de diverses sources afin de développer une compréhension approfondie de ce que les élèves savent, comprennent et sont capables de faire avec leurs connaissances à la suite de leurs expériences éducatives (Huba et Freed, 2000). De même, le ministère de l'Éducation du Québec (MEQ, 2002) définit l'évaluation comme étant un processus visant à juger les connaissances et les compétences acquises par un élève dans le but de prendre des décisions. Cette compréhension générale du rôle de l'évaluation met en évidence l'importance de considérer non seulement ce qui est évalué, mais aussi la manière dont l'évaluation est menée, ce qui ouvre la réflexion sur la question de l'équité en évaluation.

L'équité en évaluation signifie qu'il faut tenir compte des caractéristiques individuelles ou « communes à certains groupes » afin de prévenir les biais favorisant ou défavorisant injustement certains élèves (MEQ, 2003). En intégrant des pratiques d'évaluation équitables, les éducateurs peuvent non seulement évaluer de manière plus authentique les compétences de leurs élèves, mais aussi prendre de meilleures décisions concernant le processus éducatif de chaque élève. Des pratiques d'évaluation équitables contribuent à ce que les élèves issus de milieux sociaux et culturels divers aient des opportunités équivalentes de démontrer leurs capacités, tout en reconnaissant et en valorisant les multiples expériences que chacun apporte à la classe (Klenowski, 2014).

Un moyen d'atteindre ces objectifs d'inclusion et d'équité est d'améliorer la validité des outils d'évaluation. La validité est un élément fondamental de la psychométrie, car elle renvoie à la capacité d'un instrument, tel qu'un test, de mesurer fidèlement ce qu'il est censé évaluer (André et al., 2015 ; Laveault et Grégoire, 1997). Bodin (2016) souligne que la validité d'une évaluation correspond à l'adéquation entre ce que l'on veut mesurer et ce qui est véritablement mesuré. Affirmer qu'une méthode d'évaluation est valide signifie donc qu'elle permet de mesurer adéquatement la compétence visée et de prendre des décisions éclairées sur le niveau de compétence de l'élève (Klenowski, 2014).

Cependant, la validité n'est pas une propriété fixe des scores d'un test, mais un processus continu d'accumulation de preuves, tout au long du cycle de vie du test (Grégoire, 2014 ; Loye, 2018). Établir la validité des scores reste un défi permanent nécessitant une démarche rigoureuse pour accumuler et analyser diverses preuves, afin de garantir que les scores possèdent un niveau de validité suffisant, sans pour autant viser une validité absolue et définitive (André et al., 2015). De même, la validité des scores doit

être évaluée dans un contexte d'utilisation spécifique, avec des objectifs précis, une population cible et un cadre donné (André et al., 2015). Par exemple, les scores à un test d'intelligence validés pour identifier des élèves surdoués ne seront pas nécessairement valides pour prédire leurs performances professionnelles futures (inspiré de Laveault et Grégoire, 1997). La validité des scores dépend donc du contexte d'utilisation des tests : des scores peuvent être valides dans un contexte, mais non valides dans une autre.

Par conséquent, la validité des scores doit être réévaluée lors d'adaptations culturelles ou linguistiques (Grégoire, 2014). Des éléments culturels et linguistiques doivent être pris en considération dans le processus de validation. Ainsi, pour prendre un exemple inspiré de Laveault et Grégoire (1997), les scores à un test validé en français au Québec ne peuvent être considérés comme valides d'emblée pour une autre population francophone, telle que la population belge, sans nouvelles preuves de validité.

2.5 Biais dans les évaluations

Une source majeure de menace pour la validité d'un test est la présence de *biais* (AERA, APA et NCME, 2014). Les biais de test se traduisent par une validité différentielle des résultats pour des groupes spécifiques d'élèves, entraînant une erreur systématique dans l'évaluation de leurs performances (Jun-bin, 2011; Ricardo, 2024). Une erreur systématique signifie dans ce contexte que le processus de réponse d'un groupe d'individus est influencé par des caractéristiques sans rapport avec la compétence évaluée, de sorte que le test tendra à significativement sous-estimer le niveau de compétence de ce groupe (AERA, APA et NCME, 2014).

Un biais systématique peut compromettre la validité des résultats en fonction, par exemple, de l'origine ethnique, du statut socioéconomique ou du genre des participants (Stetz, 2022). Un biais de contenu apparaît lorsque le matériel d'un test est plus familier à certains groupes qu'à d'autres, influençant ainsi leurs performances selon leurs expériences antérieures (Reynolds et Suzuki, 2013). Les biais culturels surviennent en raison des écarts entre les normes, les valeurs et les expériences des personnes évaluées et celles des concepteurs des tests, influençant ainsi le développement cognitif et socio-émotionnel (Gálvez-Lara et al., 2015). Les biais socio-économiques favorisent souvent les individus issus de milieux favorisés, le statut socio-économique ayant un impact sur l'accès aux ressources éducatives (Doyle et al., 2022). Les biais linguistiques se produisent lorsque les barrières linguistiques affectent les performances, surtout chez les élèves dont la langue maternelle n'est pas la même que celle utilisée pour le test (Assilamehou-Kunz et Testé, 2022). Ces biais peuvent conduire à des diagnostics erronés, une sous-

représentation de certains groupes et l'aggravation des inégalités éducatives, posant d'importants défis éthiques (Lecerf, 2014). Par conséquent, il est essentiel de reconnaître et de réduire ces biais pour assurer des pratiques d'évaluation justes et équitables.

Les biais de test sont souvent le résultat de préjugés culturels qui désavantagent les élèves en fonction de leur origine ethnique, de leur genre, de leur statut socio-économique ou de leur langue, entre autres (Muñiz et Hambleton, 1996). La présence de biais dans un test peut se manifester sous plusieurs formes, telles que les différences de moyennes entre les groupes, le fonctionnement différentiel des items (que nous abordons plus loin), les interprétations erronées des scores et les contenus discriminatoires (Jun-bin, 2011). Il convient de souligner que, bien que le fonctionnement différentiel des items (FDI) puisse être un indicateur potentiel de biais, sa présence n'implique pas nécessairement l'existence d'un biais, mais constitue un signal d'alerte nécessitant une investigation plus poussée. En effet, déterminer si un test est équitable dépend de plusieurs facteurs complexes, tels que le contexte socioculturel et la manière dont il a été conçu (Muñiz et Hambleton, 1996; French, 2020).

2.5.1 Biais linguistiques dans les évaluations

Nous définissons les biais linguistiques dans les épreuves de mathématiques comme l'impact non anticipé et non souhaité des aspects linguistiques du test, ou des caractéristiques linguistiques de l'élève sur sa performance en mathématiques (Loignon, 2021b). Selon Buono et Jang (2021), les biais linguistiques dans les évaluations font référence aux influences de la langue utilisée qui peuvent affecter la performance des élèves lors de tests de mathématiques standardisés. Bodin (2016) explique que les biais linguistiques dans les tests se présentent lorsque des barrières langagières empêchent l'élève de démontrer sa compétence en mathématiques. Il mentionne que les biais linguistiques dans les évaluations de mathématiques peuvent parfois être détectés en modifiant la méthode de questionnement. Il donne l'exemple d'un élève qui échoue à résoudre un problème présenté par écrit, mais qui y parvient lorsqu'il est accompagné d'images et d'explications orales, illustrant ainsi que le langage constituait un facteur de biais.

Les biais linguistiques peuvent provenir de deux sources principales qui fonctionnent en tandem : les facteurs liés à des caractéristiques linguistiques du test, et les facteurs davantage liés à l'élève lui-même ainsi qu'à son environnement (Loignon, 2021b). Il est important de comprendre ces deux sources potentielles de biais pour mieux saisir la complexité du problème des biais linguistiques. En ce qui concerne les biais provenant du test, les facteurs incluent la complexité linguistique des questions. À cet égard,

Buono et Jang (2021) ont défini les biais linguistiques dans les items d'évaluation en mathématiques pour les élèves dont la langue d'apprentissage est une langue seconde (L2) comme des caractéristiques de complexité linguistique qui peuvent impacter leur performance. Parmi ces caractéristiques figurent l'utilisation de langage non mathématique, la voix passive, les groupes nominaux complexes, les propositions conditionnelles et relatives, ainsi que des phrases de questions complexes. Abedi et Lord (2001) abordent l'importance de la langue dans les performances des élèves aux tests de mathématiques sur des problèmes de mots. Ils suggèrent que le biais linguistique peut se référer à l'impact de la complexité linguistique sur la compréhension et la performance des élèves dans les tests de mathématiques. Il est important de noter que certains problèmes mathématiques, contenant un vocabulaire technique et une syntaxe complexe, peuvent être plus difficiles pour les apprenants L2 que pour les apprenants dont la langue de test est la langue maternelle (L1), même s'ils ont des capacités mathématiques équivalentes (Lee et Randall, 2011). Un test conçu pour évaluer les capacités en mathématiques est considéré valide s'il mesure spécifiquement ces habiletés, et non d'autres, comme les compétences en lecture.

Des facteurs liés à l'élève peuvent aussi influencer sa performance. Parmi ceux-ci, on peut citer certains aspects inspirés de l'étude de Lee et Randall (2011). Tout d'abord, les différences culturelles peuvent jouer un rôle dans la compréhension des problèmes mathématiques, car certaines notions peuvent être plus ou moins familières en fonction du contexte culturel de l'élève. De plus, le niveau de compétence linguistique des élèves peut influencer leur capacité à comprendre et à répondre correctement aux questions, affectant ainsi les résultats des évaluations en mathématiques. De même, l'interprétation des termes mathématiques varie selon la langue maternelle des élèves, ce qui peut entraîner des malentendus ou des erreurs dans les réponses fournies. Ces biais se traduisent souvent par des écarts de performance entre les apprenants L2 et leurs pairs L1, écarts possiblement attribuables à des obstacles linguistiques, ce qui peut nuire à la validité des scores des L2 par rapport aux L1 et remettre en question l'interprétation des capacités mathématiques des L2 (Lee et Randall, 2011).

Cependant, il est important de faire la distinction entre biais et intention d'évaluation, c'est-à-dire que la complexité linguistique du test n'est pas nécessairement source de biais. À ce propos, Bodin (2016) souligne que le biais linguistique ne réside pas dans la question en soi, mais dans l'intention évaluative sous-jacente. Il explique que, si les questions posées, en accord avec les objectifs recherchés de l'évaluation, comportent des textes longs, demandant un effort pour saisir l'énoncé, repérer les informations pertinentes, sélectionner les connaissances mathématiques adéquates et interpréter les

résultats, alors la compétence linguistique est sans doute essentielle pour l'évaluation. Toutefois, cela n'est pas un biais, car dans ce cas, l'importance du langage est volontaire et admise dans le contexte de l'évaluation. De même, Loignon (2021b) distingue la variance désirable, lorsque l'évaluation de la compétence langagière est un objectif intentionnel du test, et la variance indésirable, la part de la variance des scores attribuable à des processus linguistiques non liés à l'objectif principal du test. Dans notre étude, l'intérêt principal réside dans la variance non anticipée liée à ces facteurs linguistiques, plus spécifiquement dans la détection de FDI attribuables à la langue dans les tests mathématiques du PISA Canada. Cette étape est essentielle pour identifier, dans une phase ultérieure, les items dont le fonctionnement différentiel pourrait révéler des biais linguistiques potentiels. Il est essentiel de prendre en compte les biais linguistiques dans les évaluations de mathématiques pour assurer leur validité. Cependant, il est tout aussi important de différencier ces biais des intentions supposées de l'évaluation afin de ne pas nuire aux objectifs pédagogiques visés par ces évaluations.

2.6 Fonctionnement différentiel d'item

L'analyse du fonctionnement différentiel des items (FDI) est une méthode utilisée pour comparer les performances sur les items de test entre différents groupes, tels que les groupes linguistiques, culturels ou ethniques. Le FDI permet de déterminer si un item de test fonctionne de la même manière pour tous les groupes de participants, indépendamment de leurs caractéristiques individuelles (Zumbo, 2007). Le FDI se produit lorsque des individus de groupes différents, mais de même niveau de compétence, répondent différemment à un item. Cette transition reflète un changement d'approche dans l'évaluation de l'équité des tests, passant d'une simple comparaison entre groupes à une analyse plus approfondie et contextuelle des différences de performance entre groupes. Il est essentiel de distinguer clairement entre le FDI et le biais d'item, car bien que ces deux notions soient souvent associées, elles ne sont pas équivalentes sur le plan conceptuel. Comme le soulignent Gómez-Benito et al. (2018), le FDI se limite à identifier statistiquement des différences de performance entre groupes à item égal, tandis que le biais d'item suppose non seulement la présence d'un FDI, mais également une analyse interprétative approfondie, impliquant soit une sous-représentation du construit visé, soit l'influence de facteurs non pertinents pour ce construit. Conformément à cette distinction, la présence de FDI ne constitue pas une preuve de biais en soi, mais un signal nécessitant une investigation qualitative pour établir si l'item désavantage injustement certains groupes. L'analyse FDI permet d'identifier les questions spécifiques qui pourraient fonctionner différemment selon les groupes linguistiques, révélant ainsi des écarts de performance qui ne seraient pas apparents par une simple analyse des scores moyens (Martinková et al., 2017). Dans le cas des adaptations

linguistiques, le FDI aide à distinguer les biais liés à la langue des différences réelles de capacité. Cette approche place le FDI au cœur des questions de validité des scores, en examinant l'interprétation et l'utilisation des résultats dans divers contextes. Lorsque des items présentent un DIF, une analyse qualitative est nécessaire pour déterminer s'il s'agit d'un biais réel ou d'une différence légitime de performance. Ainsi, le FDI offre des pistes pour comprendre l'origine de ces écarts et adapter les méthodes d'évaluation en conséquence.

Zumbo (2007) structure l'évolution de l'analyse de FDI en trois générations ou cadres conceptuels. La première génération s'intéressait surtout aux biais des items dans une perspective d'équité des tests, cherchant à déterminer si un groupe performait au même niveau qu'un autre. L'analyse de première génération employait des méthodes relativement simples, comme la procédure de Mantel-Haenszel, qui utilise un tableau de contingence pour comparer les taux de réponses correctes entre les groupes pour chaque item, en stratifiant les élèves par groupe et par score total. La seconde génération a introduit des méthodes plus complexes, basées sur la théorie de réponse à l'item, qui prend en compte la difficulté des items pour mieux estimer l'habileté des élèves. Ces deux premières générations se concentraient sur l'identification d'items biaisés sans chercher à comprendre les causes sous-jacentes du FDI.

La troisième génération de méthodes d'analyse de FDI, selon Zumbo et al. (2015), se distingue par son approche écosystémique, inspirée du modèle de Bronfenbrenner (Bronfenbrenner, 1979; Kaushik et al., 2023). Elle considère les interactions complexes entre les personnes et leur environnement, prenant en compte différents niveaux écologiques, du microsystème au macrosystème, pour expliquer de quelle manière les environnements sociaux et culturels influencent les réponses aux items. Cette approche permet une compréhension plus complète des facteurs qui contribuent au FDI, en intégrant les influences de l'école, des enseignants, des parents, et d'autres contextes sociaux (Kaushik et al., 2023). Zumbo (2007) distingue également l'impact de l'item (différences réelles de capacité) du biais d'item (différences dues à des facteurs non pertinents). Les méthodes de détection du FDI se sont sophistiquées au fil du temps, intégrant des approches statistiques variées telles que les tableaux de contingence, les modèles de régression, ainsi que la modélisation des réponses aux items, notamment à travers la théorie de la réponse aux items (TRI). Il convient de souligner que les méthodes fondées sur la TRI s'inscrivent elles aussi dans une logique de régression, et ne constituent donc pas un cadre entièrement distinct, mais plutôt une extension de ces modèles adaptée à la nature des données de test (Zumbo, 2007). Ces méthodes se sont également étendues aux items polytomiques, qui offrent plus de deux options de réponse, contrairement

aux items dichotomiques. Parmi ces approches, on trouve les méthodes Mantel-Haenszel et de régression logistique.

L'évolution des méthodes du FDI reflète une tendance vers une compréhension plus approfondie des raisons du FDI, en intégrant diverses approches statistiques pour replacer le FDI dans un contexte plus large. Nous abordons dans la section suivante les méthodes traditionnelles les plus courantes de détection du FDI, afin d'en donner une vue d'ensemble et d'expliquer nos choix méthodologiques dans cette étude.

2.7 Aperçu des études sur le FDI associé à la langue

Cette section vise à présenter un survol de travaux empiriques ayant exploré le fonctionnement différentiel des items (FDI) en relation avec des variables linguistiques. L'objectif est de situer notre propre étude dans le paysage des recherches existantes, et de montrer en quoi les enjeux de langue sont centraux dans l'analyse du FDI en contexte éducatif.

Les études employant une approche de FDI sont nombreuses. Parmi celles-ci, on peut citer les travaux de Chen et ses collaborateurs en 2020. Ils ont proposé une méthode pour détecter le fonctionnement différentiel des items dans le contexte de l'évaluation de la compétence linguistique fonctionnelle en anglais. Leur analyse des données d'évaluation est basée sur les scores de tâches continues. Les résultats ont démontré l'efficacité de leur méthode pour identifier le FDI dû à des formations variées dans la tâche de score continu d'écriture en anglais. Ainsi, leur étude souligne l'importance de l'approche du FDI pour assurer l'équité et la validité des évaluations.

De même, dans l'étude de Buono et Jang (2021), les chercheurs ont utilisé le concept de FDI dans leur analyse pour déterminer si les items du test mathématique fonctionnaient de manière équitable pour différents groupes linguistiques d'élèves. Une autre étude qui adopte l'approche du FDI est celle de Wang et al. (2023b) qui stipulent que le FDI est souvent utilisé lors de la sélection des items pour garantir leur équité et leur validité dans tous les groupes de test. Ils définissent le FDI comme étant des procédures qui évaluent si les caractéristiques d'un item de test diffèrent pour différents groupes de personnes, à condition que les différences globales de performance aient été prises en compte. Cela signifie prendre en considération les différences générales de performance, telles que les caractéristiques démographiques, comme le sexe, l'âge, le statut socio-économique, et d'autres caractéristiques, comme le niveau de capacités, entre les différents groupes et les ajuster afin de garantir que les comparaisons entre les groupes

sont équitables ce que l'on appelle *le pairage des groupes* ou *matching.* Cela se fait en utilisant le *score de propension*, calculé pour chaque individu à partir de ses caractéristiques observables. Cela permet de comparer les résultats entre les groupes équilibrés, fournissant ainsi des estimations plus fiables des effets causaux (Hancock et al., 2018). Nous aborderons la notion du pairage des groupes et le score de propension de manière plus approfondie dans une partie ultérieure de ce chapitre afin de mieux comprendre leur utilité dans l'analyse de FDI.

En somme, les études sur le FDI démontrent son importance pour identifier les biais dans les tests. Cette méthode est essentielle pour notre étude, qui vise à évaluer l'équité des tests mathématiques pour les élèves multilingues. En ajustant les évaluations, cette méthode permet de mieux refléter les capacités réelles de tous les élèves, indépendamment de leur langue maternelle. Le FDI est une approche fréquemment utilisée dans la validation des tests, contribuant ainsi à garantir que les évaluations sont justes et équitables pour tous les individus évalués, tout en réduisant les risques de discrimination. Le Tableau 2.1 propose une synthèse des principales études ayant exploré le FDI en lien avec la variable linguistique, afin d'illustrer les approches méthodologiques utilisées et les résultats obtenus dans ce champ de recherche.

Tableau 2.1: ***Principales études antérieures sur le FDI lié à la langue***

Auteur(s) et année	Contexte / Domaine d'évaluation	Méthode utilisée	Principaux résultats
Chen et al. (2020)	Évaluation de la compétence linguistique fonctionnelle en anglais	Analyse basée sur des scores de tâches continues	Mise en évidence du FDI dû à des formations variées, soulignant l'importance d'ajuster pour l'équité
Buono & Jang (2021)	Test de mathématiques, comparaison de groupes linguistiques	Approche FDI classique	Identification de différences de fonctionnement des items selon la langue des élèves
Wang et al. (2023b)	Sélection d'items pour garantir l'équité	Méthodes de FDI avec ajustement des différences globales de performance	Importance du pairage (matching) pour isoler le FDI et assurer des comparaisons justes
Hancock et al. (2018)	Analyses méthodologiques (score de propension)	Pairage et score de propension	Méthode permettant d'obtenir des estimations causales plus fiables en équilibrant les groupes

2.8 Aperçu des méthodes les plus courantes pour l'analyse du FDI

Diverses méthodes ont été proposées pour identifier les items présentant un FDI, notamment le SIBTEST (Bolt et Stout, 1996), la régression logistique (Swaminathan et Rogers, 1990), l'analyse FDI multiniveau (Kamata, 2001), le modèle de Rasch (Rasch, 1960), le modèle à deux paramètres logistiques (2PL) de la TRI (Baker et Kim, 2004), ainsi que la procédure de Mantel-Haenszel (Mantel et Haenszel, 1959 ; Holland et Thayer, 1988).

2.8.1 SIBTEST

Le test de biais simultané des items (SIBTEST) est une méthode largement utilisée pour détecter le fonctionnement différentiel des items (FDI) dans les tests (Jiang et Stout, 1998). Cette approche non

paramétrique compare les performances de deux groupes, un groupe de référence et un groupe focal, en vérifiant si des différences persistent après appariement des candidats à partir d'une « sous-épreuve de correspondance », censée mesurer uniquement la compétence visée (Bolt et Stout, 1996). Pour contrôler les différences systématiques de compétences entre les groupes, le SIBTEST utilise une correction par régression qui ajuste les scores moyens, permettant d'estimer le score vrai des candidats à capacité équivalente (Bolt et Stout, 1996; Gierl et al., 2004). Parmi les principaux avantages du SIBTEST figurent sa capacité à limiter les erreurs et sa flexibilité, qui permet une analyse confirmatoire ou exploratoire pour l'analyse des items (Bolt et Stout, 1996). Cependant, la méthode présente également certaines limites. Elle nécessite des calculs statistiques approfondis, ce qui la rend plus complexe à mettre en œuvre que d'autres méthodes (Bolt et Stout, 1996). De plus, le SIBTEST a été moins étudié dans des situations où de nombreux items présentent un FDI, ce qui pourrait limiter sa fiabilité dans ces contextes (Gierl et al., 2004). Enfin, cette méthode repose sur un modèle multidimensionnel, ce qui peut poser un problème si le test ne comporte pas de dimensions secondaires ou si les items n'évaluent pas des habiletés comparables (Bolt et Stout, 1996). Par ailleurs, le SIBTEST peut être sensible aux distributions d'aptitudes différentes entre groupes, ce qui implique que des données manquantes pourraient potentiellement affecter la précision de l'estimation de l'aptitude, surtout si elles altèrent la comparabilité des distributions d'aptitudes entre groupes (Gierl et al., 2004). Compte tenu de la présence de données manquantes dans les données PISA, l'utilisation du SIBTEST ne serait pas appropriée dans le cadre de notre étude.

2.8.2 Analyse du FDI par régression logistique et approche multiniveau

La régression logistique, largement utilisée en recherche en éducation, permet d'examiner la relation entre une variable dépendante binaire ou dichotomique (par exemple, réussite ou échec) et diverses variables prédictives continues ou catégorielles (Niu, 2020; Peng et al., 2002). Elle fonctionne en transformant la probabilité d'un événement en logit, soit une transformation logarithmique du rapport des cotes de l'événement, ce qui permet d'établir une relation linéaire entre les variables indépendantes et la variable dépendante. Parmi ses avantages, cette méthode évite les hypothèses classiques de normalité et d'homoscédasticité des résidus, accepte différents types de variables prédictives, et fournit des résultats faciles à interpréter à l'aide des rapports de cotes (odds ratios) (Wurtz, 2008; Peng et al., 2002). Elle est donc bien adaptée aux contextes où l'on cherche à prédire des résultats dichotomiques, comme le succès ou l'échec académique. Dans l'analyse du FDI, elle consiste à vérifier si l'appartenance au groupe de référence ou au groupe focal influence la probabilité de réussite à un item, une fois le niveau de compétence contrôlé. Cependant, cette approche nécessite des échantillons de taille suffisante pour

garantir la stabilité des résultats et peut devenir difficile à interpréter lorsqu'elle inclut des interactions complexes entre variables (Niu, 2020; Peng et al., 2002).

L'approche multiniveau, parfois considérée comme une extension de la régression logistique, est particulièrement utile dans les contextes où les données présentent une structure hiérarchique, comme des élèves regroupés dans des classes ou des écoles (Liu et Bradley, 2021; Ravand, 2015; O'Connell et McCoach, 2008; Pastor, 2003). Elle permet d'analyser simultanément plusieurs niveaux d'organisation des données, ce qui en fait une méthode précieuse pour les évaluations à grande échelle, comme PISA, où les réponses sont imbriquées dans des niveaux multiples (Raudenbush et Bryk, 2002). Dans le cadre du FDI, cette approche permet de modéliser les variations des paramètres d'items entre les groupes, tout en tenant compte de la dépendance des données et en évitant la nécessité de séparer strictement les échantillons (Liu et Bradley, 2021). Appliquée au FDI, elle permet également de modéliser la différence de performance entre le groupe de référence et le groupe focal tout en tenant compte de la structure hiérarchique des données. Elle améliore la précision des estimations et réduit les biais associés aux violations de l'indépendance des observations, et permet une exploration détaillée des sources de variance à chaque niveau (Ravand, 2015). De plus, elle conserve la flexibilité de la régression logistique tout en intégrant une structure hiérarchique, ce qui facilite l'analyse de questions complexes en éducation.

Cependant, malgré ces avantages, l'approche multiniveau présente certaines limites importantes. Elle nécessite une expertise technique avancée et repose sur des hypothèses rigoureuses relatives à la structure des données. Cela devient rapidement exigeant en termes de ressources computationnelles, surtout dans des contextes de grande ampleur comme PISA (O'Connell et McCoach, 2008; Pastor, 2003). Dans notre étude, l'application de cette méthode n'est pas réalisable en raison de la complexité des modèles, de la taille importante de l'échantillon, ainsi que des contraintes liées à l'accès et à la manipulation des bases de données multiniveaux de PISA.

2.8.3 Modèle de Rasch

Le modèle de Rasch, fondé sur les idées initiales de Georg Rasch (Rasch, 1960), représente une approche probabiliste essentielle pour l'analyse des évaluations, reposant sur le principe que la probabilité de répondre correctement à une question dépend de la capacité de la personne et de la difficulté de l'item (Zamora-Araya et al., 2018). Ce modèle calcule la probabilité qu'une personne réponde correctement à un item en fonction de deux paramètres principaux : le niveau de trait de la personne et la difficulté de l'item,

tous deux exprimés sur une échelle logit (logarithme naturel du rapport des cotes) permettant leur comparaison directe (Zamora-Araya et al., 2018). Dans ce contexte, les items d'ancrage jouent un rôle fondamental en fournissant un cadre de référence pour le trait mesuré et en assurant une calibration correcte des items sur le continuum du trait latent (Zamora-Araya et al., 2018). Le modèle de Rasch est caractérisé par une courbe logistique qui s'apparente à une distribution normale cumulative, ce qui donne aux items une propriété d'additivité conjointe lorsqu'ils respectent les formes logistiques imposées par le modèle de Rasch (Perline et al., 1979). Cette propriété est avantageuse dans la mesure où elle permet une mesure plus cohérente sur une échelle d'intervalle, facilitant ainsi l'interprétation des scores (Törmäkangas, 2011).

Le modèle de Rasch n'est pas en soi une méthode de détection du FDI. Pour l'analyse du FDI, il est utilisé en comparant les paramètres estimés (difficulté des items, par exemple) entre le groupe de référence et le groupe focal. Toutefois, il est possible d'examiner le FDI en comparant les paramètres estimés par le modèle entre différents groupes (Engelhard, 2013). Parallèlement, les indices statistiques dérivés des résidus (comme les statistiques infit et outfit) permettent d'identifier des réponses inattendues (Bond et Fox, 2013). Ces indices peuvent être utiles dans une optique diagnostique, mais, comme le souligne Engelhard (2013), ils présentent des limitations majeures pour l'analyse du FDI: leur vulnérabilité aux artefacts de mesure et leur incapacité à distinguer les différents types de FDI.

Parmi les principales limitations du modèle de Rasch figure son incapacité à détecter le FDI non uniforme, c'est-à-dire lorsque l'effet différentiel varie selon le niveau de compétence des répondants (par exemple, un item favorise un groupe lorsque le niveau de compétence est faible, mais favorise l'autre groupe lorsque le niveau de compétence est élevé) (Mellenbergh, 1989). De plus, le modèle repose sur l'hypothèse d'unidimensionnalité, qui peut ne pas être applicable dans tous les contextes (Zamora-Araya et al., 2018). En fait, l'absence d'items d'ancrage non biaisés dans les données PISA rend impossible l'application du modèle de Rasch pour l'analyse du FDI dans notre étude.

2.8.4 Modèle à deux paramètres logistiques (2PL) de la TRI

La théorie de réponse aux items (TRI) constitue un cadre psychométrique essentiel permettant de modéliser la relation entre les caractéristiques des items et la probabilité de réponse correcte, tout en situant les individus et les items sur un même continuum latent (Van der Linden et Hambleton, 1997). Parmi les modèles TRI, le modèle à deux paramètres logistiques (2PL) (Baker et Kim, 2004) représente une

avancée par rapport au modèle de Rasch en intégrant deux paramètres clés: la difficulté de l'item et un paramètre de discrimination, qui mesure dans quelle mesure l'item différencie les répondants en fonction de leur niveau de compétence (Embretson et Reise, 2013). Dans le cas du FDI, le modèle 2PL permet d'examiner si les paramètres d'un item (difficulté et discrimination) diffèrent systématiquement entre le groupe de référence et le groupe focal. Cette caractéristique fait du modèle 2PL un outil pertinent pour analyser le FDI, puisqu'il permet de détecter à la fois le FDI uniforme et le FDI non uniforme. Le FDI uniforme se produit lorsqu'un item est toujours plus facile pour un groupe que pour l'autre, quel que soit le niveau de compétence. En revanche, le FDI non uniforme survient lorsqu'un item avantage un groupe à un faible niveau de compétence, mais l'autre groupe à un niveau de compétence plus élevé (Zumbo, 2007). Cependant, ce modèle présente des limitations. Comme tous les modèles de la TRI, le 2PL nécessite des items d'ancrage non biaisés pour établir une métrique commune. L'absence d'items d'ancrage dans les données PISA compromet l'application de cette méthode pour l'analyse du FDI dans notre étude.

2.8.5 Méthode de Mantel-Haenszel

La méthode de Mantel-Haenszel (MH) est utilisée pour détecter le FDI dans les items d'un test entre deux groupes : un groupe de référence « neutre » et un groupe focal, dans notre cas respectivement le groupe L1 et le groupe L2. La méthode de Mantel-Haenszel (MH) est utilisée pour détecter le FDI dans les items d'un test en comparant les performances du groupe de référence et du groupe focal. La méthode MH permet d'identifier de manière fiable les items potentiellement biaisés, contribuant ainsi à renforcer la validité des évaluations éducatives. Adaptée à la détection du FDI par Holland et Thayer (1988), la méthode MH compare les chances de réussite à un item entre les deux groupes, après avoir apparié les participants selon leur score total au test (Clauser et Mazor, 1998). Sur le plan opérationnel, la méthode MH s'appuie sur la construction de tableaux de contingence stratifiés (par exemple, par niveau de compétence), permettant une analyse détaillée des performances entre les deux groupes. Bien que souvent illustrée par des matrices 2x2 pour simplifier son interprétation, la méthode MH peut s'appliquer à des structures plus complexes, notamment lorsqu'elle est étendue à des analyses multivariées ou à des stratifications multiples (Agresti, 2018). La statistique de Mantel-Haenszel permet ensuite de calculer un rapport des cotes, ou *odds ratio* (α), représentant le rapport entre les chances de réussite à l'item pour le groupe de référence et celles du groupe focal. Le rapport des cotes (odds ratio) indique si, à compétence équivalente, un item est plus favorable au groupe de référence ou au groupe focal. Ce rapport sert à quantifier l'écart dans la probabilité de réponse correcte à un item entre les groupes (Holland et Thayer, 1988; Clauser et Mazor, 1998). Reconnue pour sa robustesse et sa simplicité, la méthode MH est largement utilisée dans la

littérature scientifique, notamment pour l'évaluation objective du FDI (Akcan et Kabasakal, 2023). Cependant, elle présente certaines limites. Par exemple, elle repose sur l'hypothèse selon laquelle le score total constitue un bon indicateur du trait latent sous-jacent, ce qui peut ne pas être toujours exact, notamment dans des tests multidimensionnels. La méthode MH ne peut détecter que le DIF uniforme (un biais constant entre groupes), mais reste aveugle au DIF non uniforme (où le biais varie selon le niveau de compétence des répondants), une lacune majeure pour l'analyse d'items complexes (Rogers et Swaminathan, 1993). De plus, son efficacité est optimale avec de grands échantillons; les résultats deviennent moins fiables lorsque le nombre de participants est faible (Millsap, 2011). Enfin, la méthode MH nécessite souvent des items d'ancrage jugés non biaisés pour établir une référence fiable. Or, dans certains contextes, il peut être difficile, voire impossible, d'identifier ou d'accéder à de tels items, ce qui limite son application.

D'ailleurs, les études de simulation ont démontré l'efficacité variable des méthodes de détection du FDI selon les conditions expérimentales. Par exemple, l'étude de Magis et al. (2011) souligne qu'aucune méthode ne s'avère systématiquement supérieure aux autres, malgré des différences méthodologiques significatives. Un autre exemple est fourni par Cuevas et Cervantes (2012), qui ont réalisé une étude de simulation comportant 270 conditions expérimentales distinctes et 200 réplications par condition. Leurs résultats révèlent que l'efficacité de la régression logistique pour détecter le DIF uniforme (DRU) et non uniforme (DRN) est influencée par plusieurs variables, telles que la taille de l'échantillon, le ratio entre groupes et certains paramètres d'items (difficulté, discrimination). Cette étude a également permis d'établir des points de coupure spécifiques pour les mesures $R^2\Delta$, optimisant ainsi la détection tout en réduisant les erreurs de type I et II, que les tests statistiques traditionnels ne parviennent pas à contrôler adéquatement.

2.8.6 Comparaison des méthodes de détection du FDI

Les données de PISA présentent des défis, notamment des données manquantes et l'absence d'items d'ancrage. Notre méthode doit être bien adaptée aux situations avec des données manquantes, sans nécessiter d'éléments d'ancrage (*anchor items*), tout en restant fiable avec des groupes de tailles inégales. Les données de PISA présentent donc la particularité de contenir des groupes de tailles inégales. Cela est essentiel dans notre étude, où le groupe de référence correspond aux élèves L1 et le groupe focal aux élèves L2, dont la taille d'échantillon est plus réduite. En effet, ces contraintes ont guidé le choix d'une méthode capable de s'adapter à ces particularités tout en maintenant la robustesse des résultats. Parmi

les méthodes couramment utilisées, plusieurs n'ont pas été retenues pour des raisons spécifiques. Le modèle de Rasch a été écarté, car il repose sur une hypothèse d'unidimensionnalité non vérifiée dans les données PISA et nécessite des items d'ancrage non biaisés, qui sont absents dans notre base de données. Le SIBTEST s'est avéré inapproprié en raison de sa sensibilité aux données manquantes, problématique fréquente dans les enquêtes PISA. Quant à l'analyse multiniveau, bien que théoriquement adaptée à la structure hiérarchique des données, son application n'a pas été retenue en raison de sa complexité et des contraintes logicielles et computationnelles. Le modèle 2PL n'a pas été retenu non plus, car son implémentation nécessite des items d'ancrage non biaisés pour une calibration correcte des échelles, éléments indisponibles dans notre jeu de données. La méthode de Mantel-Haenszel, bien que simple à implémenter, n'a pas été retenue dans notre étude, car elle ne détecte que le DIF uniforme et reste dépendante de la disponibilité d'items d'ancrage, tout en présentant des limites dans le traitement des groupes de taille inégale. Enfin, la régression logistique, bien qu'elle permette de détecter le DIF uniforme et non uniforme, a été écartée en raison des complexités d'interprétation des interactions et des contraintes analytiques qu'elle impose. Pour surmonter ces défis, nous proposons l'utilisation du pairage par score de propension, une technique efficace pour atténuer les déséquilibres entre les groupes et réduire les biais potentiels liés aux différences de taille d'échantillon. En équilibrant les covariables observées, cette méthode améliore la comparabilité entre les groupes, permettant ainsi d'obtenir des analyses plus robustes, même dans des conditions d'échantillons imparfaits. Bien que la régression logistique puisse être une méthode potentiellement appropriée pour l'analyse du FDI dans le cadre de notre étude, nous avons choisi d'utiliser des tests statistiques conventionnels pour les proportions (test de proportion) pour une analyse simple du FDI. Ce choix s'explique par le fait que notre approche ne se focalise pas principalement sur la méthode de détection du FDI en tant que telle, mais plutôt sur le pairage des données comme élément central. Dans la logique du FDI, il est essentiel de comparer les performances entre groupes uniquement à habileté égale, c'est-à-dire en tenant compte du score ou du niveau de compétence global des répondants. Le test de proportion, tel que présenté par Laurencelle (2021), constitue un outil efficace pour comparer les taux de réussite entre groupes et identifier le FDI. Concrètement, les proportions ne sont pas calculées de manière brute, mais conditionnellement aux strates d'habileté formées après le pairage. Chaque proportion reflète donc la probabilité de réussite dans un groupe donné, à niveau de compétence comparable, ce qui correspond au cœur même de la logique du FDI. En testant l'hypothèse nulle d'égalité des proportions contre l'hypothèse alternative d'une différence significative, cette méthode offre une approche simple, mais rigoureuse pour détecter des écarts potentiels dans le fonctionnement des items entre groupes pairés.

L'utilisation des tests de proportion permet une analyse simplifiée du FDI, tout en assurant une comparaison rigoureuse des groupes à l'aide de méthodes statistiques robustes et facilement interprétables. Appliquée dans ce contexte spécifique du pairage des données, cette approche est particulièrement adaptée pour explorer avec précision l'influence des variables linguistiques dans notre recherche. Pour renforcer la validité de nos résultats, nous complétons cette analyse par des tests de sensibilité. Comme l'ont démontré Liu et al. (2013), ces analyses permettent d'évaluer la robustesse des résultats dans des recherches non expérimentales en tenant compte de facteurs non contrôlés, ce qui renforce la validité de nos conclusions concernant les effets linguistiques. Dans cette perspective, le Tableau 2.2 présente de façon synthétique les raisons pour lesquelles les autres méthodes décrites précédemment n’ont pas été retenues pour l’analyse du FDI dans cette étude.

Tableau 2.2 : ***Raisons d'exclusion des méthodes de détection du FDI dans l'étude***

Méthode	Raisons d'exclusion
SIBTEST	- Sensible aux données manquantes (présentes dans PISA)
Régression logistique	- Difficultés d'interprétation liées aux interactions complexes entre variables. - Méthode non utilisée ici malgré ses avantages, en raison de la complexité analytique.
Analyse multiniveau (modèle hiérarchique)	- Exigeante en termes de ressources computationnelles, surtout dans des contextes de grande ampleur comme PISA. -L'application de cette méthode n'est pas réalisable en raison de la complexité des modèles, de la taille importante de l'échantillon, ainsi que des contraintes liées à l'accès et à la manipulation des bases de données multiniveaux de PISA.
Modèle de Rasch	- Ne permet pas de détecter le FDI non uniforme. - Repose sur l'unidimensionnalité (non-garantie dans PISA) - Impossible à appliquer sans items d'ancrage non biaisés (absents des données PISA)
Modèle 2PL (TRI)	- Requiert des items d'ancrage non biaisés pour calibrer les échelles. Ces items ne sont pas disponibles dans les données PISA
Méthode de Mantel-Haenszel (MH)	- Ne détecte que le FDI uniforme (non uniforme ignoré)[1] - Requiert des items d'ancrage fiables (absents des données PISA)

[1] Nous tenons à remercier M. Sébastien Béland pour cette précision méthodologique pertinente.

2.9 Pairage des données

Le pairage est une méthode statistique essentielle dans les études observationnelles, utilisée pour créer une situation de quasi-randomisation dans les échantillons où la randomisation n'est pas possible (Arikan et al., 2018). Introduite par Rosenbaum et Rubin en 1983, cette technique permet de réduire les biais en équilibrant les covariables entre les groupes de traitement et de contrôle. Chen et al. (2020), notamment, ont fait appel au pairage pour une détection du FDI tenant compte de covariables.

L'idée derrière cette approche est de permettre une comparaison entre les individus des groupes traités et non traités en minimisant les différences dues aux covariables. En appariant les individus à partir de scores de propension similaires, cette méthode vise à équilibrer les groupes pour que les différences observées dans les résultats soient attribuables aux traitements et non aux covariables (Olmos et Govindasamy, 2015). Le pairage par score de propension constitue une approche précieuse pour les études observationnelles, permettant de limiter les biais et de produire des comparaisons plus rigoureuses entre groupes, contribuant ainsi à renforcer la validité et la fiabilité des résultats (Wang et al., 2013).

Le pairage peut être effectué en utilisant diverses mesures estimant la similitude (ou la distance) entre les individus, la plus courante étant le score de propension. Le score de propension se définit généralement comme la probabilité de recevoir un traitement spécifique en fonction d'un ensemble de caractéristiques observables, telles que le sexe ou le statut socio-économique (Zhao et al., 2021). On peut obtenir le score de propension en effectuant une régression logistique, où la variable dépendante est une variable dichotomique indiquant le groupe (dans notre cas, L1 ou L2) et les variables indépendantes sont les covariables qu'on juge avoir une association statistique avec le fait d'appartenir à un ou l'autre groupe. Le modèle de régression logistique résultant permet d'estimer le score de propension, c'est-à-dire la probabilité d'appartenir à l'un ou l'autre groupe considérant les covariables.

2.9.1 Méthodes de pairage

Il existe différentes manières d'apparier les individus selon leur score de propension ou d'autres mesures de leur similitude. Parmi ces méthodes, le pairage exact consiste à apparier les individus possédant des valeurs identiques pour chaque covariable. Bien qu'elle assure un excellent équilibre, cette méthode peut réduire considérablement la taille de l'échantillon analysable, en particulier lorsque les covariables sont

nombreuses (Stuart, 2010), puisqu'il devient improbable de trouver pour chaque individu du premier groupe un individu identique dans le second groupe.

Une autre méthode courante est le pairage par voisin le plus proche (*nearest neighbor matching*), dans lequel chaque individu du groupe traité est apparié avec l'individu du groupe contrôle dont le score de propension est le plus proche (Arikan et al., 2018). Cette méthode, parfois désignée sous le nom de pairage avare (*greedy matching*), forme les premières paires disponibles sans procéder à une optimisation globale du pairage. Cette approche reste l'une des plus couramment utilisées, mais elle peut laisser persister des différences entre les individus appariés, particulièrement si la distribution des scores diffère entre les groupes (Arikan et al., 2018). Bien qu'efficace pour de grands échantillons, elle peut en effet produire des pairages moins précis, surtout si les distributions de scores de propension sont inégales entre les groupes (Olmos et Govindasamy, 2015).

Le pairage optimal est similaire à la méthode du voisin le plus proche, mais utilise un algorithme pour tester de nombreuses paires possibles afin de minimiser la somme des distances entre les scores de propension des paires. Cela permet d'obtenir un équilibre optimal, particulièrement utile dans les petits échantillons (Olmos et Govindasamy, 2015).

Le pairage génétique, plus récent, utilise des algorithmes évolutionnaires pour optimiser le pairage en pondérant l'importance relative des covariables, et s'avère précieux dans des contextes où la complexité des données exige une grande flexibilité (Zhao et al., 2021).

D'autres méthodes de pairage n'utilisent pas le score de propension. Le pairage exact, par exemple, implique d'apparier les individus possédant des valeurs identiques pour chaque covariable. Bien qu'il assure un excellent équilibre, cette méthode peut limiter la taille de l'échantillon analysable, surtout si les covariables sont nombreuses (Stuart, 2010). La librairie *matchit* pour le langage R (Stuart et al., 2011) permet un compromis où certaines covariables serviront à faire un pairage exact alors que d'autres covariables serviront à un pairage plus flexible, par exemple, par voisin le plus proche.

Lors du pairage, l'utilisation d'un seuil de précision appelé étrier (*caliper*) impose une limite à la différence de score de propension entre deux individus pouvant être appariés, par exemple une marge maximale de 0,2 écart-type. Cela permet de minimiser le risque de différences résiduelles entre paires, bien que cette

méthode puisse aussi limiter le nombre de paires possibles (Zhao et al., 2021; Olmos et Govindasamy, 2015).

Le pairage présente plusieurs avantages en rendant les groupes plus comparables et en augmentant ainsi la validité interne des analyses (Austin, 2011). Cependant, ses limites incluent l'incapacité de contrôler les variables non observées et la possibilité de réduire la taille des échantillons analysables, ce qui peut diminuer la puissance statistique de l'étude (Arikan et al., 2018; Zhao et al., 2021). De plus, certaines méthodes peuvent s'avérer techniquement complexes à mettre en œuvre, surtout si les distributions de scores de propension ne se chevauchent pas suffisamment entre les groupes (Olmos et Govindasamy, 2015).

2.10 Synthèse

Ce chapitre a abordé plusieurs concepts clés essentiels à notre recherche sur les biais linguistiques dans l'évaluation des mathématiques. Nous avons commencé par justifier notre choix de concentrer l'étude sur les mathématiques au secondaire et exploré les facteurs influençant la réussite dans cette matière. L'utilisation des données canadiennes de PISA 2012 a été expliquée, soulignant la pertinence de ce contexte bilingue et multiculturel pour notre étude. Un aperçu historique de l'influence de la langue sur l'apprentissage des mathématiques a été présenté, mettant en lumière les défis spécifiques rencontrés par les apprenants de langue seconde. Nous avons examiné la littérature scientifique traitant de l'impact des caractéristiques linguistiques des items sur la performance des élèves en mathématiques. Nous avons également défini l'évaluation des apprentissages et son importance, puis exploré le concept de validité des scores en évaluation. Nous avons ensuite examiné les notions de biais dans les tests, en nous concentrant particulièrement sur les biais linguistiques. Un concept central de notre étude, le fonctionnement différentiel des items (FDI), a été expliqué en détail, y compris son historique et les méthodes courantes d'analyse telles que Mantel-Haenszel, le modèle de Rasch, SIBTEST, la régression logistique, l'analyse FDI multiniveau et le modèle 2PL (TRI). Nous avons discuté des avantages et des limites de ces méthodes, justifiant ainsi notre choix d'utiliser une approche de pairage basée sur le score de propension, plus adaptée à notre contexte de recherche.

Ces concepts sont directement liés à notre objectif principal : identifier le FDI d'origine linguistique dans les items mathématiques de PISA 2012 au Canada. Notre étude vise à révéler les disparités potentielles affectant les élèves de langue seconde lors des évaluations. Pour ce faire, nous examinerons les écarts de

performance aux items entre les groupes L1 et L2, tout en développant une méthodologie de pairage adaptée aux données complexes de PISA, contribuant ainsi à combler une lacune méthodologique dans la littérature existante. Pour atteindre l'objectif de cette recherche, nous avons défini des objectifs spécifiques :

Premier objectif spécifique : Examiner le fonctionnement différentiel des items (FDI) selon le statut linguistique (L1/L2). Cela nous permettra de comprendre si les items avantagent ou désavantagent les élèves selon leur groupe linguistique auquel ils appartiennent, indépendamment de leurs véritables capacités en mathématiques.

Deuxième objectif spécifique : Élaborer une approche *« de pairage des groupes »* ou *« matching »* pour examiner le FDI dans les enquêtes internationales à grande échelle, telles que PISA, caractérisées par leur complexité et la présence de nombreuses données manquantes. Cette méthodologie n'exige pas d'« anchor items » et doit rester fiable avec des tailles d'échantillon relativement petites. L'élaboration de cette méthodologie est essentielle pour garantir la précision et la fiabilité de nos analyses.

CHAPITRE 3

MÉTHODOLOGIE

L'objectif général de cette étude est d'analyser l'impact potentiel du statut linguistique des élèves sur leurs performances aux évaluations de mathématiques de l'enquête PISA 2012 au Canada. Plus précisément, nous cherchons à déterminer si des FDI d'origine linguistique existent dans ces évaluations, susceptibles de désavantager les élèves apprenant dans une langue seconde (L2) par rapport à ceux dont la langue première est la langue d'enseignement (L1). Cette recherche vise à alimenter le débat sur l'équité des évaluations standardisées dans un contexte de diversité linguistique en expansion, tout en abordant les défis méthodologiques inhérents à la conception par livret (*booklet design*) de l'enquête PISA. En outre, nous explorons l'efficacité des approches de pairage pour identifier le fonctionnement différentiel d'items, afin d'améliorer la précision et la fiabilité de l'analyse des biais linguistiques dans les évaluations internationales.

Nous avons vu dans la revue de littérature que plusieurs études ont abordée la présence potentielle de biais linguistiques dans les évaluations de mathématiques de l'enquête PISA, biais qui pourraient désavantager les élèves apprenant dans une langue seconde (L2) par rapport à leurs pairs, dont la langue première est la langue d'enseignement (L1). Toutefois, les études menées dans le contexte canadien demeurent peu nombreuses. Par conséquent, il est possible de s'appuyer sur les méthodologies précédemment utilisées dans ces études pour déterminer celle qui sera employée dans cette recherche qui concerne le Canada.

Dans cette section, nous examinerons les éléments méthodologiques mis en place pour étudier la question de recherche. Cette approche vise à répondre de manière rigoureuse aux objectifs de l'étude. Tout d'abord, la stratégie analytique impliquée dans l'étude est décrite, suivie d'une présentation de la description des données et des variables utilisées. Ensuite, cette section se conclut par une description des procédures d'analyse employées pour interpréter les résultats de cette étude.

3.1 Stratégie analytique

La présente étude se concentre sur l'analyse du fonctionnement différentiel des items (FDI) dans le contexte des épreuves de mathématiques de l'enquête PISA 2012. Cette édition du PISA avait placé les

mathématiques au cœur de son évaluation, offrant ainsi un terrain d'étude riche pour notre recherche. L'échantillon étudié se compose d'adolescents canadiens âgés de 15 ans, stratifiés selon leur statut linguistique en deux catégories : langue maternelle (L1) et langue seconde (L2). Cette classification permet d'examiner les potentielles différences de performance liées au contexte linguistique des élèves.

La méthodologie adoptée repose principalement sur la méthode de pairage pour l'analyse du FDI. Cette approche est particulièrement pertinente pour notre objectif de recherche, car elle offre la possibilité de prendre en compte les variables confondantes susceptibles d'influencer l'identification du FDI. Elle permet de contrôler les facteurs potentiels de confusion, de comparer des élèves ayant des niveaux de compétence équivalents et d'isoler les effets spécifiques au FDI de l'item. De plus, elle présente l'avantage d'être robuste face aux données manquantes, une caractéristique fréquente dans les enquêtes internationales qui, comme PISA, utilisent une conception par livret de test.

Par la suite, pour évaluer rigoureusement l'impact du pairage, nous avons mené une analyse comparative systématique des résultats avant et après pairage. Cette démarche s'est articulée autour de deux axes principaux: la réalisation de tests de proportion comparant les taux de réussite entre les groupes L1 et L2 à chaque étape, et une analyse de sensibilité pour évaluer la robustesse du FDI observé face à des covariables non observées. La comparaison des taux de réussite entre les deux groupes linguistiques a été effectuée à deux étapes distinctes : avant et après le processus de pairage. Cette double analyse nous permet d'évaluer l'impact du pairage sur l'identification du FDI. En d'autres termes, cela permet de vérifier si, à niveau de compétence égale, les élèves du groupe L2 sont désavantagés par rapport à ceux du groupe L1 sur certains items. La comparaison entre avant et après le pairage permet de mesurer si les différences initiales dans les performances sont dues à des différences globales de compétence qui seront prélevées après l'étape de pairage ou à un FDI propre à l'item lui-même.

Les résultats de notre étude seront présentés de manière détaillée et structurée. Nous dresserons des listes exhaustives des items de mathématiques du PISA 2012, chacun accompagné de statistiques précises indiquant la présence ou l'absence de FDI. Ces statistiques offrent une vue d'ensemble complète et nuancée du fonctionnement différentiel des items selon le statut linguistique des élèves canadiens.

3.2 Description des données et des variables

L'analyse de cette recherche repose sur les données issues de l'enquête PISA 2012. Elle s'inscrit dans le cadre d'une analyse secondaire, utilisant des données existantes sans collecte de nouvelles informations, ce qui oriente les choix méthodologiques et les types d'analyses possibles. Cette section présente d'abord les caractéristiques générales du dispositif PISA et de l'échantillon retenu, avant de détailler les variables utilisées pour l'étude et les choix de recodage effectués afin de répondre aux objectifs de recherche.

3.2.1 Présentation des données PISA 2012

Le Programme international pour le suivi des acquis des élèves (PISA) est une évaluation internationale à grande échelle menée par l'Organisation de coopération et de développement économiques (OCDE) tous les trois ans depuis 2000 (OCDE, 2013). Cette enquête vise à mesurer les compétences des jeunes de 15 ans dans divers domaines, notamment en mathématiques, en lecture et en sciences. Le PISA se distingue par sa conception innovante de l'évaluation qui va au-delà de la simple restitution des acquis scolaires. En effet, il cherche à déterminer la capacité des élèves à s'appuyer sur leurs connaissances et à les appliquer dans de nouvelles situations inspirées du monde réel (OCDE, 2013).

L'ensemble de données qui résulte du PISA permet aux chercheurs d'étudier la réussite scolaire et l'appartenance à un groupe sous différents angles (OCDE, 2014). Dans cette étude, l'échantillon canadien du PISA 2012 a été utilisé. 84 items codés ou recodés de manière dichotomique de l'évaluation des mathématiques du PISA 2012 ont été inclus pour détecter les effets du FDI.

La structure de l'évaluation PISA 2012 est complexe et rigoureuse. Les épreuves, d'une durée totale de deux heures par élève, sont principalement de type « papier-crayon » et comprennent à la fois des items à choix multiples et des questions ouvertes nécessitant une argumentation. La batterie complète d'items représente 390 minutes de test, réparties en 13 carnets différents *« booklet design »*. Chaque élève répond uniquement à une fraction des items, selon diverses combinaisons. Dans certains pays, des épreuves informatisées supplémentaires de 40 minutes en mathématiques et en compréhension de l'écrit ont également été administrées (OCDE, 2013). Au Canada, le PISA 2012 a été mis en œuvre par l'entremise d'une évaluation sur papier (CMEC, 2013).

Pour le cycle PISA 2012 au Canada, l'échantillon comprend environ 21 000 élèves de 15 ans, issus d'environ 900 écoles réparties dans les 10 provinces canadiennes. L'évaluation a été menée en anglais et en français,

selon le système scolaire respectif de chaque région (CMEC, 2013). Cette large couverture géographique et linguistique assure une représentativité significative de la diversité éducative canadienne.

Les données PISA sont riches et multidimensionnelles, incluant non seulement les résultats des épreuves cognitives, mais aussi une vaste gamme d'informations contextuelles. Ces dernières sont recueillies par le biais de questionnaires administrés aux élèves et aux chefs d'établissement, d'une durée d'environ 30 minutes chacun. Le questionnaire destiné aux élèves couvre des aspects tels que leur milieu familial, leurs attitudes envers l'apprentissage, leurs habitudes et leur mode de vie à l'école et à la maison. Le questionnaire destiné aux directions d'établissement porte sur des éléments comme la qualité des ressources de l'école, le mode de gestion et de financement, les processus de prise de décision et les contenus des programmes d'enseignement (OCDE, 2013). En plus de ces questionnaires principaux, le PISA 2012 propose trois questionnaires optionnels : un sur la maîtrise de l'informatique, qui explore l'accès et l'usage des technologies de l'information et de la communication (TIC) par les élèves ; un sur le parcours scolaire, qui recueille des informations sur d'éventuelles interruptions de scolarité et la préparation à la carrière professionnelle ; et un questionnaire destiné aux parents, qui aborde des thématiques telles que leur opinion sur l'établissement fréquenté par leur enfant et leur implication dans la vie scolaire, le soutien qu'ils apportent à l'apprentissage de leur enfant, leurs attentes concernant la future profession de leur enfant, et leur éventuelle ascendance allochtone (OCDE, 2013). Pour la présente étude, les données ont été obtenues par téléchargement à partir du site officiel de l'OCDE, qui les met à disposition à des fins de recherche. Il est important de noter que ces données sont anonymisées et ne contiennent aucune information permettant d'identifier un répondant en particulier, conformément aux normes éthiques de la recherche. Pour garantir la qualité et l'intégrité des données obtenues, une procédure de vérification a été mise en place. Celle-ci a consisté en une visualisation minutieuse des données, permettant de s'assurer de leur correspondance avec la documentation détaillée fournie par l'OCDE. Les données incluses dans cette étude proviennent du PISA. Par conséquent, l'étude n'a pas nécessité d'approbation éthique. Enfin, le choix des données du test administré en anglais était motivé par deux facteurs principaux : la plus grande quantité de données disponibles et la possibilité de comparaison avec la littérature récente sur le FDI, comme exemple les travaux de Liu et Bradley (2021) et de Buono et Jang (2021), qui se concentrent sur les tests en anglais.

3.2.2 Variables utilisées et recodage

Cette étude utilise diverses variables issues des données PISA 2012, certaines disponibles directement, d'autres calculées ou recodées selon les besoins de l'analyse. Le premier groupe de variables est constitué des items de mathématiques. Un processus de recodage a été appliqué aux items non dichotomiques afin de simplifier l'analyse du FDI. Tous les items ont été recodés en utilisant une échelle binaire : 0 pour l'échec et 1 pour le succès. Ce recodage a été effectué de manière systématique, y compris pour les items partiellement réussis ou non complétés. Ainsi, les items pour lesquels l'élève n'avait obtenu qu'une partie des points ont été considérés comme des échecs et codés 0. De même, les items laissés en blanc alors qu'ils figuraient dans le livret de l'élève ont également été codés 0, signifiant un échec. Cette approche de dichotomisation a été choisie, car elle simplifie considérablement l'analyse du FDI, permettant une évaluation plus directe des différences de performance entre les groupes linguistiques. Il est à noter que cette méthode de recodage est couramment utilisée dans la littérature sur l'analyse du FDI, renforçant ainsi la pertinence et la comparabilité de notre approche méthodologique avec d'autres études dans ce domaine. Nous citons à ce titre l'article d'Akcan et Kabasakal (2023), qui ont également dichotomisé les items pour analyser le FDI.

Nous avons de plus créé une variable dichotomique pour le statut d'immigration des élèves (IMMIG_DICHO). Les élèves de premières et deuxièmes générations d'immigrants ont été classés comme immigrants, tandis que les élèves nés au Canada ont été classés comme « non immigrants ». Ce regroupement visait à faciliter la procédure de pairage, car il permet de réduire la variabilité entre les groupes comparés, permettant un pairage plus précis des élèves immigrants et non immigrants sur des caractéristiques similaires, conformément aux exigences de la méthode de correspondance par score de propension.

La variable principale de regroupement est le statut linguistique (STATUT_LING), que nous avons créé à partir de la question contextuelle sur la langue parlée à la maison (ST25Q01). Elle comporte deux niveaux possibles : L1 = la langue à la maison est la même que la langue du test et L2 = la langue à la maison est une autre langue. Le groupe focal est constitué des élèves dont la langue du test est une langue seconde (L2), tandis que le groupe de référence est composé des élèves dont la langue du test est leur langue maternelle (L1). Cette variable de regroupement permet d'identifier les sous-groupes d'élèves pour lesquels nous souhaitons analyser le FDI (Differential Item Functioning), et c'est à partir de cette distinction que le pairage et l'analyse de FDI seront effectués.

Plusieurs covariables ont été utilisées pour le pairage afin de contrôler les variables de confusions potentielles et d'isoler les effets spécifiques au FDI de l'item. Elles ont été employées pour réduire les variations entre les élèves lors de l'identification du FDI. Ces covariables ont été sélectionnées, car elles se sont avérées être des prédicteurs importants influençant la performance en mathématiques des élèves, comme démontré dans le chapitre précédent. Le Tableau 1 présente le sommaire des variables utilisées pour contrôler les facteurs potentiels de confusion et isoler les effets spécifiques au FDI dans l'analyse.

Au niveau de l'élève, les variables individuelles suivantes ont été appariées : L'âge (AGE) représente l'âge de chaque élève participant à l'étude, tandis que le sexe (ST04Q01) indique si l'élève est masculin ou féminin. Le livret auquel l'élève a répondu dans l'enquête PISA (BOOKID) est utilisé pour contrôler la variabilité des questions et garantir que l'on compare des élèves qui ont répondu aux mêmes items.

Les variables familiales de l'élève comprennent également plusieurs éléments importants. Le statut socio-économique (ESCS) est un indice composite qui combine des informations liées aux facteurs économiques, sociaux et culturels, incluant l'éducation des parents, l'occupation et les possessions du ménage. Le niveau d'éducation de la mère (MISCED) et du père (FISCED) est représenté par des catégories allant de 0 à 6: (0 = 'None' = Aucun), (1= 'ISCED 1' = enseignement primaire), (2 = 'ISCED 2' = enseignement secondaire inférieur), (3 = 'ISCED 3B, C'= enseignement secondaire supérieur), (4 = 'ISCED 3A, ISCED 4' = enseignement post-secondaire non universitaire), (5 = 'ISCED 5B' = maîtrise ou équivalent), et (6 = 'ISCED 5A, 6' = doctorat ou équivalent, ou post-doctorat). Le statut d'immigration (IMMIG) représente la situation de l'élève par rapport à son statut d'immigration, avec les catégories : 1= Native, 2 = Deuxième génération, et 3 = Première génération.

Au niveau de l'école, plusieurs variables ont été étudiées comme sources potentielles de FDI après contrôle des variables de l'élève. La langue dans laquelle le test est administré (TESTLANG) indique le système scolaire anglophone ou francophone avec les catégories : 313 = Anglais; 493 = Français. Le Tableau 3.1 présente la nature des variables retenues pour cette étude. Celles-ci ont été identifiées afin de guider leur intégration dans le processus de pairage et d'assurer que les comparaisons entre groupes reposent sur des données correctement codées et structurées.

Tableau 3.1 : *Variables utilisées en plus des items de mathématiques*

Variable	Description	Codage	Nature
ST04Q01	*Gender* : sexe de l'élève (masculin/féminin)	1-Féminin 2-Masculin	Discrète, dichotomique
AGE	*age of student* : âge de l'élève	Valeur numérique réelle	Continue
ST25Q01	*Language at home* : langue principalement parlée à la maison.	1-Langue du test 2- Autre langue	Discrète, dichotomique
ESCS	*Index of economic, social and cultural status* : indice regroupant éducation, occupation des parents et biens du ménage.	Valeur numérique composite	Continue
MISCED	*Educational level of Mother* : niveau d'éducation de la mère.	0- Aucun 1-ISCED1 [2] 2-ISCED 2 [3] 3-ISCED 3B, C [4] 4-ISCED 3A, ISCED 4 [5] 5-ISCED 5B [6] 6-ISCED 5A, 6 [7]	Discrète, ordinale polychotomique
FISCED	*Educational level of Father* : niveau d'éducation du père.	0- Aucun 1- ISCED 1 2- ISCED 2 3- ISCED 3B, C 4- ISCED 3A, ISCED 4 5-ISCED 5B 6- ISCED 5A, 6	Discrète, ordinale polychotomique
IMMIG	*Immigration status* : statut d'immigration de l'élève.	1-Natif 2-Deuxième génération 3-Première génération	Discrète, catégorielle polychotomique
TESTLANG	*Language of Test* : langue du test administré.	313- Anglais 493- Français	Discrète, dichotomique
BOOKID	*Booklet* : livret auquel l'élève a répondu.	1 à 13	Discrète, identifiant numérique

[2] Enseignement primaire
[3] Enseignement secondaire inférieur
[4] Enseignement secondaire supérieur
[5] Enseignement post-secondaire non universitaire ou universitaire
[6] Maîtrise ou équivalent
[7] Doctorat ou équivalent ou post-doctorat

Les variables continues (par exemple, l'âge et l'indice ESCS) ont été traitées comme telles dans le modèle, tandis que les variables catégorielles et ordinales ont été codées de façon appropriée selon leur structure. Afin d'évaluer l'équilibre obtenu après le pairage, des courbes de densité ont été générées pour les variables continues et des diagrammes en barres ont été utilisés pour les variables discrètes. Une fois le pairage effectué, les tests de proportion ont été appliqués sur les items dichotomisés afin de comparer les taux de réussite entre les deux groupes. À cette étape, le rôle des covariables est indirect, mais demeure essentiel, puisqu'elles ont permis de neutraliser l'effet de facteurs potentiellement confondants dans la comparaison des groupes.

3.2.3 Variables considérées, mais non utilisées

Dans la revue de littérature du chapitre précédent, nous avons identifié d'autres variables ayant un impact potentiel sur la performance des élèves en mathématiques. Au niveau individuel, par exemple, l'auto-efficacité en mathématiques, mesurée par la variable MATHEFF, reflète le degré de confiance et la perception de compétence des élèves concernant leurs habiletés mathématiques. Parallèlement, l'anxiété liée aux mathématiques, représentée par la variable ANXMAT, indique le stress ressenti par les élèves lorsqu'ils abordent cette matière. De même, la motivation pour l'apprentissage des mathématiques, mesurée par INSTMOT, permet d'évaluer l'intérêt et l'engagement des élèves envers cette discipline. Enfin, les items de lecture, qui peuvent refléter la maîtrise de la langue, car la lecture et la compréhension d'un texte sont souvent considérées comme les premières formes de maîtrise linguistique essentielles aux fonctions cognitives (Chen, 2010). Du côté de l'établissement scolaire, des variables contextuelles jouent également un rôle. La variable STUDREL mesure la perception des élèves quant à leurs relations avec leurs enseignants, tandis que la variable MTSUP évalue le soutien et l'attitude des enseignants envers l'apprentissage des mathématiques. Cependant, dans les données canadiennes de PISA 2012, environ 35 % des informations relatives à ces variables sont manquantes. Par conséquent, nous avons décidé de ne pas inclure ces variables comme covariables dans le processus de pairage.

De même, les items de lecture du PISA auraient pu contribuer à notre démarche en aidant à isoler l'effet de la langue sur les questions de mathématiques. Toutefois, pour l'édition 2012 du PISA, environ 31 % des élèves du Canada n'avaient pas de questions de lecture dans leur cahier, ce qui aurait entrainé une perte

importante de données pour notre étude puisqu'il aurait fallu exclure ces élèves. Nous n'avons donc pas inclus d'items de lecture ou de dérivés de ces items comme covariables.

3.2.4 Test de proportion

Selon Laurencelle (2021), le test de proportion est une méthode statistique fondamentale qui permet d'évaluer s'il existe une différence significative entre les proportions de deux groupes. Ce test repose sur une hypothèse nulle, qui stipule qu'il n'y a pas de différence significative entre les proportions des groupes comparés, tandis que l'hypothèse alternative suggère qu'une différence existe (Laurencelle, 2021). Pour être valide, le test nécessite des données catégorielles, par exemple des données binaires ou dichotomiques (succès/échec) et des observations indépendantes dans chaque groupe. Le test de proportion nécessite généralement un échantillon de taille suffisamment grande pour que l'approximation normale soit valide. Il implique le calcul de la valeur p pour déterminer si les différences observées sont dues au hasard. Un seuil de signification (souvent $\alpha = 0,05$) est fixé pour décider si l'hypothèse nulle doit être rejetée (Laurencelle, 2021).

Fielding (1974) souligne que le test de proportion est particulièrement utile dans diverses applications pratiques, notamment pour l'analyse des tableaux de contingence. Il met en avant l'importance du rapport de cotes (odds ratio) comme mesure d'association dans les comparaisons entre groupes. Selon Fan et ses collaborateurs (2017), le rapport de cotes est une mesure utilisée pour quantifier l'association entre deux variables. Il permet de comparer la probabilité qu'un événement se produise dans chaque groupe. Un odds ratio supérieur à 1 suggère que l'événement est plus probable dans un groupe par rapport à l'autre, tandis qu'un odds ratio inférieur à 1 indique que l'événement est moins probable. Bien que cette mesure donne une indication de l'intensité de l'association entre deux variables, elle peut être moins intuitive à interpréter, surtout lorsqu'elle est supérieure à 1.

Laurencelle (2021) met en évidence plusieurs avantages majeurs du test de proportion. Il est relativement simple à comprendre et à appliquer, ce qui le rend accessible même pour les chercheurs débutants. De plus, il offre des résultats faciles à interpréter, ce qui facilite la prise de décision. Il est également robuste face à certaines violations de la normalité, notamment lorsqu'il est appliqué à de grands échantillons. Cependant, Laurencelle (2021) identifie plusieurs limitations importantes. Le test est particulièrement sensible à la taille de l'échantillon, les petits échantillons pouvant produire des résultats moins fiables. Il repose également sur l'hypothèse d'indépendance des échantillons, qui n'est pas toujours vérifiée en

pratique. Sa limitation principale réside dans le fait qu'il est conçu essentiellement pour comparer deux groupes, ce qui rend nécessaires d'autres approches pour des conceptions plus complexes.

3.2.5 Analyse de sensibilité

Dans les recherches non expérimentales, où il n'est pas toujours possible de contrôler chaque facteur susceptible d'altérer les conclusions, l'analyse de sensibilité aide les chercheurs à tester la robustesse des résultats face à des variables non mesurées. Elle est couramment employée dans les sciences sociales pour évaluer si des biais potentiels pourraient compromettre la validité des inférences causales (Liu et al., 2013).

Pour des données dichotomiques appariées, l'analyse de sensibilité se concentre sur les paires discordantes, c'est-à-dire les paires où l'individu du groupe focal et son équivalent dans le groupe apparié donnent des réponses différentes. Suivant Rosenbaum (2002b), on définit T+ comme les paires où l'individu du groupe focal a une réponse positive alors que son équivalent dans le groupe apparié a une réponse négative, et T− comme les paires où c'est l'individu apparié qui a une réponse positive et le groupe focal une réponse négative. Sous l'hypothèse nulle, le nombre de paires T+ devrait être équivalent au nombre de paires T-. Cette hypothèse est vérifiée à l'aide d'un test binomial. Ensuite, pour évaluer la robustesse des résultats face à des facteurs confondants non mesurés, on introduit un paramètre de sensibilité, Γ (gamma), qui représente le niveau maximal de biais que les résultats peuvent supporter sans perdre leur signification statistique, les chercheurs peuvent estimer dans quelle mesure un résultat pourrait être influencé par des variables non observées (Hasegawa et Small, 2017). À mesure que Γ augmente, on ajuste les probabilités associées à T+ et T- pour simuler un possible déséquilibre causé par une covariable non observée. L'objectif est de déterminer la plus grande valeur Γ pour laquelle la différence observée entre les groupes demeure statistiquement significative. Cette valeur maximale de Γ indique jusqu'à quel point les conclusions sont robustes face à d'éventuels facteurs confondants. Par exemple, un test significatif sous un Γ de 1,5 indiquerait que, même si une covariable faisait augmenter les chances du groupe focal de répondre correctement à un item de 50 % par rapport au groupe de référence (rapport des cotes de 1,5), l'effet différentiel observé resterait significatif. Cela signifie que l'item présente un fonctionnement différentiel (DIF) qui ne peut pas être entièrement expliqué par ce niveau de biais potentiel.

Comme le précise Rosenbaum (2015), l'analyse de sensibilité remplace les affirmations qualitatives concernant l'existence de biais non mesurés par une quantification objective du niveau de biais nécessaire

pour modifier les conclusions d'une étude observationnelle. L'étude de Hasegawa et Small (2017) souligne que l'analyse de sensibilité est particulièrement utile lorsque les biais potentiels, ou « biais de confusion », ne sont pas directement observables. Cette analyse consiste à ajuster les résultats en fonction de divers scénarios hypothétiques afin de mesurer dans quelle mesure les conclusions de l'étude changeraient en présence de tels biais.

Les avantages de l'analyse de sensibilité sont nombreux. Elle renforce la crédibilité des conclusions en confirmant leur robustesse malgré la présence de biais potentiels non observés, notamment dans les études non expérimentales (Keele, 2010). Sa flexibilité lui permet de s'appliquer à divers modèles et contextes d'étude, ce qui en fait un outil polyvalent pour évaluer la fiabilité de résultats obtenus à partir de données appariées (Keele, 2010). De plus, l'analyse de sensibilité offre aux chercheurs la possibilité de quantifier l'impact potentiel de biais non mesurés en testant différentes valeurs de Γ, ce qui apporte une perspective nuancée sur les conclusions des études (Liu et al., 2013). Cependant, l'analyse de sensibilité présente aussi certaines limites. La complexité de sa mise en œuvre nécessite une solide maîtrise des modèles statistiques avancés, ce qui peut représenter un obstacle pour les utilisateurs moins expérimentés (Hasegawa et Small, 2017). En outre, comme l'efficacité de cette analyse dépend de la qualité des variables confondantes observées, des données insuffisantes peuvent entraîner des conclusions biaisées et même trompeuses (Keele, 2010).

3.3 Procédure d'analyse

Cette étude vise à déterminer si 84 items de l'évaluation PISA 2012 en mathématiques présentent un fonctionnement différentiel des items (FDI) entre les élèves de langue première (L1) et ceux de langue seconde (L2). La population ciblée est composée d'élèves canadiens ayant passé l'examen de mathématiques PISA 2012 en anglais ou en français. Les données des deux langues du test ont été analysées séparément. Le groupe focal inclut les étudiants de langue seconde (L2), tandis que le groupe de référence est formé d'étudiants de langue première (L1). Ces deux groupes ont été définis en fonction des réponses au questionnaire de l'élève administré avec le test, de sorte qu'ils ne se distinguent que par leur statut linguistique, soit en tant que langue première ou langue seconde. Pour garantir que les résultats de FDI ne sont pas influencés par des différences de compétence entre les groupes, un échantillonnage apparié a été utilisé. Cette technique associe chaque candidat du groupe focal (L2) à un ou plusieurs candidats du groupe de référence (L1) ayant des scores similaires. Pour chaque L2 ayant un score donné, plusieurs L1 ayant le même score sont tirés au hasard pour former un échantillon apparié. Cette méthode

vise à obtenir des distributions équivalentes des scores de compétence en mathématiques, assurant ainsi que les habiletés sont évaluées de manière comparable dans les deux groupes.

Notre méthode s'inspire des travaux de Chen et Zumbo (2020), bien qu'ils aient utilisé des items d'ancrage et les modèles linéaires mixtes pour l'analyse du FDI. Notre approche, quant à elle, repose sur un pairage par score de propension pour former des groupes comparables, équilibrant ainsi les covariables des deux groupes. Notre première étape consiste à stratifier les élèves selon leur statut linguistique (L1 ou L2), basé sur la question contextuelle de langue parlée à la maison (ST25Q01). Le groupe de référence (L1) comprend les élèves pour lesquels la langue du test correspond à leur langue maternelle, tandis que le groupe focal (L2) inclut ceux pour lesquels la langue du test est une langue seconde.

3.3.1 Pairage

La procédure de pairage a été réalisée avec la librairie *matchit* pour le langage R (Stuart et al., 2011). Le livret de l'évaluation (BOOKID) a été utilisé comme variable de pairage exact, c'est-à-dire que pour chaque élève L2, un élève L1 devait être sélectionné parmi ceux ayant eu le même livret de test. Cela garantit que chaque étudiant apparié répondait aux mêmes items que son pair. Les covariables utilisées incluaient l'âge (AGE), le sexe (ST04Q01), le statut socio-économique (ESCS), le niveau d'éducation de la mère et du père (MISCED et FISCED), et le statut d'immigration (IMMIG_DICHO). Pour ces covariables, des techniques de pairage optimal ont été appliquées. Le pairage optimal prend en compte la distance globale pour créer des groupes équilibrés. Le pairage a été réalisé selon un ratio 1:1 sans remise, en utilisant un caliper de 0,2 pour garantir que les élèves appariés étaient comparables sur toutes les covariables observables. Cette procédure assure que les différences observées entre les groupes L1 et L2 reflètent davantage un effet linguistique qu'une disparité liée aux covariables sociodémographiques

3.3.2 Évaluation de la qualité du pairage

L'évaluation de la qualité du pairage constitue une étape essentielle pour garantir la validité des analyses comparatives entre les groupes L1 et L2, notamment dans le cadre de l'analyse du FDI. Conformément aux travaux de Zhang et al. (2019) et aux recommandations de Chen et Zumbo (2020), nous avons adopté une approche méthodologique combinant plusieurs techniques complémentaires.

Premièrement, une fois le pairage effectué, les chercheurs examinent l'équilibre des distributions des covariables afin de vérifier si le processus a permis de réduire les différences initiales entre le groupe focal et le groupe de référence. Pour cela, des analyses graphiques ont été réalisées, incluant des diagrammes à barres pour les variables catégorielles ou nominales, ainsi que des courbes de densité pour les variables continues. Ces figures présentent les distributions de chaque covariable dans les deux groupes (L1 et L2), avant et après le pairage, et permettent d'évaluer visuellement si les différences initiales ont été atténuées. Comme le suggère Chen et Zumbo (2020), cette évaluation visuelle des distributions globales des covariables permet d'identifier d'éventuels déséquilibres persistants. De plus, toujours selon la méthode de Zhang et al. (2019), l'équilibre des covariables entre les groupes L1 et L2 a également été vérifié à l'aide de diagrammes de type Love plots, qui affichent les différences de moyennes standardisées des covariables ainsi que les rapports de variances, fournissant ainsi une évaluation complémentaire de l'équilibre atteint.

Deuxièmement, Rubin (2001) propose des diagnostics plus précis, basés sur des critères quantitatifs stricts, incluant la mesure des différences standardisées entre les moyennes des covariables, les rapports de variances entre les groupes, ainsi que l'évaluation de la variance des résidus. Pour que le pairage soit considéré comme efficace, les écarts moyens standardisés doivent généralement être inférieurs à 0,25, tandis que les rapports de variances doivent se situer entre 0,5 et 2 pour chaque covariable.

Dans notre étude, nous appliquons ces recommandations en recourant à des outils graphiques tels que les courbes de densité, les diagrammes en barres et les Love plots pour visualiser les différences standardisées de moyenne avant et après le pairage. De plus, nous analysons l'évolution des rapports de variance afin d'évaluer l'effet du pairage sur la distribution des covariables (Stuart et Rubin, 2008).

3.3.3 Tests de proportion et analyse de sensibilité

L'approche retenue reconnaît que les résultats ne sont pas une vérité absolue, mais représentent une estimation des différences de performance entre élèves L1 et L2 dans le contexte de cette étude. Cette perspective s'inscrit dans une posture post-positiviste, cohérente avec l'usage de méthodes quantitatives telles que le pairage et les tests de proportion, tout en considérant que les interactions sociales et linguistiques peuvent moduler les résultats observés. Nous avons appliqué des tests de proportion sur échantillons indépendants pour comparer les taux de réussite des items entre les groupes L1 et L2 avant pairage, puis avons répété cette procédure sur les données pairées. Ces tests offrent une manière simple

de détecter des différences significatives entre les taux de réussite selon le statut linguistique. Bien que ces tests permettent d'identifier des différences statistiquement significatives ($p < 0,05$), nous avons complété cette analyse par le calcul des tailles d'effet *h* de Cohen, suivant les recommandations de Cohen (1988) et des recherches récentes en méthodologie (Sullivan et Feinn, 2012; Wasserstein et al., 2019). Concrètement, nous avons retenu un seuil minimal de $h \geq 0,20$, correspondant à un effet de petite taille selon les paramètres standards, pour identifier les écarts pertinents sur le plan pratique. Cette approche duale (valeur p + taille d'effet) répond aux limites soulevées par Lakens (2013) concernant l'interprétation isolée des valeurs p, en permettant de distinguer les différences statistiquement significatives, mais d'ampleur négligeable ($p < 0,05$, mais $h < 0,20$) de celles présentant une réelle importance éducative ($p < 0,05$ et $h \geq 0,20$). Ainsi, notre protocole garantit que seules les disparités à la fois significatives et substantielles entre les groupes linguistiques sont prises en compte dans l'interprétation des résultats.

Pour les items ayant affiché une différence significative entre les groupes sur les données pairées, nous avons donné suite avec une analyse de sensibilité (Rosenbaum, 2005). Cette démarche, également employée par Chen et al. (2020), visait à évaluer la robustesse du FDI observé face à des covariables non observées. Rappelons que cette analyse permet d'évaluer si la différence observée (FDI) reste significative en tenant compte de l'influence potentielle de covariables non mesurées. Elle estime, en logit (logarithme du rapport des cotes), l'influence qu'une covariable non observée devrait avoir pour expliquer complètement le FDI, cette force étant exprimée par un paramètre Γ (gamma). L'analyse de sensibilité par la méthode du gamma (Γ) est une préoccupation centrale dans les études observationnelles (Rosenbaum, 2002a). Le seuil $\Gamma = 1$ indique l'absence de biais caché, tandis qu'un Γ significativement supérieur à 1 ($p < 0,05$) suggère qu'une variable omise devrait modifier les rapports de cotes d'au moins cette valeur pour remettre en cause la significativité des résultats (DiPrete et Gangl, 2004). Le gamma indique ainsi dans quelle mesure les résultats sont sensibles à l'impact de facteurs confondants non pris en compte. Cette approche, développée initialement en épidémiologie (Rosenbaum, 2010), s'appuie sur le principe que des Γ élevés signifient que la différence observée est peu susceptible d'être expliquée par une variable omise, ce qui renforce la robustesse des conclusions. Dans notre étude, nous avons retenu un seuil de $\Gamma = 1,2$ pour identifier un fonctionnement différentiel (DIF) significatif. Ce choix s'appuie sur les travaux de Rosenbaum (2002a), qui démontrent qu'un $\Gamma \geq 1,2$ indique qu'une variable non mesurée devrait modifier les rapports de cotes d'au moins 20% pour remettre en question nos conclusions.

3.4 Synthèse

En somme, notre méthodologie s'inspire de Chen et al. (2020) et permet, en utilisant stratégiquement le pairage, une comparaison approfondie des performances en tenant compte des variables confondantes potentielles. Nous visons ainsi à produire une estimation plus précise du fonctionnement différentiel des items (FDI) lié à la langue, contribuant à améliorer la validité de l'évaluation pour les élèves des groupes linguistiques différents.

Ce chapitre a décrit en détail les données PISA utilisées, les variables sélectionnées, le processus de recodage des items, ainsi que les procédures d'analyse. Il présente la méthodologie complète de notre étude, qui vise à analyser le fonctionnement différentiel des items (FDI) dans les épreuves de mathématiques de l'enquête PISA 2012 au Canada, en comparant les performances des élèves de langue première (L1) et de langue seconde (L2). La recherche utilise une approche de pairage par score de propension, des tests de proportion et une analyse de sensibilité pour détecter le FDI. Cette méthodologie rigoureuse est conçue pour isoler les effets spécifiques du FDI tout en contrôlant les variables confondantes potentielles, offrant ainsi une analyse approfondie des différences de performance liées au statut linguistique des élèves.

CHAPITRE 4

RÉSULTATS

Notre étude sur les biais linguistiques dans les évaluations mathématiques de PISA 2012 au Canada nécessite une analyse rigoureuse des données pour répondre aux questionnements soulevés concernant l'équité des évaluations. Ce chapitre présente les résultats des analyses menées pour examiner l'influence du statut linguistique des élèves sur leur performance aux items de mathématiques. Les analyses s'appuient sur la méthode de pairage par score de propension pour l'identification du fonctionnement différentiel d'items (FDI), tout en tenant compte des défis particuliers posés par la conception par livret de l'enquête PISA. Dans un premier temps, nous vérifions la qualité du pairage effectué, c'est-à-dire l'équilibre entre les groupes linguistiques (L1 et L2) afin d'assurer la validité des analyses comparatives ultérieures. Ensuite, nous présentons les résultats des analyses de FDI, en mettant en évidence les items qui manifestent un fonctionnement différentiel entre les apprenants L1 et L2.

4.1 Pairage et évaluation de la qualité

Le pairage des élèves par statut linguistique (L1 ou L2) a été effectué à l'aide de la version 4.5.5 de la librairie *MatchIt* pour le langage R (Ho et al., 2011). Conformément aux travaux d'Arikan (2018), la méthode « Optimal » a été utilisée avec une distance *glm* (scores de propension) pour apparier les covariables démographiques et socio-économiques (AGE, ESCS, FISCED, MISCED, ST04Q01, IMMIG_DICHO). Un pairage « exact » a été appliqué spécifiquement pour la variable BOOKID, garantissant ainsi que les élèves comparés ont répondu aux mêmes questions du test PISA.

Avant pairage, l'échantillon contenait 14634 élèves du Canada ayant effectué l'épreuve PISA 2012 en langue anglaise, soit 12876 anglophones (statut L1) et 3007 francophones ou allophones (statut L2). Le statut d'immigrant différait grandement entre les groupes : 10,0% des élèves L1 étaient issus de l'immigration (1283 sur 12876), alors que 50,1% des élèves L2 étaient issus de l'immigration (1508 sur 3007); cette différence était significative selon le test du chi-carré, $X2(1) = 3107$, $p < 0{,}001$. L'application de tests *t* a révélé une différence statistiquement significative pour la covariable ESCS (statut socio-économique) entre le groupe L1 (M = 0,475; ET = 0,826) et L2 (M = 0,475; ET = 0,878), $t(3745) = 1{,}63$, $p < 0{,}001$, $d = 0{,}224$. Une différence significative pour la covariable MISCED (niveau d'éducation de la mère) a

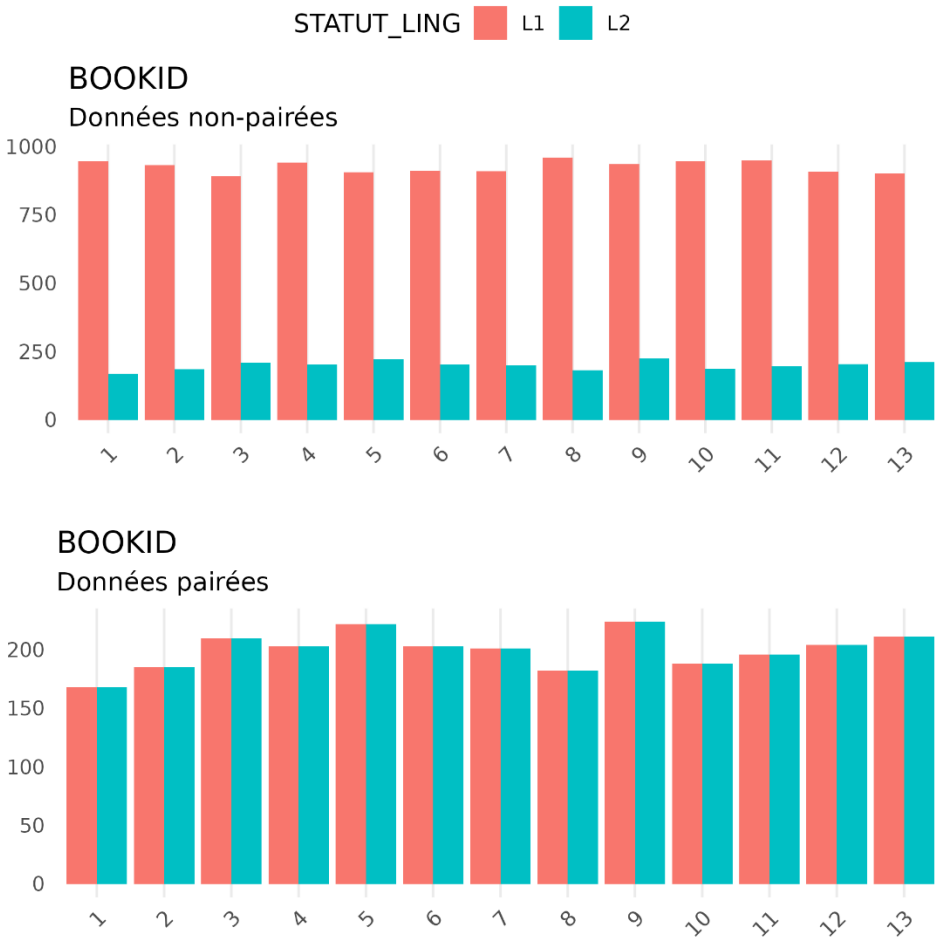
également été observée entre le groupe L1 (M = 4,93; ET = 1,07) et le groupe L2 (M = 4,72; ET = 1,38), $t(3303) = 7{,}30$, $p < 0{,}001$, $d = 0{,}170$. Les groupes non-pairés étaient équivalents en ce qui a trait à la scolarité du père (FISCED), $p = 0{,}649$, et avaient un âge équivalent au moment du test, $p = 0{,}104$.

Après application du pairage avec la méthode « optimal », l'échantillon apparié comptait 2597 élèves L1 appariés à 2597 élèves L2 en fonction des caractéristiques sociodémographiques retenues et du livret (*booklet*) de test qui leur a été administré. Le statut d'immigrant différait significativement entre les groupes L1 (48,7% d'élèves issus de l'immigration) et L2 (58,1%), $X2(1) = 44{,}9$, $p < 0{,}001$. Cette différence de proportion (9,4 points de pourcentage) était cependant réduite par comparaison aux données non-pairées (40,1 points de pourcentage), la Figure 4.1 illustre la réduction de l'écart de proportion. Considérant un seuil nominal de $p < 0{,}05$, des tests *t* n'ont révélé aucune différence significative entre les groupes pairés à l'égard du niveau socio-économique, de la scolarité de la mère ou du père, et de l'âge.

Figure 4.1 : *Variable IMMIG_DICHO (statut d'immigrant) avant et après pairage*

Enfin, la procédure de pairage ayant effectué un appariement exact sur la variable BOOKID (livret de test), les résultats illustrés à la Figure 4.2 montrent que, de manière prévisible, le nombre d'élèves L1 et L2 est équivalent pour chacun des livrets[8].

Figure 4.2 ***: Variable BOOKID avant et après pairage***



En plus de la comparaison des moyennes et proportions, nous avons effectué une évaluation plus qualitative de l'équivalence des distributions statistiques au moyen de la méthode de comparaison graphique, suivant les recommandations de Chen et Zumbo (2020). Nous n'avons pas relevé d'écart préoccupant entre les groupes L1 et L2 en comparant les distributions des variables continues ESCS,

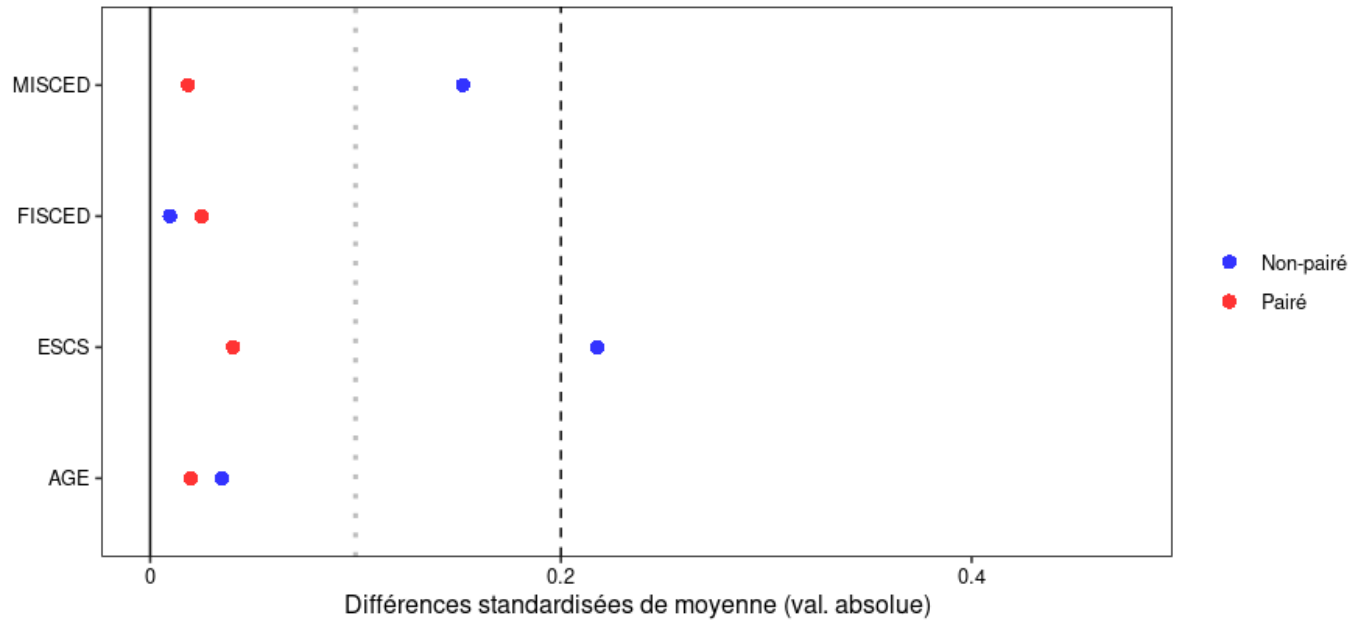
[8] La distribution de chacune des covariables selon le statut linguistique, avant et après pairage, est présentée à l'annexe A.

FISCED, MISCED et AGE par cette méthode. Nous avons de plus employé des graphes de Love pour comparer les moyennes et variances (Rubin, 2001; Stuart et Rubin, 2008).

4.1.1 Évaluation de la qualité du pairage

Afin d'évaluer la qualité du pairage, la Figure 4.3 illustre un graphe de Love comparant les écarts standardisés entre les moyennes des groupes L1 et L2, avant et après pairage. L'axe des X représente les différences standardisées absolues entre les groupes pour les variables continues, ou les différences de proportion pour les variables dichotomiques (statut d'immigration) ou catégorielles (BOOKID). Plus les points sont proches de la ligne verticale du zéro (la référence parfaite), meilleur est l'équilibre entre les groupes. Une différence standardisée absolue inférieure à 0,1 est généralement considérée comme indicatrice d'un bon équilibre (Austin, 2009; Granger et al., 2020; Zhang et al., 2019). Après pairage, la plupart des covariables présentent des différences standardisées de la moyenne inférieures à 0,1, ce qui reflète un bon équilibre entre les groupes[9].

Figure 4.3 : ***Graphe de Love – comparaison des différences de moyennes avant et après pairage***

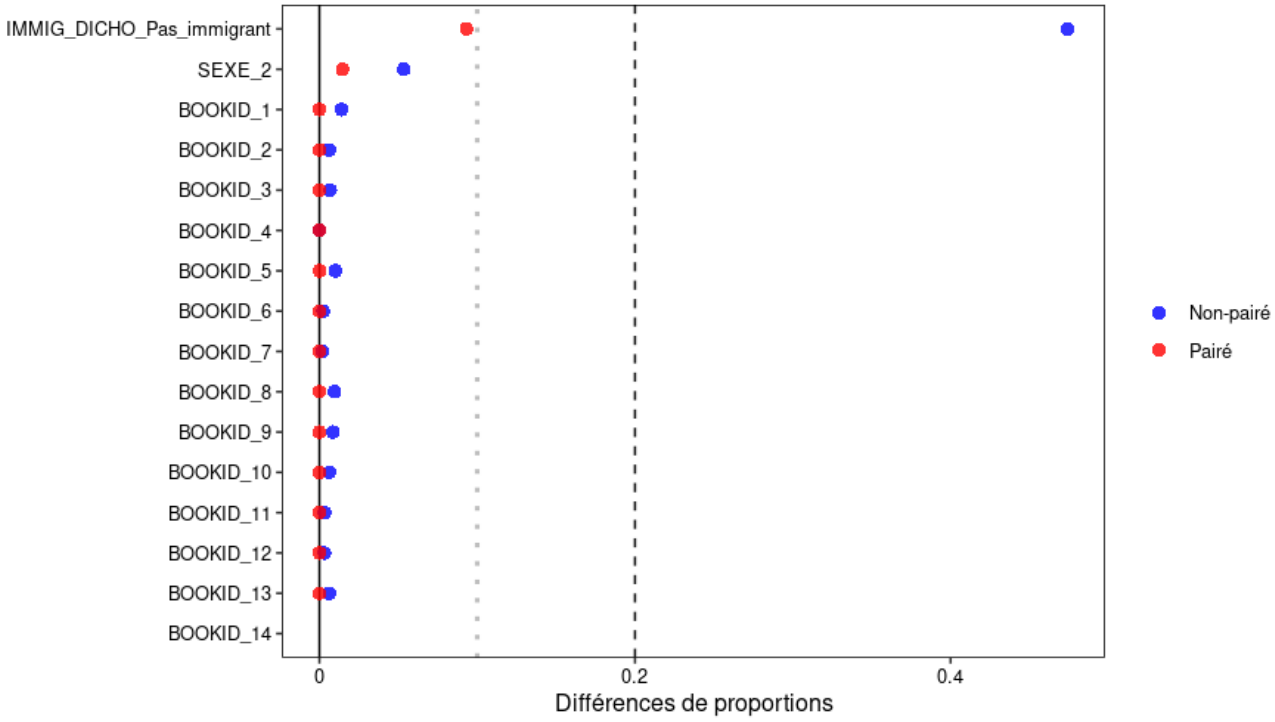


Note. Les lignes pointillées verticales indiquent les seuils de 0,1 (excellent équilibre) et de 0,2 (équilibre acceptable) proposés par Austin (2009). * Indique que la différence de proportion entre les groupes est montrée.

[9] Les différences standardisées des moyennes des covariables, avant et après pairage, sont présentées à l'annexe B.

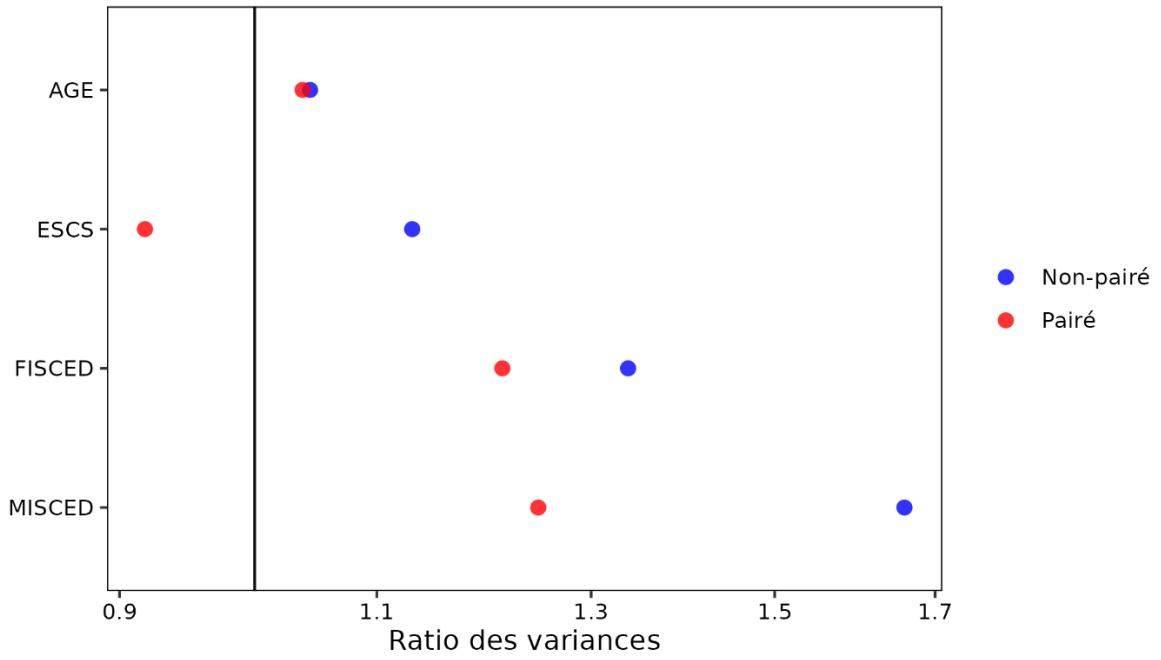
La Figure 4.4 montre la comparaison des proportions pour les variables dichotomiques (IMMIG_DICHO et SEXE) et catégorielles (BOOKID). Seule la covariable indiquant le statut d'immigrant (IMMIG_DICHO), présente encore un léger déséquilibre entre les groupes, avec une différence standardisée près du seuil de 0,1, mais cela demeure dans les limites acceptables.

Figure 4.4 : ***Graphe de Love – comparaison des proportions***



La Figure 4.5 montre le ratio des variances des groupes L1 et L2 avant et après pairage. Plus ce ratio est près de 1, plus il est possible d'affirmer que les groupes ont une variance équivalente. Les ratios de variance devraient être entre 0,5 et 2 pour chaque covariable (Zhang et al., 2019). On observe que le processus de pairage a eu pour effet de ramener les ratios des variances plus près de 1, et que les variances des groupes pairés L1 et L2 ont une équivalence acceptable. Cela reflète une grande consistance entre les groupes, ce qui renforce la fiabilité et la robustesse de l'analyse.

Figure 4.5 : ***Graphe de Love – ratios des variances avant et après pairage***



4.2 Tests de proportions simples

Afin de comparer les performances des groupes L1 et L2 sur les items de mathématiques, les tests de proportions simples ont été appliqués. Ils permettent d'évaluer si la différence observée entre les taux de réussite des deux groupes est statistiquement significative. L'analyse a été conduite en deux étapes : d'abord sur les données non pairées, puis sur les données pairées, afin de mettre en évidence l'apport du processus de pairage dans la neutralisation des biais liés aux covariables.

4.2.1 Tests de proportions sur données non-pairées

Le Tableau 4.1 présente les résultats obtenus à partir des tests de proportions sur les données non pairées. Seuls les items montrant une différence significative, avec une taille d'effet h ≥ 0,2 (soit au minimum un effet de petite ampleur selon Cohen, 1988), y sont rapportés. Les résultats des tests de proportions sur les données non-pairées sont présentés en entier à l'annexe C.

Tableau 4.1 ***: Résultats des tests de proportions sur données non-pairées (p < 0,05 et h ≥ 0,2)***

Item	n		Taux de réussite		*p*	*h*
	L1	L2	L1	L2		
PM571Q01	3677	804	0,53	0,40	< 0,001	0,26
PM909Q02	3737	810	0,71	0,59	< 0,001	0,25
PM034Q01T	3667	824	0,47	0,36	< 0,001	0,22
PM953Q02	3688	811	0,61	0,50	< 0,001	0,22
PM446Q01	3684	823	0,80	0,71	< 0,001	0,21
PM496Q02	3677	804	0,69	0,59	< 0,001	0,21
PM155Q04T	3667	824	0,65	0,55	< 0,001	0,20
PM408Q01T	3684	823	0,47	0,37	< 0,001	0,20
PM982Q04	3794	769	0,51	0,41	< 0,001	0,20

L'analyse du Tableau 4.1 met en évidence des différences notables de performance entre les groupes L1 et L2 sur plusieurs items de mathématiques, avec un avantage pour le groupe L1. Par exemple, l'item PM571Q01 présente un taux de réussite de 53 % pour les élèves L1 contre 40 % pour les élèves L2, avec une taille d'échantillon respectivement de 3 677 et 804 et une taille d'effet (*h*) de 0,26, ce qui suggère un avantage considérable pour les élèves de langue maternelle. De même, l'item PM909Q02 suit une tendance similaire, avec une proportion de 71 % de réussite chez les L1 contre 59 % chez les L2, une taille d'échantillon respectivement de 3 737 et 810, et une taille d'effet h = 0,25, illustrant une difficulté relative plus marquée pour les élèves de langue seconde.

Parmi les autres items, PM034Q01T et PM953Q02 affichent également des écarts marqués (*h* **≥** 0,22), confirmant une tendance où les élèves L2 ont une performance systématiquement inférieure à celle des élèves L1. L'item PM446Q01, bien qu'ayant des proportions de réussite plus élevées dans les deux groupes (80 % pour L1 et 71 % pour L2), présente tout de même un effet différentiel significatif, suggérant que même les items plus faciles ne garantissent pas une équité parfaite entre les groupes linguistiques.

La cohérence des résultats à travers les différents items, avec des échantillons relativement grands (allant de 3 794 à 3 667 pour L1 et de 824 à 769 pour L2), souligne que, lorsqu'on ne tient pas compte des variables

socioéconomiques, on observe une tendance claire où les élèves L1 ont de meilleures chances de réussir les items de mathématiques que les élèves L2. Ces résultats renforcent l'idée d'un impact linguistique sur la performance en mathématiques, les élèves de langue seconde étant désavantagés sur l'ensemble des items identifiés.

4.2.2. Tests de proportions sur données pairées

Le Tableau 4.2 présente les résultats obtenus à partir des tests de proportion sur les données pairées de manière à mitiger l'effet des variables socioéconomiques. L'analyse de ces tests confirme la persistance d'un écart de performance significatif entre les groupes L1 et L2, bien que légèrement réduit par rapport aux résultats obtenus sur les données non-pairées. Nous montrons au Tableau 4.2 uniquement les items ayant une différence statistiquement significative et une taille d'effet ($h \geq 0{,}2$) notable, suggérant un impact linguistique ayant des implications réelles sur la réussite des items. Les résultats des tests de proportions sur les données pairées sont présentés en entier à l'annexe D.

Tableau 4.2 : ***Résultats des tests de proportions sur données pairées (p < 0.05 et h ≥ 0.2)***

Item	n		Taux de réussite		*p*	*h*
	L1	L2	L1	L2		
PM571Q01	804	804	0,51	0,40	<0,001	0,22
PM155Q02D	711	706	0,80	0,71	<0,001	0,21
PM909Q02	810	810	0,69	0,59	<0,001	0,21
PM918Q01	784	784	0,90	0,83	<0,001	0,21
PM00FQ01	784	784	0,52	0,42	<0,001	0,20

L'examen détaillé des résultats met en lumière certains items pour lesquels l'écart entre les groupes demeure particulièrement marqué, malgré l'application du pairage. Par exemple, l'item PM571Q01 affiche une proportion de réussite de 51 % pour les élèves L1 contre 40 % pour les élèves L2, avec une taille d'échantillon de 804 pour chaque groupe et une taille d'effet *h* de 0,22, traduisant un désavantage important pour les élèves dont la langue du test est une langue seconde. De même, l'item PM909Q02, avec une réussite de 69 % pour L1 et 59 % pour L2, une taille d'échantillon de 810 pour chaque groupe et une taille d'effet (*h)* de 0,21, suit une tendance similaire, indiquant un écart persistant malgré le pairage. Certains items, bien que présentant des taux de réussite plus élevés, continuent de montrer des différences

significatives. C'est le cas de PM918Q01, où 90 % des élèves L1 réussissent contre 83 % des élèves L2 (h = 0,21, pour un échantillon de 784 dans chaque groupe), et de PM155Q02D (80 % contre 71 %, h = 0,21, pour un échantillon de 711 en L1 et 706 en L2). L'item PM00FQ01, bien qu'ayant des taux de réussite plus modérés (52 % pour L1 et 42 % pour L2, h = 0,20, avec un échantillon de 784 élèves dans chaque groupe), s'inscrit dans la même tendance de désavantage pour les élèves L2. Ainsi, même après pairage des groupes, l'effet différentiel en fonction de la langue demeure, renforçant l'hypothèse selon laquelle les élèves dont la langue du test est une langue seconde rencontrent des obstacles linguistiques affectant leur performance en mathématiques.

Il convient de préciser la stratégie retenue pour le traitement des comparaisons multiples. Aucune correction pour comparaisons multiples n'a été appliquée aux p-values dans cette étude, l'objectif étant avant tout d'identifier des tendances potentielles de fonctionnement différentiel des items (FDI). Cette décision repose sur le fait que l'interprétation ne s'appuie pas uniquement sur la significativité statistique, mais également sur la taille d'effet (h de Cohen), permettant ainsi de distinguer les différences statistiquement significatives, mais négligeables de celles ayant une réelle importance pratique. L'intégration conjointe de ces deux critères contribue à réduire le risque de faux positifs, inhérent à la multiplication des tests, tout en maintenant la sensibilité nécessaire pour détecter des écarts pertinents sur le plan éducatif.

4.3 Analyse de sensibilité

L'analyse de sensibilité vise à évaluer la solidité des différences observées entre les groupes L1 et L2 face à d'éventuels biais liés à des variables non mesurées. Nous avons employé deux indicateurs clés pour cette évaluation : la valeur p associée à un Γ (gamma) de 1, qui est l'estimation brute de la différence entre les groupes sans ajustement pour covariables non observées, et le plus grand Γ significatif, défini selon un seuil de $p < 0,05$, lequel estime l'influence qu'une variable non mesurée devrait exercer pour rendre cette différence non significative (Rosenbaum, 2005). Une valeur plus élevée de Γ suggère donc une robustesse plus forte de l'effet observé, le rendant moins susceptible d'être dû à un biais non contrôlé (Rosenbaum, 2007).

Le Tableau 4.3 montre le résultat des analyses de sensibilité pour les items qui semblent afficher un FDI selon cette méthode ($p < 0,05$ lorsque Γ = 1, et présence d'un Γ significatif supérieur à 1). Les valeurs Γ sont un rapport de cote et s'interprètent comme suit : le groupe L1 a Γ fois plus de chances de réussir l'item que

le groupe L2. Par exemple, à l'item PM571Q01 le groupe L1 a un taux de réussite de 51%, pour un rapport de cotes de 1,041, et le groupe L2 a un taux de réussite de 40%, ce qui correspond à des rapports de cotes de 0,667. Le plus grand Γ significatif est de 1,3, ce qui signifie qu'une covariable non observée devrait diminuer le rapport des cotes du groupe L1, ou augmenter le rapport des cotes du groupe L1 par un facteur de 1,3 pour rendre non-significative la différence entre les groupes. Effectivement, pour le groupe L1 cette déviation entrainerait un nouveau rapport de cotes de 1,014 / 1,3 = 0,801, l'équivalent d'un taux de réussite de 44,5%, ce qui est environ 6,5 points de pourcentage sous le taux observé de 51%. Pour le groupe L2, cette déviation signifierait une augmentation du rapport de cotes à 0,667 x 1,3 = 0,867, ou un taux de réussite de 46,4%, soit environ 6,4 points de pourcentage au-dessus du taux observé de 40%. Ces marges indiquent dans quelle mesure une covariable non observée devrait faire dévier les scores afin que la différence observée pour l'item PM571Q01 soit jugée non-significative.

Tableau 4.3 : *Résultats de l'analyse de sensibilité (max Γ > 1,0)*

Item	*p* | Γ = 1	max Γ | *p* < 0,05
PM918Q01	< 0,001	1,4
PM571Q01	< 0,001	1,3
PM00FQ01	< 0,001	1,2
PM155Q02D	< 0,001	1,2
PM909Q01	0,00266	1,2
PM909Q02	< 0,001	1,2
PM034Q01T	< 0,001	1,1
PM408Q01T	0,00314	1,1
PM446Q01	0,00312	1,1
PM496Q02	0,00368	1,1
PM905Q01T	0,00606	1,1
PM906Q01	< 0,001	1,1
PM906Q02	0,00154	1,1
PM915Q02	< 0,001	1,1
PM919Q01	0,00427	1,1
PM919Q02	0,00477	1,1
PM924Q02	< 0,001	1,1
PM953Q02	0,00284	1,1
PM953Q04D	0,00317	1,1
PM954Q01	0,00137	1,1
PM982Q01	0,01001	1,1
PM982Q04	0,00170	1,1
PM998Q02	0,00328	1,1

Note. Valeur *p* lorsque gamma est maintenu à 0, et plus grande valeur gamma significative ($p < 0,05$). Le tableau affiche seulement les items pour lesquels le plus grand gamma significatif était supérieur à 1,0.

Les résultats montrent les items du Tableau 4.3 présentant des différences significatives ($p < 0,05$) entre L1 et L2, ce qui suggère la présence d'un FDI[10]. Cependant, la robustesse de ces différences varie en fonction des items, comme l'illustre la valeur de gamma significative maximale (max Γ | $p < 0,05$). L'item PM918Q01 se distingue par une différence montrant la plus grande robustesse avec Γ = 1,4, ce qui signifie qu'une variable non mesurée devrait avoir une influence suffisamment forte pour diviser par 1,4 le rapport des cotes du groupe L1, ou multiplier par 1,4 le rapport des cotes du groupe L2, afin d'annuler la significativité de la différence observée. Cet item est donc particulièrement résistant aux biais potentiels. À un niveau légèrement inférieur, PM571Q01 affiche un Γ = 1,3, indiquant une robustesse élevée : bien que moins stable que PM918Q01, il demeure insensible aux biais non mesurés. Certains items, comme PM00FQ01, PM155Q02D, PM909Q01 et PM909Q02, présentent un Γ = 1,2, suggérant une stabilité modérée. Si leurs différences entre L1 et L2 restent statistiquement significatives, elles sont toutefois plus sensibles aux influences de variables non contrôlées. En particulier, PM909Q01, avec une valeur p de 0,00266, légèrement plus élevée que les autres, indique, bien que toujours significative, une significativité plus faible par rapport aux autres items de cette catégorie. Enfin, les items dont Γ ≤ 1,1, tels que PM034Q01T, PM408Q01T, PM446Q01, etc. sont ceux dont les résultats semblent le moins robustes. Bien que la différence entre les groupes L1 et L2 demeure significative ($p < 0,05$) lorsque gamma est de 1, ces items sont sensibles aux influences de variables non mesurées. Cela signifie que l'effet du FDI observé pour ces items pourrait être plus facilement réduit par des facteurs externes, rendant leur interprétation plus délicate en termes de stabilité des différences entre les groupes.

Afin de mieux visualiser l'ensemble des résultats issus des analyses statistiques, le Tableau 4.4 croise les résultats des tests de proportions non-pairées et pairées, ainsi que les valeurs maximales de sensibilité (Γ), ce qui permet d'évaluer la stabilité des effets observés. Il fournit également des informations complémentaires sur les caractéristiques des items (contenu, type, processus mathématique et contexte), facilitant ainsi une interprétation hypothétique des écarts de performance entre les groupes linguistiques.

[10] Les résultats pour l'ensemble des items sont présentés à l'annexe E.

Tableau 4.4 : *Sommaire des résultats*

Identifiant de l'item	Analyse non-pairée	Analyse pairée	Analyse de sensibilité (max Γ)	Catégorie de contenu	Type d'item	Processus mathématique	Contexte
PM918Q01	Non	Oui	1,4	Données et incertitudes	Choix multiple simple	Interpréter	Sociétal
PM571Q01	Oui	Oui	1,3	Changement et relations	Choix multiple simple	Interpréter	Scientifique
PM00FQ01	Non	Oui	1,2	Expert espace et forme	Réponse construite	Formuler	Personnel
PM155Q02D	Non	Oui	1,2	Changement et relations	Choix multiple simple	Employer	Scientifique
PM909Q02	Oui	Oui	1,2	Quantité	Choix multiple simple	Employer	Sociétal
PM034Q01T	Oui	Non	1,1	Espace et forme	Réponse construite Auto-codée	Formuler	Professionnel
PM408Q01T	Oui	Non	1,1	Données et incertitudes	Choix multiple complexe	Interpréter	Sociétal
PM446Q01	Oui	Non	1,1	Changement et relations	Manuel de réponse construite	Formuler	Scientifique
PM496Q02	Oui	Non	1,1	Quantité	Manuel de réponse construite	Employer	Sociétal
PM953Q02	Oui	Non	1,1	Expert données et incertitudes	Réponse construite	Interpréter	Scientifique
PM982Q04	Oui	Non	1,1	Données et incertitudes	Choix multiple simple	Formuler	Sociétal

Parmi les onze items analysés, cinq se distinguent par un FDI détecté après pairage et soutenu par une valeur de sensibilité élevée ($\Gamma \geq 1{,}2$), ce qui renforce leur caractère potentiellement inéquitable. Il s'agit des items PM918Q01 ($\Gamma = 1{,}4$), PM571Q01 ($\Gamma = 1{,}3$), ainsi que PM00FQ01, PM155Q02D et PM909Q02 ($\Gamma = 1{,}2$ chacun).

L’item PM918Q01, un choix multiple simple portant sur les données et incertitudes dans un contexte sociétal, mobilise le processus « Interpréter » ; ce type de tâche peut désavantager les élèves de langue seconde (L2) en raison de la complexité textuelle qu’elle implique. De même, PM571Q01, également un item d’interprétation, mais en contexte scientifique, présente une surcharge langagière possible liée à l’usage d’un vocabulaire spécialisé. L’item PM00FQ01, de type réponse construite, exige la formulation d’une solution géométrique en contexte personnel, un format particulièrement exigeant pour les élèves L2 puisqu’il demande non seulement des compétences en mathématiques, mais aussi une capacité à exprimer clairement une démarche. Enfin, les items PM155Q02D et PM909Q02, tous deux des choix multiples simples mobilisant le processus « Employer » en contextes scientifique et sociétal respectivement, présentent également un FDI post-pairage avec une robustesse modérée ($\Gamma = 1,2$). Bien que leur format semble moins exigeant sur le plan rédactionnel, il est possible que le vocabulaire ou les formulations spécifiques des énoncés contribuent à accentuer les écarts de performance entre les groupes linguistiques. Ces résultats soulignent l’influence non négligeable de la langue d’évaluation sur la performance des élèves et confirment la nécessité d’intégrer une analyse linguistique dans les processus de développement et de validation des tests standardisés, afin d'en garantir l'équité.

CHAPITRE 5

DISCUSSION

Cette recherche s'intéresse à l'influence des facteurs linguistiques sur l'évaluation des compétences en mathématiques des élèves dans PISA 2012, en ciblant les adolescents canadiens âgés de 15 ans. L'objectif principal était de déterminer dans quelle mesure la langue d'apprentissage, lorsqu'elle diffère de la langue maternelle, affecte la performance aux tests standardisés, soulevant des enjeux d'équité puisque ces tests reposent sur des normes linguistiques pouvant désavantager les élèves L2. L'étude visait à identifier et analyser d'éventuels effets de FDI liés à la langue afin d'établir si les résultats reflètent des obstacles linguistiques plutôt que des différences de compétence. Un enjeu méthodologique majeur concernait la gestion des données manquantes générées par la structure en livrets de PISA qui, bien qu'elle permette d'administrer un grand nombre d'items, introduit des données partiellement observées pouvant affecter l'analyse du FDI et la comparabilité des résultats. Pour y remédier, des techniques de pairage avancées ont été utilisées afin de comparer des élèves aux profils similaires, différant uniquement par leur statut linguistique, ce qui a permis une analyse rigoureuse des effets de la langue tout en minimisant l'influence d'autres variables confondantes et en renforçant la validité des conclusions.

Les items de PISA ont été recodés en format dichotomique (0 = échec, 1 = succès), y compris pour les items partiellement réussis ou non complétés, simplifiant l'analyse du FDI et permettant une comparaison directe entre groupes linguistiques (Akcan et Kabasakal, 2023). Le pairage , réalisé avec la librairie MatchIt pour R (Stuart et al., 2011), a neutralisé l'effet de variables confondantes et montré un bon équilibre (Zhang et al., 2019; Rubin, 2001), puis les tests de proportion et la taille d'effet h de Cohen, complétés par une analyse de sensibilité selon Rosenbaum (2005), ont confirmé que les différences observées reflètent véritablement un FDI plutôt que des disparités globales d'habileté ou de contexte socio-économique, tout en minimisant l'impact des données manquantes. Par la suite, un aperçu des résultats est présenté en lien avec les objectifs de la recherche, afin d'examiner comment les facteurs linguistiques se traduisent dans la performance aux tests de mathématiques.

5.1 Aperçu des résultats en lien avec les objectifs de la recherche

Les résultats de cette étude confirment l'existence systématique d'un FDI entre les groupes linguistiques L1 et L2 dans l'évaluation PISA 2012 en mathématiques. Tous les items analysés présentent des différences statistiquement significatives ($p < 0,05$), avec des degrés variables de robustesse face aux FDI non mesurés. En réponse à la question principale de recherche concernant l'influence des facteurs linguistiques sur l'évaluation des habiletés en mathématiques, l'étude démontre clairement que la langue du test a un impact mesurable sur la performance des élèves. Cet impact affaiblit la validité des scores et soulève des questions d'équité, particulièrement pour les élèves dont la langue d'apprentissage diffère de leur langue maternelle.

Les tendances générales observées révèlent que l'influence de la langue n'est pas uniforme à travers tous les items, comme en témoigne la variabilité des valeurs de robustesse (Γ) allant de 1,1 à 1,4. Certains items (ex: PM918Q01 et PM571Q01) montrent un écart plus robuste (Γ = 1,4 et 1,3), indiquant que les différences entre groupes L1 et L2 demeurent stables même en présence de variables non mesurées. D'autres items présentent une robustesse plus modérée (Γ = 1,2) ou faible (Γ ≤ 1,1), suggérant que les différences observées pourraient être influencées par des facteurs externes non pris en compte. Ces résultats confirment l'hypothèse initiale selon laquelle les tests standardisés peuvent défavoriser certains groupes d'élèves non pas en raison de leurs habiletés mathématiques, mais à cause de leur maîtrise de la langue d'apprentissage. Par ailleurs, la gestion des données manquantes liées à la méthode de livret (*booklet design*) de PISA constitue un enjeu méthodologique central. Nos résultats montrent que la gestion des données manquantes liées au livret de PISA a été efficacement prise en compte grâce au pairage, qui a permis d'isoler l'effet de la langue, de limiter l'influence des variables confondantes et de renforcer la validité des analyses.

5.2 Analyse approfondie des résultats et mise en relation avec la littérature

Dans cette partie nous abordons les résultats clés et nous les mettons en relation avec les concepts théoriques et les études du cadre théorique.

5.2.1 Présence et évolution du FDI après pairage

L'analyse de la présence et de l'évolution du FDI constitue une étape clé pour comprendre l'influence réelle des facteurs linguistiques. Cette section examine d'abord les résultats obtenus avant et après pairage, puis

met en lumière l'effet du pairage sur la détection du FDI. Enfin, une comparaison est proposée avec d'autres études ayant mobilisé des approches similaires afin de mieux situer les résultats de cette recherche dans la littérature existante.

5.2.1.1 Explication des résultats obtenus avant et après pairage

Nos résultats concernant l'échantillon avant pairage révèlent des différences significatives entre les groupes linguistiques qui justifient l'application d'une procédure d'appariement. L'échantillon initial comprenait 14634 élèves canadiens ayant participé à l'épreuve PISA 2012, comprenant 12876 anglophones (statut L1) et 3007 francophones ou allophones (statut L2). Une disparité importante a été observée concernant le statut d'immigrant : seulement 10,0% des élèves L1 étaient issus de l'immigration contre 50,1% des élèves L2, une différence statistiquement significative. De plus, des tests t ont révélé des différences significatives entre les groupes concernant le statut socio-économique (ESCS) et le niveau d'éducation de la mère (MISCED), avec des tailles d'effet respectives de $d = 0,224$ et $d = 0,170$. Ces disparités initiales sont problématiques pour l'analyse du FDI, car elles pourraient constituer des variables confondantes susceptibles de fausser l'interprétation des résultats, conformément aux préoccupations soulevées par Chen et al. (2020).

L'application de la méthode d'appariement optimal a considérablement amélioré l'équivalence des groupes. L'échantillon apparié comprend maintenant 2597 élèves dans chaque groupe linguistique, appariés selon les caractéristiques sociodémographiques et le livret de test. Bien qu'une différence significative persiste au niveau du statut d'immigrant (48,7% pour L1 contre 58,1% pour L2), l'écart a été largement réduit, passant de 40,1 à 9,4 points de pourcentage. Les tests t n'ont révélé aucune différence significative entre les groupes appariés concernant le niveau socio-économique, la scolarité des parents et l'âge, ce qui montre l'efficacité du pairage. Cette amélioration de l'équilibre est importante, car elle permet, comme le soulignent Arikan et al. (2018), d'isoler plus précisément les effets potentiels du FDI en contrôlant simultanément plusieurs caractéristiques de base. Notre évaluation qualitative utilisant des comparaisons graphiques, conformément aux recommandations de Chen et al. (2020), n'a pas révélé d'écart préoccupant entre les distributions des variables continues. Les graphes de Love montrent que la plupart des covariables présentent des différences standardisées inférieures à 0,1 après pairage, seuil généralement considéré comme indicateur d'un bon équilibre selon plusieurs chercheurs (Austin, 2009; Granger et al., 2020; Zhang et al., 2019). De même, les ratios de variance entre les groupes se sont rapprochés de 1 après pairage, confirmant une équivalence acceptable et renforçant la fiabilité de notre

analyse. Ces constats nous permettent de conclure que le pairage a atteint son objectif principal : rendre les groupes plus comparables et réduire l'influence de variables confondantes. Même si certaines différences, notamment liées au statut d'immigrant, persistent, elles sont suffisamment atténuées pour permettre une interprétation plus fiable des résultats. Cela confirme l'importance de recourir à une telle méthode dans le cadre d'analyses de FDI, car sans ce contrôle rigoureux, les conclusions auraient pu refléter davantage les disparités sociodémographiques que de véritables effets linguistiques.

5.2.1.2 Effet du pairage sur la détection du FDI

L'impact du pairage sur la détection du FDI dans notre étude est décisif et s'aligne avec les observations rapportées dans la littérature scientifique. En équilibrant efficacement les covariables entre nos groupes linguistiques, nous avons pu éliminer ou réduire considérablement l'influence de facteurs confondants potentiels qui auraient pu être incorrectement attribués au FDI. Cette approche méthodologique rigoureuse rejoint les recommandations de Zumbo (2007), qui souligne l'importance de comprendre si les différences observées dans les résultats des tests entre les groupes sont dues à de véritables différences d'aptitudes ou à des FDI sous-jacents dans les items. En réduisant l'écart entre les groupes concernant le statut d'immigrant, le statut socio-économique et le niveau d'éducation parental, nous avons créé des conditions plus favorables pour détecter avec précision les items présentant un fonctionnement différentiel authentique.

La méthode d'appariement optimal que nous avons employée s'est avérée particulièrement efficace pour maintenir une taille d'échantillon adéquate sans perte significative de représentativité, comme le suggèrent Arikan et al. (2018). Après le pairage, chaque groupe (L1 et L2) comptait 784 participants pour la majorité des analyses, ce qui reflète un échantillon équilibré et suffisamment large pour garantir la robustesse des comparaisons. Cette caractéristique est essentielle, car elle préserve la puissance statistique nécessaire pour détecter le FDI tout en assurant l'équilibre des covariables. Notre analyse rejoint également les conclusions de Wang et al. (2013) concernant l'importance d'un équilibre adéquat entre les groupes pour minimiser les biais dans la détection du FDI. Comme le montrent nos résultats avec des différences standardisées majoritairement inférieures à 0,1 après pairage, nous avons atteint un niveau d'équilibre considéré comme optimal dans la littérature. Cette réduction des différences entre les groupes nous a permis d'effectuer une analyse du FDI plus précise, en diminuant le risque d'identifier faussement des items comme présentant un FDI alors qu'ils reflètent en réalité des différences sociodémographiques préexistantes. Notre approche s'inscrit ainsi dans ce que Chen et al. (2020) décrivent comme une avancée

vers des affirmations causales plus crédibles concernant les effets du FDI, en éliminant les différences préexistantes entre les groupes avant l'analyse.

5.2.1.3 Comparaison avec les études ayant utilisé des approches similaires pour contrôler des variables confondantes

Notre méthodologie d'appariement présente des similitudes et des différences notables avec d'autres études ayant utilisé des approches comparables pour contrôler les variables confondantes dans l'analyse du FDI. Tout comme Chen et al. (2020), nous avons utilisé des méthodes d'appariement optimal pour garantir que nos groupes soient équilibrés en termes de covariables. Cette approche est particulièrement appréciée dans la littérature pour son efficacité à créer des groupes comparables. Notre évaluation de la qualité du pairage, incluant des inspections visuelles et le calcul des différences standardisées, s'aligne également avec leur méthodologie. Toutefois, contrairement à leur étude qui a révélé que l'appariement complet optimal produisait un meilleur équilibre des covariables que l'appariement par paires optimal, notre recherche s'est concentrée principalement sur l'appariement optimal sans comparer systématiquement différentes méthodes d'appariement.

En comparaison avec l'étude d'Arikan et al. (2018), qui a exploré diverses techniques d'appariement pour l'analyse FDI dans le contexte des évaluations PISA, notre recherche présente des parallèles intéressants. Ces chercheurs ont constaté que l'appariement optimal fournissait des résultats cohérents pour diverses mesures et était efficace pour produire des groupes équilibrés tout en maintenant la taille de l'échantillon, observations que nos résultats confirment. Cependant, contrairement à leur étude qui a comparé explicitement les méthodes d'appariement exact, du plus proche voisin et optimal, nous nous sommes concentrés sur l'application et l'évaluation d'une seule méthode. Ils ont également observé que l'efficacité des techniques d'appariement est déterminée par leur capacité à équilibrer les variables d'arrière-plan, ce que notre recherche corrobore au vu de l'amélioration significative de l'équilibre de nos covariables après pairage.

Notre approche s'inscrit également dans la lignée des recommandations de Wang et al. (2013) concernant l'importance de l'équilibre des covariables pour minimiser les biais dans l'analyse. Leur étude a indiqué qu'une différence standardisée inférieure à 0,1 indique un déséquilibre négligeable des covariables entre les groupes, seuil que nous avons adopté et largement atteint pour la plupart de nos variables. Cependant, alors que Wang et ses collègues ont exploré différentes largeurs de *caliper* dans l'appariement par score

de propension pour déterminer l'approche optimale, notre étude n'a pas varié ce paramètre méthodologique. Leur recherche a souligné l'importance de sélectionner des largeurs de *caliper* appropriées pour minimiser les biais tout en maximisant le nombre de sujets appariés, aspect que nous pourrions explorer dans des recherches futures pour raffiner davantage notre méthodologie d'appariement.

5.2.2 Impact de la langue (L1/L2) sur la performance en mathématiques

Cette section examine comment la langue d'apprentissage (L1 vs L2) influence la performance en mathématiques chez les adolescents canadiens de 15 ans dans PISA 2012. Elle présente d'abord les différences de performance observées entre les groupes linguistiques, puis explore les effets de FDI liés à la langue du test, en les comparant aux résultats et conclusions de la littérature existante sur l'impact du facteur linguistique en évaluation.

5.2.2.1 Analyse des différences de performance entre les groupes linguistiques

Notre étude a mis en évidence des différences statistiquement significatives ($p < 0,05$) entre les performances des élèves L1 et L2 dans les évaluations mathématiques. Ces résultats confirment l'existence d'un FDI systématique qui mérite une attention particulière. L'item PM918Q01 présente la plus forte robustesse avec un coefficient gamma significatif jusqu'à $\Gamma = 1,4$, indiquant qu'une variable non mesurée devrait exercer une influence plutôt considérable pour neutraliser cette différence. Cette robustesse élevée suggère que ces écarts de performance reflètent des disparités structurelles entre les groupes linguistiques plutôt que des variations aléatoires.

Nos analyses montrent une gradation dans la robustesse des différences observées. Des items comme PM571Q01 ($\Gamma = 1,3$) démontrent une résistance élevée aux facteurs externes, tandis que d'autres comme PM00FQ01, PM155Q02D, et PM909Q02 ($\Gamma = 1,2$) présentent une stabilité modérée. Cette variabilité correspond aux observations de Buono et Jang (2021) concernant l'impact différencié de la complexité linguistique sur les performances des élèves. Ces différences peuvent être liées à certaines particularités des items, telles que la complexité du vocabulaire, la longueur des énoncés ou la structure des questions, qui peuvent rendre la compréhension plus difficile pour les élèves L2. Autrement dit, un élève L2 pourrait se tromper sur un item à cause de la langue ou de la formulation de la question, et non parce qu'il ne maîtrise pas les compétences mathématiques évaluées. Cette variabilité indique que l'impact du facteur linguistique n'est pas uniforme et dépend des caractéristiques spécifiques de chaque item, comme la

présence de termes techniques ou de formulations syntaxiques plus complexes. Nos résultats corroborent également ceux de Liu et Bradley (2021), qui ont identifié l'un des mêmes items (PM909Q02) comme particulièrement difficile pour les apprenants de l'anglais dans les évaluations PISA 2012.

5.2.2.2 Explication des FDI possibles liés à la langue du test et comparaison avec la littérature existante sur l'impact du facteur linguistique en évaluation

La présence de FDI liée à la langue du test peut être expliquée par plusieurs facteurs linguistiques spécifiques qui créent des obstacles à la performance des élèves L2. Bien que le FDI lié à la langue du test constitue un phénomène complexe dont l'explication requiert une analyse approfondie des mécanismes sous-jacents, plusieurs facteurs linguistiques interconnectés contribuent à créer des obstacles systématiques pour les élèves apprenant dans une langue seconde (L2), entraînant des écarts de performance significatifs lors des évaluations mathématiques.

La complexité des énoncés mathématiques représente un premier obstacle majeur pour les apprenants L2. Comme l'ont montré Buono et Jang (2021), les caractéristiques linguistiques des items peuvent considérablement influencer leur accessibilité. Leur analyse méthodique a révélé que parmi 28 items mathématiques étudiés, 11 présentaient une complexité linguistique élevée. À noter, sept de ces items à forte charge linguistique manifestaient un FDI statistiquement significatif au détriment des apprenants d'anglais langue seconde. Cette relation n'est pas accidentelle, mais reflète plutôt comment l'interaction entre formulations linguistiques complexes et exigences cognitives élevées amplifie les défis rencontrés par ces élèves. Les caractéristiques particulièrement problématiques incluent les présentations abstraites, l'utilisation d'un vocabulaire peu fréquent, les constructions nominales étendues, les structures interrogatives complexes et l'emploi de la voix passive.

De plus, comme l'ont établi Lee et Randall (2011) dans leurs travaux fondamentaux, la complexité linguistique inhérente aux tests mathématiques crée une double exigence cognitive pour les élèves L2, qui doivent simultanément traiter le contenu mathématique et surmonter les barrières linguistiques. Notre étude corrobore cette hypothèse en révélant des niveaux variables de robustesse du FDI, avec des coefficients gamma (Γ) variant entre 1,1 et 1,4 selon les items. Cette variabilité suggère que certaines formulations comportent des caractéristiques linguistiques particulièrement désavantageuses pour les apprenants L2, créant ainsi une barrière cognitive supplémentaire qui interfère avec leur capacité à démontrer pleinement leurs habiletés mathématiques.

L'exposition différentielle aux programmes de mathématiques constitue un deuxième facteur déterminant dans l'émergence du FDI. Lee et Randall (2011) ont observé que les élèves en difficulté linguistique sont souvent placés dans des classes de langue seconde qui limitent leur accès au contenu mathématique standard. Ces dispositifs, bien qu'ils visent à renforcer les compétences linguistiques, réduisent souvent le temps consacré aux mathématiques au profit de l'enseignement linguistique. Ce déséquilibre peut créer un décalage progressif dans l'acquisition des connaissances et compétences mathématiques. De plus, le passage entre classes de langue seconde et classes régulières peut engendrer des lacunes d'apprentissage, en raison des différences de rythme et de contenu entre ces programmes.

Les facteurs institutionnels et les politiques éducatives pourraient constituer un troisième déterminant essentiel des écarts de performance, comme l'ont démontré Liu et Bradley (2021). Leur recherche révèle que le type d'établissement scolaire exerce une influence significative sur certains items présentant un FDI, soulignant ainsi l'importance du contexte éducatif dans la formation des compétences mathématiques. Cette variabilité interinstitutionnelle peut s'expliquer par plusieurs aspects : les approches pédagogiques distinctives adoptées par chaque établissement, les variations dans la formation et l'expérience professionnelle des enseignants, l'organisation spécifique des programmes et la répartition du temps d'enseignement, ainsi que les politiques linguistiques mises en œuvre au niveau institutionnel.

Ces différents facteurs potentiels interagissent de manière complexe, créant des conditions systématiquement défavorables pour les apprenants L2 lors des évaluations mathématiques. La persistance du FDI, même après avoir contrôlé diverses variables contextuelles, illustre la nature fondamentale de ces obstacles linguistiques et souligne la nécessité d'adopter des approches plus équitables en matière d'évaluation éducative.

Face à ces défis, plusieurs solutions ont été proposées pour réduire les biais linguistiques dans les évaluations. L'une d'entre elles est l'utilisation de questions « décomposées », qui consistent à diviser une question complexe en plusieurs questions simples pour aider les élèves à mieux comprendre. En effet, Riccardi et al. (2020) ont montré que les apprenants L2 obtiennent de meilleurs résultats avec ce type de questions. Une autre approche bénéfique est la réduction de la complexité linguistique. À ce sujet, Loignon (2021a) a présenté l'outil ALSI (Analyseur Lexico-syntaxique Intégré) pour modéliser la complexité des textes en français. Il souligne que l'estimation de la complexité linguistique est essentielle pour l'élaboration d'évaluations. Cela permet de contrôler la variation indésirable due à la langue et de fournir

aux élèves des textes appropriés dans leurs évaluations. Des outils de ce type pourraient être utilisés pour valider les épreuves et les tests, afin de s'assurer que les évaluations mesurent réellement ce qu'elles sont censées mesurer. Des outils similaires existent pour l'anglais : par exemple, Coh-Metrix et MultiazterTest permettent d'analyser la lisibilité et la cohésion des textes, facilitant l'identification des passages complexes pour les apprenants L2 (Odo, 2018; Bengoetxea & Gonzalez-Dios, 2021). Ces outils peuvent être utilisés pour valider les épreuves et les tests, afin de s'assurer que les évaluations mesurent réellement ce qu'elles sont censées mesurer, en réduisant les biais liés à la langue.

5.3. Limites et avenues de recherches futures

5.3.1 Limites de l'étude

L'analyse repose sur les données de PISA 2012, qui représentaient la version la plus récente de PISA focalisée sur les mathématiques au moment de l'initiation de cette étude. Cependant, les données de PISA 2022, également centrées sur les mathématiques, sont maintenant disponibles et offriraient une perspective plus actuelle. De plus, notre étude s'appuie sur le test de proportion qui implique l'hypothèse d'indépendance locale des items, c'est-à-dire que les réponses à chaque question sont supposées indépendantes entre elles une fois que l'influence du trait mesuré est contrôlée. Cette hypothèse peut ne pas être entièrement respectée, ce qui limite la portée de nos conclusions. Par ailleurs, la dichotomisation des items (succès/échec) a simplifié l'analyse, mais s'est faite au prix d'une perte d'information. En transformant les réponses en variables binaires, certaines nuances liées à la difficulté graduelle des items ont pu être masquées, affectant potentiellement la détection du FDI.

Une des limites de notre étude réside dans le fait que la variabilité du niveau linguistique au sein du groupe L2 n'a pas été prise en compte. Ce groupe n'est pas homogène : il inclut des apprenants ayant des compétences linguistiques diverses, allant de débutants à quasi-bilingues. Cette hétérogénéité n'a pas été prise en compte dans l'analyse, ce qui pourrait influencer la précision des résultats et la détection du FDI. Les données PISA 2012 sont limitées à un échantillon particulier d'élèves et ne prennent pas en compte certaines variables importantes, telles que le niveau de bilinguisme, qui pourraient avoir un impact significatif sur les résultats.

En ce qui concerne le processus de pairage, celui-ci a permis d'améliorer l'équilibre entre les groupes. Toutefois, bien que l'intégration du statut d'immigration comme covariable ait contribué à cette

amélioration, elle n'a pas complètement éliminé les disparités. En effet, bien que l'écart initial de 40,1 points de pourcentage ait été réduit, une différence subsiste : 48,7 % des participants du groupe L1 sont immigrants, contre 58,1 % dans le groupe L2. Cette persistance des écarts souligne une limite aux méthodes d'appariement utilisées pour l'analyse du FDI. Bien que les méthodes de pairage permettent de créer des groupes plus comparables en contrôlant certaines variables, elles demeurent sensibles aux biais cachés. Si des covariables essentielles ne sont pas incluses ou ne peuvent être mesurées, l'équivalence entre les groupes peut être compromise, influençant ainsi les résultats obtenus. Ce risque, propre aux études observationnelles, peut être atténué par une sélection rigoureuse des covariables en s'appuyant sur la littérature et les modèles théoriques, mais il ne peut être totalement éliminé. Dans notre étude, les analyses corrélationnelles menées pour identifier les sources de FDI ne capturent pas l'ensemble des facteurs expliquant les écarts observés. Des éléments tels que le contexte scolaire, les stratégies de réponse ou encore l'influence culturelle pourraient jouer un rôle déterminant et mériteraient d'être examinés plus en profondeur. En particulier, l'exposition différentielle au contenu mathématique en fonction des parcours scolaires et linguistiques pourrait constituer un facteur clé influençant la performance des élèves. Cette observation met en évidence la nécessité de prendre en compte les politiques de placement et les pratiques pédagogiques dans les environnements multilingues afin de mieux comprendre et atténuer les inégalités observées.

Afin de compléter cette analyse, il est pertinent d'envisager des pistes d'amélioration méthodologique. Bien que cette étude ait principalement utilisé des tests de proportion pour identifier les différences de performance entre les groupes L1 et L2, il convient de noter que ces tests ne constituent pas la méthode la plus robuste pour analyser le FDI. Dans des recherches futures, des approches plus solides, telles que la régression logistique ou la procédure de Mantel-Haenszel, pourraient être employées afin de confirmer les résultats et de mieux contrôler l'influence des covariables observables et non observables. Cette perspective permettrait de renforcer la validité des conclusions et de compléter l'interprétation des différences de performance identifiées dans la présente étude. Notons finalement que nous n'avons pas comparé les résultats à ceux qui auraient pu émerger si d'autres méthodes plus courantes de détection du FDI avaient été appliquées aux données pairées. Notre approche ne se focalise pas principalement sur la méthode de détection du FDI en tant que telle, mais plutôt sur le pairage des données comme élément central de l'analyse. L'objectif de cette étude était d'examiner l'utilisation du pairage par score de propension comme méthode permettant d'identifier le fonctionnement différentiel

des items (FDI), tout en contrôlant l'influence de variables confondantes dans des données structurées selon un format en livrets (booklet).

5.3.2 Avenues de recherches futures

Des recherches futures pourraient s'intéresser à d'autres cohortes de PISA, notamment PISA 2022, qui se concentre également sur les mathématiques et représente les données les plus récentes. Nos analyses ont débuté avant la publication des données de PISA 2022. L'analyse des données TIMSS pourrait également permettre d'approfondir la compréhension des phénomènes de FDI dans un autre contexte évaluatif. De plus, l'étude actuelle a privilégié les tests de proportions en raison de leur simplicité. Toutefois, d'autres méthodes d'analyse du FDI, telles que le modèle de Rasch ou le modèle linéaire hiérarchique généralisé (GLMM), comme utilisés par Liu et Bradley (2021), pourraient offrir des perspectives différentes et complémentaires.

Une étude approfondie du contenu des items identifiés comme présentant un FDI pourrait éclairer les sources possibles de biais. Identifier les éléments linguistiques ou culturels complexes au sein des items mathématiques permettrait d'améliorer leur équilibre et leur validité pour les populations linguistiquement diverses. Les recherches futures pourraient approfondir l'analyse du contenu spécifique des items mathématiques afin d'identifier les éléments de complexité linguistique et les potentiels biais culturels qui affectent les performances des apprenants L2. Une analyse détaillée des caractéristiques linguistiques des items présentant un FDI robuste pourrait contribuer à la conception d'évaluations plus équitables. Par ailleurs, des études explorant l'évolution des habiletés mathématiques des apprenants L2 permettraient de mieux comprendre la dynamique temporelle des écarts de performance et l'efficacité des interventions pédagogiques.

Notre étude s'est concentrée sur les tests en anglais en raison de la disponibilité de données plus riches et de la possibilité de comparer avec des travaux antérieurs. Cependant, une étude parallèle sur un échantillon francophone permettrait d'examiner si les tendances observées sont similaires dans d'autres contextes linguistiques. En définitive, une démarche plus systématique faisant appel à des simulations de type Monte-Carlo pourrait nous aider à évaluer la fiabilité de la méthode préconisée dans cette étude, par exemple en déterminant si le pairage combiné à des simples tests de proportion parvient à identifier du FDI provoqué artificiellement.

Il convient également de souligner que le FDI peut avoir une dimension intersectionnelle. Par exemple, les effets liés au statut linguistique pourraient interagir avec d'autres variables comme l'âge ou le statut socio-économique, produisant ainsi des inégalités scolaires plus complexes. Bien que cette dimension n'ait pas été explorée dans la présente étude, elle constitue une piste pertinente pour de futures recherches visant à mieux comprendre la multiplicité des facteurs qui influencent l'équité des évaluations. Enfin, une piste de recherche concerne l'utilisation des valeurs plausibles. Dans les études PISA, les performances des élèves sont généralement analysées à partir de ces valeurs, qui fournissent des estimations plus robustes du niveau de compétence. Bien que la présente recherche ne les ait pas mobilisées, centrée sur l'identification du fonctionnement différentiel des items (FDI) selon le statut linguistique, leur intégration pourrait enrichir de futures investigations en reliant le FDI à des estimations plus globales de compétence.

5.4 Implications de la recherche

L'étude menée sur l'influence des caractéristiques linguistiques dans l'évaluation des habiletés en mathématiques des élèves issus de l'enquête PISA 2012 a plusieurs implications majeures sur les plans éducatif, méthodologique et politique. Ces implications découlent directement des constats présentés dans la problématique et mettent en lumière l'importance d'adopter des approches d'évaluation plus équitables et inclusives dans un contexte marqué par une diversité linguistique croissante au Canada.

5.4.1 Implications pour l'enseignement et l'apprentissage

L'un des principaux résultats attendus de cette étude est une meilleure compréhension des obstacles que rencontrent les élèves dont la langue d'apprentissage est une langue seconde (L2) lorsqu'ils passent des évaluations de mathématiques. La présence de FDI liée aux caractéristiques linguistiques dans ces évaluations signifie que certains élèves peuvent être désavantagés, non pas en raison de leurs habiletés en mathématiques, mais de leur niveau de maîtrise de la langue d'apprentissage.

Cette constatation a des implications directes pour les enseignants et les pédagogues. Une sensibilisation plus élevée aux défis linguistiques dans l'apprentissage des mathématiques peut amener les enseignants à adapter leurs pratiques pédagogiques. Par exemple, ils pourraient simplifier le langage utilisé dans les énoncés de problèmes, fournir un soutien linguistique aux élèves L2 et favoriser des approches pédagogiques intégrant explicitement la langue et les concepts mathématiques. En rendant

l'enseignement des mathématiques plus accessible à tous les élèves, ces adaptations contribueraient à réduire les écarts de performance liés à la langue et à promouvoir une éducation plus équitable.

5.4.2 Implications pour l'élaboration des tests et des politiques d'évaluation

Les résultats de cette recherche auront également un impact sur la conception des tests standardisés et les politiques d'évaluation. L'identification d'items présentant un FDI dû à des facteurs linguistiques met en évidence la nécessité de développer des outils d'évaluation plus justes, qui permettent à tous les élèves de démontrer leurs habiletés réelles en mathématiques, indépendamment de leur niveau de maîtrise de la langue d'apprentissage. Les concepteurs d'examens pourraient ainsi être encouragés à réviser la formulation des items afin de limiter les obstacles linguistiques. Cela pourrait inclure l'utilisation d'un vocabulaire plus accessible, la simplification des structures syntaxiques complexes et l'ajout de clarifications contextuelles pour éviter toute ambiguïté. Par ailleurs, des politiques pourraient être mises en place pour tester systématiquement les évaluations standardisées à l'aide de méthodes d'analyse du FDI, afin de s'assurer que les tests ne désavantagent pas certains groupes d'élèves de manière injustifiée.

5.4.3 Implications méthodologiques et scientifiques

Sur le plan méthodologique, cette étude contribue à l'amélioration des techniques d'analyse du FDI, notamment en contexte d'enquêtes à grande échelle comme PISA. L'un des défis soulevés dans la problématique est la gestion des données manquantes générées par la méthode de livret (*booklet design).* L'approche adoptée dans cette recherche, qui repose sur l'utilisation du pairage (*matching*) pour analyser le FDI, constitue une avancée significative dans ce domaine.

En démontrant l'efficacité de la méthode de pairage dans l'identification des FDI liées aux caractéristiques linguistiques, cette étude ouvre la voie à une utilisation plus large de cette approche dans le domaine de l'évaluation éducative. Les chercheurs pourraient s'inspirer de cette approche pour analyser d'autres types de facteurs influençant la performance des élèves dans diverses disciplines et contextes éducatifs. De plus, la prise en compte des données manquantes dans ce type d'analyse pourrait inspirer de nouvelles recherches visant à optimiser les méthodes statistiques appliquées aux données complexes.

5.4.4 Implications pour l'équité en éducation

Enfin, cette étude soulève des enjeux fondamentaux en lien avec l'équité en éducation. Si les évaluations standardisées influencent l'accès à des opportunités éducatives et professionnelles, il est important qu'elles mesurent fidèlement les habiletés des élèves, sans favoriser ou désavantager certains groupes en raison de facteurs linguistiques. Cette recherche pourrait aider à promouvoir des politiques éducatives plus inclusives. Par exemple, des ajustements aux conditions de passation des tests pourraient être envisagés, comme l'introduction d'un soutien linguistique spécifique pour les élèves L2 ou l'adoption d'une approche différenciée dans l'évaluation des compétences. Une telle démarche s'inscrit dans une vision plus large de l'éducation, où l'objectif est d'assurer que tous les élèves, quel que soit leur profil linguistique, aient des chances égales de réussir.

CONCLUSION

Cette recherche visait à explorer l'influence des facteurs linguistiques sur la performance des élèves en mathématiques, en s'appuyant sur les données de l'enquête PISA 2012 et en mobilisant la méthode de fonctionnement différentiel des items (FDI). En mettant l'accent sur l'approche par pairage des données, et non exclusivement sur la détection du FDI comme fin en soi, nous avons cherché à démontrer l'utilité du score de propension pour contrôler les variables confondantes et mieux isoler l'effet du statut linguistique dans un contexte de données issues d'une structure en livret.

Les analyses ont permis d'identifier plusieurs items présentant un FDI entre les groupes d'élèves de langue maternelle (L1) et de langue seconde (L2), même après le pairage. Cela met en lumière la persistance de certains obstacles linguistiques dans les évaluations de mathématiques, susceptibles de compromettre l'équité des mesures. Les résultats soulignent également l'intérêt méthodologique de l'approche par score de propension pour affiner l'identification du FDI, surtout dans des contextes complexes où des facteurs de confusion multiples interagissent.

La contribution originale de cette étude réside ainsi dans l'intégration d'une méthode de pairage au cœur d'un dispositif d'analyse du FDI appliqué à des données internationales à structure complexe. Elle offre une perspective nuancée sur la relation entre langue du test et performance en mathématiques, en tenant compte à la fois des caractéristiques individuelles des élèves et de la structure de l'évaluation elle-même.

Sur le plan pédagogique, les résultats de cette recherche incitent à une réflexion sur les pratiques d'enseignement des mathématiques auprès des élèves pour qui la langue du test est une langue seconde. Ils invitent à considérer la langue comme un facteur central dans l'élaboration des consignes, des supports d'apprentissage et des stratégies d'enseignement. Sur le plan de l'élaboration des tests, cette étude suggère la nécessité d'un encadrement plus rigoureux des items susceptibles d'introduire un biais linguistique, même lorsque ceux-ci évaluent en apparence des compétences non langagières.

D'un point de vue méthodologique, la recherche souligne l'importance d'adopter des approches analytiques capables de contrôler les effets de confusion afin de mieux cerner les sources réelles de différenciation dans les réponses aux tests. Enfin, sur le plan politique, ces résultats appellent à une vigilance accrue quant aux conditions d'évaluation des élèves dont la langue du test est une langue

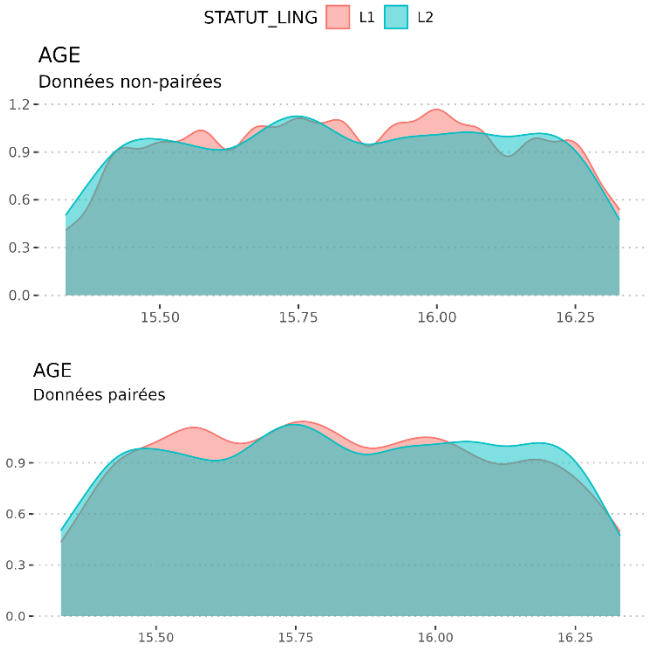
seconde, afin de s'assurer que les dispositifs d'évaluation ne reproduisent pas, ou n'aggravent pas, les inégalités déjà présentes dans le système éducatif.

Dans l'ensemble, cette étude contribue à l'avancement des connaissances dans le champ de la mesure en éducation, en démontrant comment une approche méthodologique rigoureuse peut appuyer des pratiques évaluatives plus justes et inclusives. Elle ouvre également la voie à de futures recherches visant à affiner les méthodes de détection du FDI, à explorer d'autres sources de biais potentielles, et à soutenir une meilleure compréhension des interactions entre langue et évaluation des compétences scolaires.

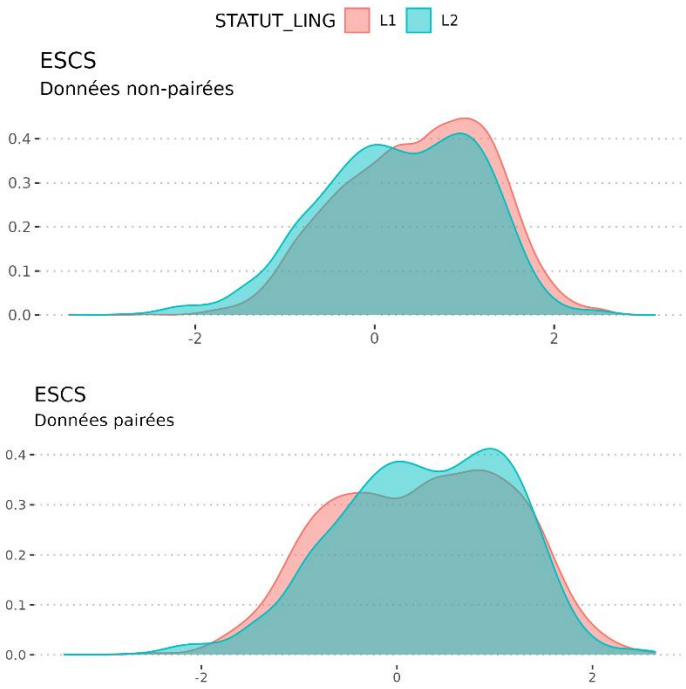
ANNEXE A

Distribution des covariables selon le statut linguistique, avant et après pairage

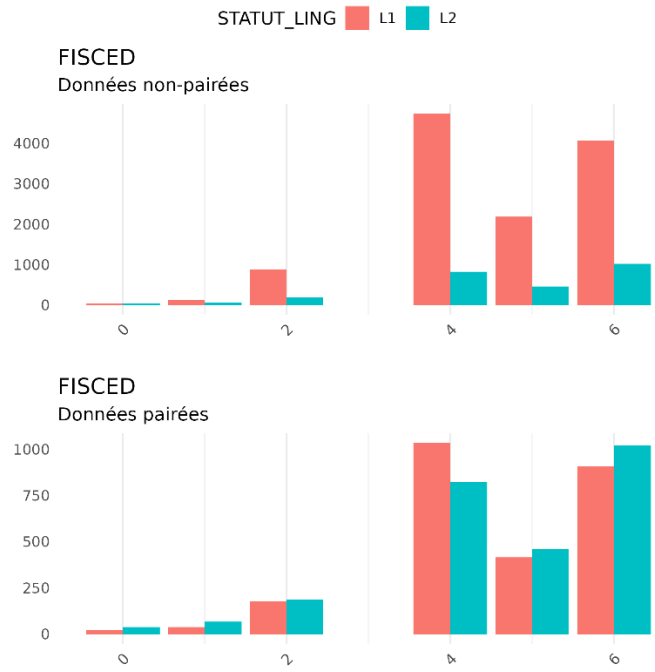
A.1. Distribution de la covariable AGE (âge) avant et après pairage



A.2. Distribution de la covariable ESCS (statut socio-économique) avant et après pairage



A.3. Distribution de la variable FISCED (niveau d'éducation du père) selon le statut linguistique, avant et après pairage

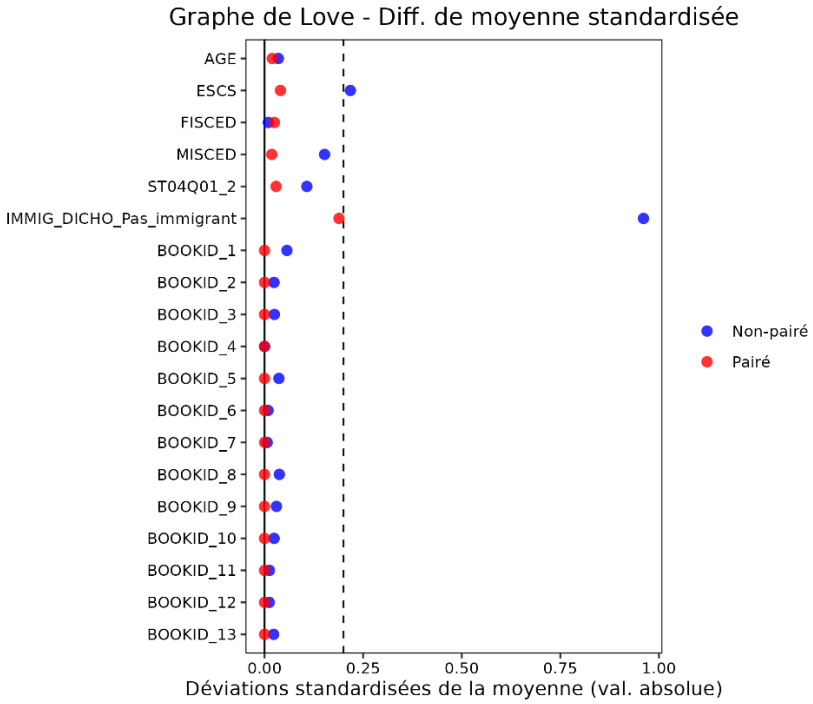


A.4. Distribution de la variable MISCED (niveau d'éducation de la mère) selon le statut linguistique, avant et après pairage

A.5. Distribution de la variable ST04Q01 (sexe) selon le statut linguistique, avant et après pairage

ANNEXE B

Différences standardisées des moyennes des covariables avant et après pairage



ANNEXE C

Résultats des tests de proportions sur les données non-pairées

Item	n_L1	n_L2	prop_L1	prop_L2	p	h
PM571Q01	3,677	804	0,53	0,40	< 0,001	0,26
PM909Q02	3,737	810	0,71	0,59	< 0,001	0,25
PM034Q01T	3,667	824	0,47	0,36	< 0,001	0,22
PM953Q02	3,688	811	0,61	0,50	< 0,001	0,22
PM446Q01	3,684	823	0,80	0,71	< 0,001	0,21
PM496Q02	3,677	804	0,69	0,59	< 0,001	0,21
PM155Q04T	3,667	824	0,65	0,55	< 0,001	0,20
PM408Q01T	3,684	823	0,47	0,37	< 0,001	0,20
PM982Q04	3,794	769	0,51	0,41	< 0,001	0,20
PM954Q01	3,688	811	0,73	0,64	< 0,001	0,19
PM909Q01	3,736	810	0,95	0,90	< 0,001	0,19
PM954Q02	3,688	811	0,38	0,29	< 0,001	0,19
PM919Q01	3,688	811	0,87	0,80	< 0,001	0,19
PM420Q01T	3,684	823	0,69	0,60	< 0,001	0,19
PM953Q04D	3,307	736	0,21	0,14	< 0,001	0,19
PM924Q02	3,690	784	0,60	0,51	< 0,001	0,18

PM998Q02	3,737	810	0,84	0,77	< 0,001	0,18
PM909Q03	3,737	810	0,39	0,31	< 0,001	0,17
PM155Q02D	3,192	706	0,78	0,71	< 0,001	0,16
PM155Q01	3,667	824	0,75	0,68	< 0,001	0,16
PM803Q01T	3,667	824	0,35	0,28	< 0,001	0,15
PM905Q01T	3,688	811	0,83	0,77	< 0,001	0,15
PM918Q02	3,690	784	0,82	0,76	< 0,001	0,15
PM955Q01	3,737	810	0,80	0,74	< 0,001	0,14
PM982Q01	3,794	769	0,88	0,83	< 0,001	0,14
PM192Q01T	3,677	804	0,48	0,41	< 0,001	0,14
PM995Q03	3,690	784	0,52	0,45	< 0,001	0,14
PM915Q02	3,794	769	0,76	0,70	< 0,001	0,14
PM447Q01	3,684	823	0,73	0,67	< 0,001	0,13
PM918Q05	3,690	784	0,82	0,77	0,0085	0,12
PM00FQ01	3,690	784	0,48	0,42	0,0019	0,12
PM411Q01	3,667	824	0,57	0,51	0,0044	0,12
PM919Q02	3,688	811	0,49	0,43	0,002	0,12
PM918Q01	3,690	784	0,87	0,83	0,0029	0,11
PM949Q01T	3,737	810	0,75	0,70	0,0107	0,11

PM992Q01	3,794	769	0,85	0,81	0,0011	0,11
PM828Q02	3,683	823	0,63	0,58	0,0148	0,10
PM982Q03T	3,794	769	0,62	0,57	0,0114	0,10
PM906Q01	3,794	769	0,58	0,53	0,013	0,10
PM923Q01	3,690	784	0,58	0,53	0,0262	0,10
PM564Q02	3,677	804	0,49	0,44	0,007	0,10
PM906Q02	3,495	716	0,54	0,49	0,0197	0,10
PM411Q02	3,667	824	0,53	0,48	0,0056	0,10
PM423Q01	3,677	804	0,79	0,75	0,0235	0,10
PM955Q02	3,737	810	0,33	0,29	0,0188	0,09
PM442Q02	3,667	824	0,40	0,36	0,0754	0,08
PM800Q01	3,684	823	0,85	0,82	0,1208	0,08
PM564Q01	3,677	804	0,46	0,42	0,0521	0,08
PM496Q01T	3,677	804	0,57	0,53	0,0576	0,08
PM955Q03	3,458	745	0,10	0,08	0,1463	0,07
PM828Q03	3,684	823	0,29	0,26	0,0426	0,07
PM982Q02	3,794	769	0,34	0,31	0,1054	0,06
PM828Q01	3,684	823	0,37	0,34	0,1978	0,06
PM305Q01	3,677	804	0,61	0,58	0,0861	0,06

PM603Q01T	3,677	804	0,46	0,43	0,068	0,06
PM273Q01T	3,684	823	0,54	0,51	0,0798	0,06
PM953Q03	3,688	811	0,52	0,49	0,0948	0,06
PM995Q01	3,690	784	0,52	0,49	0,2469	0,06
PM923Q04	3,690	784	0,18	0,16	0,1714	0,05
PM992Q02	3,794	769	0,23	0,21	0,1838	0,05
PM033Q01	3,667	824	0,77	0,75	0,1207	0,05
PM406Q01	3,677	804	0,30	0,28	0,206	0,04
PM954Q04	3,688	811	0,32	0,30	0,1487	0,04
PM559Q01	3,684	823	0,65	0,63	0,3383	0,04
PM915Q01	3,794	769	0,39	0,37	0,4034	0,04
PM905Q02	3,688	811	0,53	0,51	0,3268	0,04
PM00GQ01	3,737	810	0,08	0,07	0,443	0,04
PM155Q03D	2,777	662	0,17	0,16	0,4048	0,03
PM464Q01T	3,684	823	0,26	0,25	0,4772	0,02
PM474Q01	3,667	824	0,72	0,71	0,6798	0,02
PM998Q04T	3,737	810	0,33	0,32	0,7947	0,02
PM943Q01	3,688	811	0,50	0,49	0,5728	0,02
PM406Q02	3,677	804	0,20	0,20	0,6361	0,00

PM903Q01	3,094	686	0,17	0,17	0,6261	0,00
PM923Q03	3,690	784	0,56	0,56	0,9298	0,00
PM995Q02	3,690	784	0,03	0,03	0,5924	0,00
PM949Q03	3,617	784	0,32	0,33	0,754	-0,02
PM903Q03	3,690	784	0,31	0,32	0,7531	-0,02
PM00KQ02	3,794	769	0,14	0,15	0,6333	-0,03
PM992Q03	3,794	769	0,07	0,08	0,3407	-0,04
PM949Q02T	3,737	810	0,32	0,34	0,4863	-0,04
PM462Q01D	3,456	774	0,03	0,05	0,0899	-0,10
PM943Q02	3,688	811	0,03	0,05	0,0039	-0,10
PM446Q02	3,684	823	0,07	0,10	0,0064	-0,11

ANNEXE D

Résultats des tests de proportions sur les données pairées

Item	n_L1	n_L2	prop_L1	prop_L2	p	h
PM571Q01	804	804	0,51	0,40	< 0,001	0,22
PM155Q02D	711	706	0,80	0,71	< 0,001	0,21
PM909Q02	810	810	0,69	0,59	< 0,001	0,21
PM918Q01	784	784	0,90	0,83	< 0,001	0,21
PM00FQ01	784	784	0,52	0,42	< 0,001	0,20
PM906Q01	769	769	0,62	0,53	0,0017	0,18
PM924Q02	784	784	0,59	0,51	0,0012	0,16
PM915Q02	769	769	0,77	0,70	0,0022	0,16
PM909Q01	810	810	0,94	0,90	0,0091	0,15
PM496Q02	804	804	0,66	0,59	0,0086	0,14
PM034Q01T	824	824	0,43	0,36	0,003	0,14
PM408Q01T	823	823	0,44	0,37	0,0067	0,14
PM982Q04	769	769	0,48	0,41	0,0056	0,14
PM953Q02	811	811	0,57	0,50	0,0062	0,14
PM906Q02	717	716	0,56	0,49	0,0162	0,14
PM953Q04D	722	736	0,19	0,14	0,0385	0,14

PM919Q01	811	811	0,85	0,80	0,0106	0,13
PM803Q01T	824	824	0,34	0,28	0,0219	0,13
PM954Q01	811	811	0,70	0,64	0,0044	0,13
PM909Q03	810	810	0,37	0,31	0,0183	0,13
PM905Q01T	811	811	0,82	0,77	0,0162	0,12
PM918Q05	784	784	0,82	0,77	0,0329	0,12
PM998Q02	810	810	0,82	0,77	0,0101	0,12
PM828Q02	823	823	0,64	0,58	0,0153	0,12
PM411Q01	824	824	0,57	0,51	0,0336	0,12
PM919Q02	811	811	0,49	0,43	0,0167	0,12
PM446Q01	823	823	0,76	0,71	0,0101	0,11
PM982Q01	769	769	0,87	0,83	0,0277	0,11
PM155Q01	824	824	0,73	0,68	0,0177	0,11
PM800Q01	823	823	0,86	0,82	0,0689	0,11
PM982Q02	769	769	0,36	0,31	0,059	0,11
PM923Q04	784	784	0,20	0,16	0,0637	0,10
PM982Q03T	769	769	0,62	0,57	0,0427	0,10
PM155Q04T	824	824	0,60	0,55	0,0321	0,10
PM564Q02	804	804	0,49	0,44	0,0278	0,10

PM918Q02	784	784	0,80	0,76	0,0382	0,10
PM423Q01	804	804	0,79	0,75	0,1224	0,10
PM955Q01	810	810	0,78	0,74	0,0701	0,09
PM954Q02	811	811	0,33	0,29	0,097	0,09
PM305Q01	804	804	0,62	0,58	0,1269	0,08
PM192Q01T	804	804	0,45	0,41	0,1183	0,08
PM603Q01T	804	804	0,47	0,43	0,0881	0,08
PM923Q01	784	784	0,57	0,53	0,1551	0,08
PM995Q03	784	784	0,49	0,45	0,1423	0,08
PM992Q01	769	769	0,84	0,81	0,1617	0,08
PM955Q03	744	745	0,10	0,08	0,3184	0,07
PM949Q01T	810	810	0,73	0,70	0,1857	0,07
PM955Q02	810	810	0,32	0,29	0,2135	0,07
PM447Q01	823	823	0,70	0,67	0,3127	0,06
PM828Q01	823	823	0,37	0,34	0,2368	0,06
PM442Q02	824	824	0,39	0,36	0,3596	0,06
PM420Q01T	823	823	0,63	0,60	0,2054	0,06
PM564Q01	804	804	0,45	0,42	0,3389	0,06
PM496Q01T	804	804	0,56	0,53	0,3414	0,06

PM995Q01	784	784	0,52	0,49	0,3905	0,06
PM995Q02	784	784	0,04	0,03	0,2207	0,05
PM155Q03D	642	662	0,18	0,16	0,443	0,05
PM033Q01	824	824	0,77	0,75	0,1848	0,05
PM828Q03	823	823	0,28	0,26	0,3171	0,05
PM903Q03	784	784	0,34	0,32	0,421	0,04
PM273Q01T	823	823	0,53	0,51	0,3238	0,04
PM905Q02	811	811	0,53	0,51	0,3981	0,04
PM411Q02	824	824	0,50	0,48	0,2571	0,04
PM953Q03	811	811	0,51	0,49	0,4269	0,04
PM992Q02	769	769	0,22	0,21	0,577	0,02
PM406Q01	804	804	0,29	0,28	0,5439	0,02
PM954Q04	811	811	0,31	0,30	0,5176	0,02
PM559Q01	823	823	0,64	0,63	0,6446	0,02
PM923Q03	784	784	0,57	0,56	0,7991	0,02
PM00GQ01	810	810	0,07	0,07	0,9238	0,00
PM00KQ02	769	769	0,15	0,15	0,8865	0,00
PM915Q01	769	769	0,37	0,37	1	0,00
PM943Q01	811	811	0,49	0,49	0,9209	0,00

PM949Q03	786	784	0,33	0,33	1	0,00
PM998Q04T	810	810	0,32	0,32	0,8729	0,00
PM949Q02T	810	810	0,33	0,34	0,9162	-0,02
PM464Q01T	823	823	0,24	0,25	0,9086	-0,02
PM406Q02	804	804	0,19	0,20	0,8995	-0,03
PM903Q01	655	686	0,16	0,17	0,7117	-0,03
PM992Q03	769	769	0,07	0,08	0,2804	-0,04
PM474Q01	824	824	0,69	0,71	0,4519	-0,04
PM462Q01D	770	774	0,04	0,05	0,5371	-0,05
PM943Q02	811	811	0,04	0,05	0,2715	-0,05
PM446Q02	823	823	0,08	0,10	0,1243	-0,07

ANNEXE E

Résultats de l'analyse de sensibilité sur les données pairées

Item	unconf_estimate	max_sig_gamma
PM00FQ01	< 0,001	1,2
PM00GQ01	0,38546	
PM00KQ02	0,39002	
PM033Q01	0,07507	
PM034Q01T	< 0,001	1,1
PM155Q01	0,00644	1,0
PM155Q02D	< 0,001	1,2
PM155Q03D	0,06326	
PM155Q04T	0,01307	1,0
PM192Q01T	0,04645	1,0
PM273Q01T	0,13774	
PM305Q01	0,05318	
PM406Q01	0,23579	
PM406Q02	0,39795	
PM408Q01T	0,00314	1,1
PM411Q01	0,01344	1,0

PM411Q02	0,10619	
PM420Q01T	0,09063	
PM423Q01	0,05203	
PM442Q02	0,15085	
PM446Q01	0,00312	1,1
PM446Q02	0,03545	1,0
PM447Q01	0,13013	
PM462Q01D	0,17484	
PM464Q01T	0,41087	
PM474Q01	0,19383	
PM496Q01T	0,14685	
PM496Q02	0,00368	1,1
PM559Q01	0,28681	
PM564Q01	0,14193	
PM564Q02	0,01124	1,0
PM571Q01	< 0,001	1,3
PM603Q01T	0,03747	1,0
PM800Q01	0,02568	1,0
PM803Q01T	0,00927	1,0

PM828Q01	0,09742	
PM828Q02	0,00526	1,0
PM828Q03	0,12759	
PM903Q01	0,18773	
PM903Q03	0,17322	
PM905Q01T	0,00606	1,1
PM905Q02	0,16676	
PM906Q01	< 0,001	1,1
PM906Q02	0,00154	1,1
PM909Q01	0,00266	1,2
PM909Q02	< 0,001	1,2
PM909Q03	0,00776	1,0
PM915Q01	0,45798	
PM915Q02	< 0,001	1,1
PM918Q01	< 0,001	1,4
PM918Q02	0,01227	1,0
PM918Q05	0,01090	1,0
PM919Q01	0,00427	1,1
PM919Q02	0,00477	1,1

PM923Q01	0,06061	
PM923Q03	0,36049	
PM923Q04	0,02072	1,0
PM924Q02	< 0,001	1,1
PM943Q01	0,42414	
PM943Q02	0,08094	
PM949Q01T	0,07077	
PM949Q02T	0,41403	
PM949Q03	0,35296	
PM953Q02	0,00284	1,1
PM953Q03	0,18673	
PM953Q04D	0,00317	1,1
PM954Q01	0,00137	1,1
PM954Q02	0,03668	1,0
PM954Q04	0,22287	
PM955Q01	0,02582	1,0
PM955Q02	0,09378	
PM955Q03	0,08532	
PM982Q01	0,01001	1,1

PM982Q02	0,02169	1,0
PM982Q03T	0,01391	1,0
PM982Q04	0,00170	1,1
PM992Q01	0,06292	
PM992Q02	0,24159	
PM992Q03	0,10325	
PM995Q01	0,15832	
PM995Q02	0,07045	
PM995Q03	0,05346	
PM998Q02	0,00328	1,1
PM998Q04T	0,39346	

BIBLIOGRAPHIE

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219–234. https://doi.org/10.1207/S15324818AME1403_2

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometric issues. *Educational Assessment, 8*(3), 231–257. https://doi.org/10.1207/S15326977EA0803_02

Abedi, J., Bailey, A., Butler, F., Castellon-Wellington, M., Leon, S., & Mirocha, J. (2005). The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives (CSE Report 663). *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*. https://eric.ed.gov/?id=ED492891

Abedi, J., & Gándara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues and Practice, 25*(4), 36–46. https://doi.org/10.1111/j.1745-3992.2006.00077.x

Agresti, A. (2018). Statistical methods for the social sciences (5th ed.). *Pearson*. https://elibrary.pearson.de/book/99.150005/9781292220345

Ajello, A. M., Caponera, E., & Palmerio, L. (2018). Italian students' results in the PISA mathematics test: Does reading competence matter? *European Journal of Psychology of Education, 33*(3), 505–520. https://doi.org/10.1007/s10212-018-0385-x

Akcan, R., & Kabasakal, K. A. (2023). The impact of missing data on the performance of DIF detection methods. *Journal of Measurement and Evaluation in Education and Psychology, 14*(1), 95–105. https://doi.org/10.21031/epod.1200131

Akyüz, G. (2014). The effects of student and school factors on mathematics achievement in TIMSS 2011. *Educational Sciences: Theory & Practice, 14*(1), 1–16. https://doi.org/10.15390/ES.2014.1255

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing.* American Educational Research Association.

André, N., Loye, N., & Laurencelle, L. (2015). La validité psychométrique: Un regard global sur le concept centenaire, sa genèse, ses avatars. *Mesure et Évaluation en Éducation, 37*(3), 125–148. https://doi.org/10.7202/1036330ar

Aricak, O. T., Guldal, H., & Erdogan, I. (2023). Which noncognitive features provide more information about reading performance? A data-mining approach to big educational data. *Journal of Pacific Rim Psychology*, *17*, Article e2072. https://doi.org/10.1177/18344909231164025

Arikan, S., van de Vijver, F., & Yagmur, K. (2018). Propensity score matching helps to understand sources of DIF and mathematics performance differences of Indonesian, Turkish, Australian, and Dutch students in PISA. *International Journal of Research in Education and Science*, *4*, 69-81. https://doi.org/10.21890/ijres.382936

Arthur, Y. D., Dogbe, C. S. K., & Asiedu-Addo, S. K. (2022). Enhancing performance in mathematics through motivation, peer assisted learning, and teaching quality: The mediating role of student interest. *EURASIA Journal of Mathematics, Science and Technology Education*, *18*(2), em2072.

Assilamehou-Kunz, Y., & Testé, B. (2022). Les biais linguistiques ou comment notre façon de décrire autrui peut contribuer au maintien des inégalités sociales. *In-Mind Magazine*. Récupéré de https://fr.in-mind.org/fr/article/les-biais-linguistiques-ou-comment-notre-facon-de-decrire-autrui-peut-contribuer-au-0

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine, 28*(25), 3083–3107.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*(3), 399-424. https://doi.org/10.1080/00273171.2011.568786

Aydin, O. (2024). A description of missing data in single-case experimental designs studies and an evaluation of single imputation methods. *Behavior Modification*, *48*(3), 312-359. https://doi.org/10.1177/01454455241226879

Baker, F. B., & Kim, S. H. (2004). Item response theory: Parameter estimation techniques (2de ed.). *CRC Press*. https://doi.org/10.1201/9781482276725

Banks, K., Jeddeeni, A., & Walker, C. M. (2016). Assessing the effect of language demand in bundles of math word problems. *International Journal of Testing*, *16*(4), 269–287. https://doi.org/10.1080/15305058.2015.111397

Barwell, R., Barton, B., & Setati, M. (2007). Multilingual issues in mathematics education: Introduction. *Educational Studies in Mathematics*, *64*, 113-119.

Barwell, R., Moschkovich, J., & Setati Phakeng, M. (2017). Language diversity and mathematics: Second language, bilingual, and multilingual learners. In J. Cai (Ed.), *Compendium for Research in Mathematics Education*, 583–606. National Council of Teachers of Mathematics.

Barwell, R., Wessel, L., & Parra, A. (2019). Language diversity and mathematics education: New developments. *Research in Mathematics Education*, *21*(2), 113-118.

Baten, E., Pixner, S., & Desoete, A. (2019). Motivational and math anxiety perspective for mathematical learning and learning difficulties. In A. Fritz, V. G. Haase, & P. Räsänen (Eds.), *International Handbook of Mathematical Learning Difficulties: From the Laboratory to the Classroom*, 457-467. Springer.

Bengoetxea, K., & Gonzalez-Dios, I. (2021). Multiaztertest: A multilingual analyzer on multiple levels of language for readability assessment. *arXiv preprint arXiv:2109.04870*.

Bergem, O. K., Kaarstein, H., & Nilsen, T. (2016). TIMSS 2015. In *Vi Kan Lykkes i Realfag: Resultater Og Analyser Fra TIMSS 2015*, 11–21. Universitetsforlaget. https://doi.org/10.18261/97882150279999-2016-02

Bodin, A. (2016). Le langage, un biais dans l'évaluation? *Tangente Éducation*, *38*. Récupéré le 10 juin 2024, de https://www.tangente-education.com/articles/TE38/8_Pisa_TE38_14_16.pdf

Bolt, D., & Stout, W. (1996). Differential item functioning: Its multidimensional model and resulting SIBTEST detection procedure. *Behaviormetrika*, *23*, 67-95.

Bond, T. G., & Fox, C. M. (2013). Applying the Rasch model: Fundamental measurement in the human sciences. *Psychology Press*.

Bourny, G., Dupe, C., Robin, I., & Rocher, T. (2001). Les élèves de 15 ans: Premiers résultats d'une évaluation internationale des acquis des élèves (PISA). *Note d'information - Direction de la programmation et du développement, 52,* 1–6.

Bronfenbrenner, U. (1979). The ecology of human development: Experiments by nature and design. *Harvard University Press*, *2*, 139-163.

Buono, S., & Jang, E. E. (2021). The effect of linguistic factors on assessment of English language learners' mathematical ability: A differential item functioning analysis. *Educational Assessment*, *26*(2), 125-144.

Castelli, F., Glaser, D. E., & Butterworth, B. (2006). Discrete and analogue quantity processing in the parietal lobe: A functional MRI study. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(12), 4693–4698. https://doi.org/10.1073/pnas.0600444103

Cheema, J. R. (2019). Cross-country gender DIF in PISA science literacy items. *European Journal of Developmental Psychology*, *16*(2), 152-166.

Chen, F. (2010). Differential language influence on math achievement [Thèse de doctorat, The University of North Carolina at Greensboro].

Chen, M. Y., Liu, Y., & Zumbo, B. D. (2020). A propensity score method for investigating differential item functioning in performance assessment. *Educational and Psychological Measurement*, *80*(3), 476-498.

Cheng, Q., Wang, J., Hao, S., & Shi, Q. (2014). Mathematics performance of immigrant students across different racial groups: An indirect examination of the influence of culture and schooling. *Journal of International Migration and Integration*, *15*, 589-607.

Chinn, S. (2009). Mathematics anxiety in secondary students in England. *Dyslexia*, *15*(1), 61-68. https://doi.org/10.1002/dys.381

Christofides, L.N., & Swidinsky, R. (2010). The economic returns to the knowledge and use of a second official language: English in Quebec and french in the Rest-of-Canada. *Canadian Public Policy, 36*, 137 - 158.

Chronaki, A., & Planas, N. (2018). Language diversity in mathematics education research: A move from language as representation to politics of representation. *ZDM*, *50*(6), 1101-1111. https://doi.org/link.springer.com/article/10.1007/s11858-018-0942-4

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. An NCME Instructional Module. *Educational Measurement: Issues and Practice*, *17*(1), 31-44.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). *Routledge*. https://doi.org/10.4324/9780203771587

Conseil des ministres de l'Éducation, Canada (CMEC). (2019). À la hauteur : Résultats canadiens de l'étude PISA 2018 de l'OCDE – Le rendement des jeunes de 15 ans du Canada en compétences globales. https://www.cmec.ca/Publications/Lists/Publications/Attachments/419/PISA2018_GC_Report_FR.pdf

Conseil des ministres de l'Éducation, Canada (CMEC). (2020). Cadre pancanadien pour les compétences globales selon la perspective du système. https://www.cmec.ca/Publications/Lists/Publications/Attachments/403/Cadre%20pancanadien%20pour%20les%20comp%C3%A9tences%20globales_FR.pdf

Conseil des ministres de l'Éducation, Canada (CMEC). (2024). Elementary-Secondary Education. *CMEC.* Récupéré 13 avril 2024, de https://www.cmec.ca/680/Elementary-Secondary_Education.html

Cordero, J. M., Cristóbal, V., & Santín, D. (2018). Causal inference on education policies: A survey of empirical studies using PISA, TIMSS and PIRLS. *Journal of Economic Surveys*, *32*(3), 878-915.

Council of Ministers of Education, Canada (CMEC). (2013). Measuring up: Canadian results of the OECD PISA study—The performance of Canada's youth in mathematics, reading, and science. 2012 first results for Canadians aged 15. *CMEC*.

Cuevas, M., & Cervantes, V. H. (2012). Differential item functioning detection with logistic regression. *Mathématiques et sciences humaines. Mathematics and Social Sciences*, (199), 45–59.

Das, G. C., & Sinha, S. (2017). Effect of socioeconomic status on performance in mathematics among students of secondary schools of Guwahati city. *IOSR Journal of Mathematics*, *13*(1), 26-33.

Davis, S. (2023). Multilingual learners in canadian french immersion programs: Looking back and moving forward. *The Canadian Modern Language Review*, *79*(3), 163-180.

Dehaene, S., Spelke, E. S., Pinel, P., Stanescu, R., & Tsivkin, S. (1999). Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science*, *284*(5416), 970–974. https://doi.org/10.1126/science.284.5416.970

Dehaene, S., Molko, N., Cohen, L., & Wilson, A. J. (2004). Arithmetic and the brain. *Current Opinion in Neurobiology*, *14*(2), 218–224. https://doi.org/10.1016/j.conb.2004.03.008

Delage, V., Daker, R. J., Trudel, G., Lyons, I. M., & Maloney, E. A. (2024). It is a "small world": Relations between performance on five spatial tasks and five mathematical tasks in undergraduate students. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*.

Demonty, I., Blondin, C., Matoul, A., Baye, A., & Lafontaine, D. (2013). La culture mathématique à 15 ans. Premiers résultats de PISA 2012 en Fédération Wallonie-Bruxelles. *Cahiers des Sciences de l'Éducation*, *34*.

Dempster, E. R. & Reddy, V. (2007). Item readability and science achievement in TIMSS 2003 in South Africa. *Science Education*, *91*(6), 906-925. https://doi.org/10/cd687q

Dinglasan, B. L., & Patena, A. (2013). Students performance on departmental examination: Basis for math intervention program. *University of Alberta School of Business Research Paper,* 2013-1308.

Dion-Viens, D. (2023). Taux d'échec trop élevé en maths : Québec a fait passer des élèves qui ont eu 55 %. *Journal de Québec*. Récupéré de https://www.journaldequebec.com/2023/01/26/des-eleves-ont-reussi-lexamen-de-math-avec-55

DiPrete, T. A., & Gangl, M. (2004). Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological Methodology, 34* (1), 271–310.

Doyle, L., Easterbrook, M.J., & Harris, P.R. (2022). Roles of socioeconomic status, ethnicity and teacher beliefs in academic grading. *The British Journal of Educational Psychology, 93*, 91 - 112.

Durand-Guerrier, V., & Chellougui, F. (2015). Aspects culturels et langagiers dans l'enseignement des mathématiques - Compte rendu du Groupe de Travail n° 8. In L. Theis (Ed.), Pluralités culturelles et universalité des mathématiques : Enjeux et perspectives pour leur enseignement et leur apprentissage – *Actes du colloque EMF2015 – GT8*, 711–717.

Education Quality and Accountability Office. (2020). Cadre d'évaluation des compétences en mathématiques, 9e année. Récupéré de https://www.eqao.com/wp-content/uploads/2020/12/annual-report-2019-2020.pdf

Embretson, S. E., & Reise, S. P. (2013). Item response theory for psychologists. *Psychology Press.*

Enders, C. K. (2022). Applied missing data analysis. *Guilford Press*.

Engelhard Jr., G. (2013). Invariant measurement: Using Rasch models in the social, behavioral, and health sciences. *Routledge*.

Fan, C., Wang, L., & Wei, L. (2017). Comparing two tests for two rates. *The American Statistician*, *71*(3), 275–281. https://doi-org/10.1080/00031305.2016.1246263

Fielding, A. (1974). Review of statistical methods for rates and proportions, by J. L. Fleiss. *Journal of the Royal Statistical Society. Series A (General)*, *137*(2), 270–271. https://doi.org/10.2307/2344565

French, B.F. (2020). Test bias. in F. Maggino (Éd.), *Encyclopedia of Quality of Life and Well-Being Research*. *Springer, Cham.* https://doi.org/10.1007/978-3-319-69909-7_2998-2

Frey, A., & Bernhardt, R. (2012). On the importance of using balanced booklet designs in PISA. *Psychological Test and Assessment Modeling*, *54*(4), 397-417.

Gálvez-Lara, M., Moriana, J. A., Vilar-López, R., Fasfous, A. F., Hidalgo-Ruzzante, N., & Pérez-García, M. (2015). Validation of the cross-linguistic naming test: A naming test for different cultures? A preliminary study in the Spanish population. *Journal of Clinical and Experimental Neuropsychology*, *37*(1), 102-112.

Gayat, E., & Porcher, R. (2012). Comparaison de l'efficacité de deux thérapeutiques en l'absence de randomisation: Intérêts et limites des méthodes utilisant les scores de propension. *Réanimation, 21*(1), 109–116.

Gierl, M. J., Gotzmann, A., & Boughton, K. A. (2004). Performance of SIBTEST when the percentage of DIF items is large. *Applied Measurement in Education*, *17*(3), 241-264.

Gómez-Benito, J., Sireci, S., Padilla, J. L., Hidalgo, M. D., & Benítez, I. (2018). Differential Item Functioning: Beyond validity evidence based on internal structure. *Psicothema*, *30*(1), 104-109.

Goodrich, J. M., Koziol, N. A., & Yoon, H. (2021). Are translated mathematics items a valid accommodation for dual language learners? Evidence from ECLS-K. *Early Childhood Research Quarterly*, *57*, 89-101.

Gouvernement du Canada, Canadian Heritage (2021). Official language minority communities with at least one school in the minority language. *Government of Canada*. Récupéré le 3 juin 2024, de https://www.canada.ca/en/canadian-heritage/services/official-languages-bilingualism/publications/minority-communities.html

Granger, E., Watkins, T., Sergeant, J. C., & Lunt, M. (2020). A review of the use of propensity score diagnostics in papers published in high-ranking medical journals. *BMC Medical Research Methodology, 20*, 1–9.

Grégoire, J. (2014). La validité des scores. *Universalis*. Récupéré le 4 juin 2024, de https://www.universalis.fr/encyclopedie/psychometrie/4-la-validite-des-scores/

Hancock, G. R., Mueller, R. O., & Stapleton, L. M. (2018). The reviewer's guide to quantitative methods in the social sciences. *Routledge*. https://doi.org/10.4324/9781315755649

Grimm, K. J. (2008). Longitudinal associations between reading and mathematics achievement. *Developmental Neuropsychology*, *33*(3), 410-426.

Haag, N., Heppt, B., Stanat, P., Kuhl, P., & Pant, H. A. (2013). Second language learners' performance in mathematics: Disentangling the effects of academic language features. *Learning and Instruction*, *28*, 24-34. https://doi.org/10.1016/j.learninstruc.2013.04.001

Hasegawa, R., & Small, D. (2017). Sensitivity analysis for matched pair analysis of binary data: From worst case to average case analysis. *Biometrics*, *73*(4), 1424-1432.

Helwig, R., Rozek-Tedesco, M. A., Tindal, G., Heath, B., & Almond, P. J. (1999). Reading as an access to mathematics problem solving on multiple-choice tests for sixth-grade students. *The Journal of Educational Research*, *93*(2), 113-125.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software, 42*(8), 1–28. https://doi.org/10.18637/jss.v042.i08

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, & H. I. Braun, *Test Validity* (pp. 129-145). Lawrence Erlbaum.

Huba, M. E., & Freed, J. E. (2000). Learner-centered assessment on college campuses: Shifting the focus from teaching to learning. *Allyn & Bacon*.

Inci, A. M., Peker, B., & Kucukgencay, N. (2023). Realistic mathematics education. *Current Studies in Educational Disciplines*, 66-83.

Jaegers, D., & Lafontaine, D. (2020). Aspirer à une carrière mathématique: quel rôle jouent le soutien et les attentes de l'enseignant chez les filles et les garçons? *Revue française de pédagogie*, 31-47.

Jiang, H., & Stout, W. (1998). Improved Type I error control and reduced estimation bias for DIF detection using SIBTEST. *Journal of Educational and Behavioral Statistics*, *23*(4), 291-322.

Jun-bin, Z. (2011). The Types of Assessment Bias and the Ways of Identification in American Examinations. *Comparative Education Review*.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*(1), 79-93.

Kasimov, M. (2019). International assessment research: The teaching of PISA and PIRLS materials in general secondary schools, scientific methods. *Theoretical & Applied Science, 80*, 612-617.

Kaushik, D., Garg, M., & Mishra, A. (2023). Exploring developmental pathways: An in-depth analysis of Bronfenbrenner's bioecological model. *International Journal for Multidisciplinary Research.* https://doi.org/10.36948/ijfmr.2023.v05i06.9154

Kazima, M. (2007). Malawian students' meanings for probability vocabulary. *Educational Studies in Mathematics*, *64*, 169-189.

Keele, L. (2010). An overview of rbounds: An R package for Rosenbaum bounds sensitivity analysis with matched data (White paper, 1, 15). *Columbus, OH*.

Kirwan, L. (2015). Mathematics curriculum in Ireland: The influence of PISA on the development of project maths. *International Electronic Journal of Elementary Education*, *8*(2), 317-332.

Kogu, Z.N. (2017). Relationship between students' mathematics achievement and career aspirations in secondary schools of Kandara Sub-county, Kenya. [Mémoire de maîtrise, Université de Kenyatta].

Klenowski, V. (2014). Towards fairer assessment. *The Australian Educational Researcher, 41*(4), 445-470.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

Lane, F., To, Y., Shelley, K., & Henson, R. (2012). An illustrative example of propensity score matching with education research. *Career and technical education research*, *37*(3), 187–212. https://doi.org/10.5328/cter37.3.187

Larwin, K. H. (2010). Reading is fundamental in predicting math achievement in 10th graders? *International Electronic Journal of Mathematics Education*, *5*(3), 131-145.

Laurencelle, L. (2021). Le traitement statistique des proportions incluant l'analyse de variance, avec des exemples. The statistical handling of proportions including analysis of variance, with worked out examples. *Quantitative Methods Psychology,17*, 272-285.

Lauwerier, T., & Akkari, A. (2020). Modèles nationaux d'intégration de migrant· e· s et résultats PISA: éléments pour un débat. *Schweizerische Zeitschrift für Bildungswissenschaften*, *42*(1), 84-107.

Laveault, D., & Grégoire, J. (1997). Introduction aux théories des tests en sciences humaines. *Bruxelles: De Boeck Université*.

Lecerf, T. (2014). Influence de la culture et du milieu socioéconomique sur les performances cognitives. *Revue de Neuropsychologie*, *6*(3), 191-200.

Lee, M. K., & Randall, J. (2011). Exploring language as a source of DIF in a math test for English language learners. In *NERA Conference Proceedings 2011* (p. 20). https://opencommons.uconn.edu/nera_2011/20

Lee, J., & Stankov, L. (2018). Non-cognitive predictors of academic achievement: Evidence from TIMSS and PISA. *Learning and Individual Differences*, *65*, 50-64.

LeFevre, J. A., Fast, L., Skwarchuk, S. L., Smith-Chant, B. L., Bisanz, J., Kamawar, D., & Penner-Wilger, M. (2010). Pathways to mathematics: Longitudinal predictors of performance. *Child Development*, *81*(6), 1753-1767. https://doi.org/10.1111/j.1467-8624.2010.01508.x

Le Mener, M., Meuret, D., & Morlaix, S. (2017). L'accroissement de l'effet de l'origine sociale sur la performance scolaire: par où est-il passé? *Revue française de sociologie*, *58*(2), 207–231.

Lemke, J. L. (2003). Mathematics in the middle: Measure, picture, gesture, sign, and word. In M. Anderson, A. Sáenz Ludlow, S. Zellweger & V. V. Cifarelli (Eds.), *Educational Perspectives on Mathematics as Semiosis: From Thinking to Interpreting to Knowing* (pp. 215-234). Brooklyn, NY; Ottawa, Ontario: *Legas*

Little, R. J. A., & Rubin, D. B. (2019). Statistical analysis with missing data (3rd ed.). *Wiley*.

Liu, W., Kuramoto, S. J., & Stuart, E. A. (2013). An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention science*, *14*, 570-580

Liu, Y., Zumbo, B., Gustafson, P., Huang, Y., Kroc, E., & Wu, A. (2019). Investigating causal DIF via propensity score methods. *Practical Assessment, Research, and Evaluation*, *21*(1), 13.

Liu, R., & Bradley, K. D. (2021). Differential item functioning among English language learners on a large-scale mathematics assessment. *Frontiers in Psychology*, *12*, 657335.

Loignon, G. (2021a). ALSI: un nouvel outil d'analyse automatisée de la complexité linguistique pour le français québécois. *Mesure et évaluation en éducation*, *44*(3), 29-57. https://doi.org/10.7202/1093065ar

Loignon, G. (2021b). Une approche computationnelle de la complexité linguistique par le traitement automatique du langage naturel et l'oculométrie. [Thèse doctorale, Université de Montréal]. Papyrus https://hdl.handle.net/1866/26189

Lory, M. P. (2023). Les approches pédagogiques qui valorisent la diversité linguistique et culturelle pour repenser de façon inclusive la francophonie en contexte minoritaire. *Arborescences*, *(13)*, 36-48. https://doi.org/10.7202/1107653ar

Loughran, J. (2014). Understanding differential item functioning for English language learners: The influence of linguistic complexity features [Doctoral dissertation, University of Kansas]. *University of Kansas.*

Loye, N. (2018). Et si la validation était plus qu'une suite de procédures techniques? *Mesure et évaluation en éducation*, *41*(1), 97-123. https://doi.org/10.7202/1055898ar

Maekawa, M., Tanaka, A., Ogawa, M., & Roehrl, M. H. (2024). Propensity score matching as an effective strategy for biomarker cohort design and omics data analysis. *PLOS ONE, 19*(5), e0302109.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior research methods*, *42*(3), 847-862.

Magis, D., Boeck, P., & Raîche, G. (2011). Comparaison empirique des méthodes classiques de détection du fonctionnement différentiel d'items en psychométrie. *Des mécanismes pour assurer la validité de l'interprétation de la mesure en éducation : La mesure, 1,* 31–49. Presses de l'Université du Québec. https://doi.org/10.1515/9782760526860-003

Mahmud, N., Pazil, N. S. M., Rosli, N. S., & Jamaluddin, S. H. (2024). Factor that affects the students' performance in mathematics subject using logistic regression. *Journal of Computing Research and Innovation*, *9*(1), 121-130

Malabayabas, M. E., Torres, P. C., Sapin, S. B., & Tamban, V. E. (2022). Mathematics performance, academic well-being, and educational aspirations of junior high school students. *International Journal on Research in STEM Education*, *4*(2), 87-102.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*(4), 719-748.

Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review*, *78*(2), 333–368 https://doi.org/10.17763/haer.78.2.70783570r1111t32

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational Assessment*, *14*(3-4), 160-179.

Martinková, P., Drabinová, A., Liaw, Y. L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE—Life Sciences Education*, *16*(2), rm2. https://doi.org/10.1187/cbe.16-10-0307

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*(2), 127–143.

Michaelides, M. P., Brown, G. T., Eklöf, H., Papanastasiou, E. C. (2019). The relationship of motivation with achievement in Mathematics. In *Motivational Profiles in TIMSS Mathematics: Exploring Student Clusters Across Countries and Time*, (pp. 9–23). https://doi.org/10.1007/978-3-030-26183-2_2

Millsap, R. E. (2011). Statistical approaches to measurement invariance. *Routledge*. https://doi.org/10.4324/9780203821961

Ministère de l'Éducation du Québec (MEQ) (2002). Cadre de référence en évaluation des apprentissages. Gouvernement du Québec.

Ministère de l'Éducation du Québec (MEQ). (2003). Être évalué pour mieux apprendre : Politique d'évaluation des apprentissages (No. 03 00062, ISBN 2-550-41407-1). *Gouvernement du Québec*. https://www.education.gouv.qc.ca/fileadmin/site_web/documents/dpse/evaluation/13-4602.pdf

Ministère de l'Éducation du Québec (MEQ). (2019). Référentiel d'intervention en mathématique. https://www.education.gouv.qc.ca/fileadmin/site_web/documents/dpse/adaptation_serv_compl/Referentiel-mathematique.PDF

Ministère de l'Éducation du Québec (MEQ). (2024). Moyenne finale et taux de réussite pour l'ensemble des programmes comportant une ou des épreuves ministérielles en 4e et 5e année du secondaire**.** Récupéré 13 avril 2024. https://www.education.gouv.qc.ca/references/indicateurs-et-statistiques/indicateurs/moyenne-taux-de-reussite-programmes-epreuves-4-5-secondaire

Mou, H., & Atkinson, M. M. (2020). Want to improve math scores? An empirical assessment of recent policy interventions in Canada. *Canadian Public Policy*, *46*(1), 107-124.

Mourji, F. & Abbaia, A. (2013). Les déterminants du rendement scolaire en mathématiques chez les élèves de l'enseignement secondaire collégial au Maroc: Une analyse multiniveau. *Revue d'économie du développement*, 21, 127-158. https://doi.org/10.3917/edd.271.0127

Moschkovich, J. (2002). A situated and sociocultural perspective on bilingual mathematics learners. *Mathematical thinking and learning*, *4*(2-3), 189-212.

Moschkovich, J. (2007). Using two languages when learning mathematics. *Educational studies in Mathematics*, *64*(2), 121-144.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2017). TIMSS 2019 assessment framework. *TIMSS & PIRLS International Study Center, Boston College*. http://timssandpirls.bc.edu/timss2019/frameworks/framework-chapters/science-framework/

Muñiz, J., & Hambleton, R. K. (1996). Directrices para la traducción y adaptación de los tests. *Papeles del psicólogo*, *66*(1), 63-70. https://www.psychologistpapers.com/abstract?pii=737

Nilsen, T., & Teig, N. (2024). Analytical framework. In effective and equitable teacher practice in mathematics and science education: A nordic perspective across time and groups of students (pp. 35-56). *Cham: Springer Nature Switzerland.*

Niu, L. (2020). A review of the application of logistic regression in educational research: Common issues, implications, and suggestions. *Educational Review*, *72*(1), 41-67.

Njomgang-Ngansop, J., & Durand-Guerrier, V. (2011). Negation of mathematical statements in French in multilingual contexts-an example in Cameroon. In *ICMI Study 21 conference-Mathematics and Language Diversity*, 268-275.

O'Connell, A. A., & McCoach, D. B. (Eds.). (2008). *Multilevel modeling of educational data*. IAP.

Odo, D. M. (2018). A comparison of readability and understandability in second language acquisition textbooks for pre-service EFL teachers. *Journal of Asia TEFL*, *15*(3), 750.

O'Halloran, K. L. (2005). *Mathematical discourse: Language, symbolism and visual images*. London: Continuum.

Olmos, A., & Govindasamy, P. (2015). Propensity scores: A practical introduction using R. *Journal of MultiDisciplinary Evaluation*, *11*(25), 68-88.

Organisation for Economic Co-operation and Development (OCDE) (2010). Education at a glance 2010: OECD indicators. *Paris: OECD*.

Organisation de coopération et de développement économiques (OCDE) (2013). *Cadre d'évaluation et d'analyse du cycle PISA 2012: Compétences en mathématiques, en compréhension de l'écrit, en sciences, en résolution de problèmes et en matières financières,* Éditions OCDE. *http://dx.doi.org/10.1787/9789264190559-fr*

Organisation for Economic Co-operation and Development (OCDE) (2014). *PISA 2012 Technical Report*, PISA, OECD Publishing, Paris. https://doi.org/10.1787/6341a959-en.

Organisation de coopération et de développement économiques (OCDE) (2016). La performance des élèves, leur statut au regard de l'immigration et leurs attitudes à l'égard de la science, in *Résultats du PISA 2015 (Volume I): L'excellence et l'équité dans l'éducation*, OECD Publishing, Paris, https://doi.org/10.1787/9789264267534-11-fr.

Organisation de coopération et de développement économiques (OCDE) (2018). Résultats du PISA 2015 (Volume III): Le bien-être des élèves, *PISA, OECD Publishing,* Paris, https://doi.org/10.1787/9789264288850-fr.

Organisation for Economic Co-operation and Development (OCDE) (2019). PISA 2018 Results (Volume II): Where All Students Can Succeed. *Programme de l'OCDE pour l'évaluation internationale des élèves (PISA)* https://www.oecd.org/en/publications/pisa-2018-results-volume-ii_b5fd1b8f-en.html

Organisation de coopération et de développement économiques (OCDE) (2020). PISA 2018 results (Volume VI): Are students ready to thrive in an interconnected world? *OECD Publishing.* https://www.oecd-ilibrary.org/education/pisa-2018-results-volume-vi_d5f68679-en

Organisation de coopération et de développement économiques (OCDE) (2023a). *À la hauteur : Résultats canadiens de l'étude du PISA 2022 de l'OCDE. Le rendement des jeunes de 15 ans du Canada en mathématiques, en lecture et en sciences*. Paris, France : OCDE. https://www.cmec.ca/docs/pisa2022/PISA-2022_Highlights_FINAL_FR.pdf

Organisation de coopération et de développement économiques (OCDE) (2023b). Compétences en mathématiques (PISA) (indicateur). doi : 10.1787/6d1f74ca-fr (récupéré le 23 mars 2023) Évaluation internationale des élèves (PISA) - Compétences en mathématiques (PISA) - OCDE Data (oecd.org)

Parmar, R. S., Cawley, J. F., & Frazita, R. R. (1996). Word problem-solving by students with and without mild disabilities. *Exceptional children*, *62*(5), 415-429.

Pastor, D. A. (2003). The use of multilevel item response theory modeling in applied research: An illustration. *Applied measurement in education*, *16*(3), 223-243.

Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The journal of educational research*, *96*(1), 3-14.

Peng, P., Lin, X., Ünal, Z. E., Lee, K., Namkung, J., Chow, J., & Sales, A. (2020). Examining the mutual relations between language and mathematics: A meta-analysis. *Psychological Bulletin*, *146*(7), 595–623. http://doi.org/10.1037/bul0000231

Perez Mejias, P., McAllister, D. E., Diaz, K. G., & Ravest, J. (2021). A longitudinal study of the gender gap in mathematics achievement: Evidence from Chile. *Educational Studies in Mathematics*, *107*(3), 583-605

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, *3*(2), 237-255.

Persson, T. (2016). The language of science and readability: correlations between linguistic features in TIMSS science items and the performance of different groups of Swedish 8th grade students. *Nordic Journal of Literacy Research*, *2*(1). https://doi.org/10.17585/njlr.v2.186

Pichette, J., Kanters, D., & Ahmed, S. (2023). Exploration de la relation entre le rendement en mathématiques au secondaire et les cheminements en EPS à l'aide de l'ensemble des données du PRC. Conseil ontarien de la qualité de l'enseignement supérieur.

Pimm, D. (1987*).* Speaking mathematically: Communication in mathematics classrooms. London, UK: *Routledge & K. Paul*. https://doi-org/10.4324/9781315278858

Purpura, D. J., & Reid, E. E. (2016). Mathematics and language: Individual and group differences in mathematical language skills in young children. *Early Childhood Research Quarterly*, *36*, 259–268.

Purpura, D. J., Logan, J. A. R., Hassinger-Dasm, B., & Napoli, A. R. (2017). Why do early mathematics skills predict later reading? The role of mathematical language. *Developmental Psychology*, *53*(9), 1633–1642.

Rasch, G. 1960. *Probabilistic models for some intelligence and attainment tests*, Copenhagen, Denmark : The danish institute of education research (Expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago, IL: The University of Chicago Press.

Raudenbush, S.W., and Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (2nd ed.). *Newbury Park, CA: Sage.*

Ravand, H. (2015). Item response theory using Hierarchical generalized linear models. *Practical Assessment, Research and Evaluation, 20*, 7.

Reynolds, C. R., & Suzuki, L. A. (2013). Bias in psychological assessment: An empirical review and recommendations. In J. R. Graham, J. A. Naglieri, & I. B. Weiner (Eds.), *Handbook of psychology: Assessment psychology* (2nd ed., pp. 82–113). John Wiley & Sons, Inc.

Riccardi, D., Lightfoot, J., Lam, M., Lyon, K., Roberson, N. D., & Lolliot, S. (2020). Investigating the effects of reducing linguistic complexity on EAL student comprehension in first-year undergraduate assessments. *Journal of English for Academic Purposes*, *43*, 100804.

Ricardo, R. (2024). Définition, types et exemples de biais d'évaluation – Leçon. Publié le 19 janvier. Récupéré le 17 avril 2024, de https://enetudiant.com/definition-types-et-exemples-de-biais-devaluation-lecon/

Rocher, T. (2009). Nécessité et complémentarité des évaluations internationales et nationales. *Revue internationale d'éducation de Sèvres*. http://journals.openedition.org/ries/5650 ; DOI : https://doi.org/10.4000/ries.5650

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*(2), 105–116.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.

Rosenbaum, P. R. (2002a). Observational studies (2nd ed.). *Springer.* https://doi.org/10.1007/978-1-4757-3692-2

Rosenbaum, P. R., (2002b). Overt bias in observational studies (pp. 71-104). *Springer New York*. https://doi.org/10.1007/978-1-4757-3692-2_3

Rosenbaum, P. R. (2005). Sensitivity analysis in observational studies. *Encyclopedia of Statistics in Behavioral Science*, *4*, 1809-1814.

Rosenbaum, P. R. (2007). Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. *Biometrics, 63*(2), 456–464.

Rosenbaum, P. R. (2010). Design of observational studies, *Springer*, *10*, 1–384.

Rosenbaum, P. R. (2015). Two R packages for sensitivity analysis in observational studies. *Observational Studies*, *1*(2), 1-17.

Roth, W. M., Ercikan, K., Simon, M., & Fola, R. (2015). The assessment of mathematical literacy of linguistic minority students: Results of a multi-method investigation. *The Journal of Mathematical Behavior*, *40*, 88-105.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2*, 169–188.

Samo, D.D. (2021). Analysis of mathematical connections ability on junior high school students. *International Journal of Educational Management and Innovation*, *2*(3), 261-271.

Schleppegrell, M. J. (2007). The linguistic challenges of mathematics teaching and learning: A research review. *Reading & Writing Quarterly*, *23*, 139–159. https://doi.org/10.1080/10573560601158461

Sells, L. W. (1973). High school mathematics as the critical filter in the job market. Proceedings of the conference on minority graduate education. *Berkeley*

Simmt, E. (2015). What Is a Canadian mathematics education? A national discourse community. *Canadian Journal of Science, Mathematics and Technology Education*, *15*(4), 407–417. https://doi.org/10.1080/14926156.2015.1091902

Spelke, E. S., & Tsivkin, S. (2001). Language and number: A bilingual training study. *Cognition*, *78*(1), 45–88. https://doi.org/10.1016/S0010-0277(00)00108-6

Solano-Flores, G., & Li, M. (2008). Examining the dependability of academic achieve- ment measures for English language learners. *Assessment for Effective Interven- tion, 33,* 135–144.

Stetz, T.A. (2022). Introduction to tests and test bias. In: Test bias in employment selection testing. classroom companion: Business. *Springer, Cham*. https://doi.org/10.1007/978-3-030-89925-7_1

Stone, J. R., Alfeld, C., Pearson, D., Lewis, M. V. and Jensen, S. (2006). Building academic skills in context: Testing the value of enhanced Math learning in CTE. University of Minnesota: National Research Center for Career and Technical Education.

Stokke, A. (2015). What to do about Canada's declining Math scores? *Labor: Human Capital eJournal*. SSRN: https://ssrn.com/abstract=2613146 or http://dx.doi.org/10.2139/ssrn.2613146

Stuart, E. A., & Rubin, D. B. (2008). Best practices in quasi-experimental designs. Best practices in quantitative methods, 155–176. *SAGE Publications*.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*(1), 1-21.

Stuart, E. A., King, G., Imai, K., & Ho, D. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of statistical software*. https://doi.org/10.18637/jss.v042.i08

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the p value is not enough. *Journal of Graduate Medical Education, 4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

Sulis, I., & Porcu, M. (2017). Detecting differential item functioning in large-scale assessments: Effects of missing data and booklet design. *Journal of Educational Measurement, 54*(1), 1–22.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, *27*(4), 361-370.

Törmäkangas, K. (2011). Advantages of the Rasch measurement model in analysing educational tests: An applicator's reflection. *Educational Research and Evaluation*, *17*(5), 307-320.

Van Der Linden, W. J., & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. *Handbook of modern item response theory* : *Springer New York*, 1-28. https://doi.org/10.1007/978-1-4757-2691-6_1

Van Rinsveld, A., Schiltz, C., Brunner, M., Landerl, K., & Ugen, S. (2016). Solving arithmetic problems in first and second language: Does the language context matter? *Learning and Instruction*, 42, 72–82. https://doi.org/10.1016/j.learninstruc.2016.01.003

Vukovic, R. K., & Lesaux, N. K. (2013). The relationship between linguistic skills and arithmetic knowledge. *Learning and Individual Differences*, *23*, 87–91. https://doi.org/10.1016/j.lindif.2012.10.007

Wang, Y., Cai, H., Li, C., Jiang, Z., Wang, L., Song, J., & Xia, J. (2013). Optimal caliper width for propensity score matching of three treatment groups: A Monte Carlo study. *PLoS one*, *8*(12), e81045. https://doi.org/10.1371/journal.pone.0081045

Wang, A. Y., Fuchs, L. S., & Fuchs, D. (2016). Cognitive and linguistic predictors of mathematical word problems with and without irrelevant information. *Learning and individual differences*, *52*, 79-87.

Wang, X. S., Perry, L. B., Malpique, A., & Ide, T. (2023a). Factors predicting mathematics achievement in PISA: A systematic review. Large-scale Assessments in Education, 11(1), 24. https://doi.org/10.1186/s40536-023-00174-8

Wang, C., Zhu, R., & Xu, G. (2023b). Using lasso and adaptive lasso to identify DIF in multidimensional 2PL models. *Multivariate Behavioral Research*, *58*(2), 387-407. https://doi.org/10.1080/00273171.2021.1985950

Wang, J. S., & Aronow, P. M. (2023). Listwise deletion in high dimensions. *Political Analysis*, *31*(1), 149-155. https://doi.org/10.1017/pan.2022.5

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p < 0.05". *The American Statistician, 73*(Suppl. 1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Wolf, M. K., & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. *Educational Assessment,* 14(3–4), 139–159. https://doi.org/10.1080/10627190903425883

Wurtz, K. (2008). A methodology for generating placement rules that utilizes logistic regression. *Journal of Applied Research in the Community College*, *16*(1), 51-57.

Young, R. (2009). Discursive practice in language learning and teaching (Vol. 58). *Malden, MA: Wiley-Blackwell.*

Zamora-Araya, J. A., Smith-Castro, V., Montero-Rojas, E., & Moreira-Mora, T. E. (2018). Advantages of the Rasch model for analysis and interpretation of attitudes: The case of the benevolent sexism subscale. *Revista Evaluar*, *18*(3). https://doi.org/10.35670/1667-4545.v25.n1

Zhang, Z., Kim, H. J., Lonjon, G., & Zhu, Y. (2019). Balance diagnostics after propensity score matching. *Annals of translational medicine*, *7*(1). https://doi.org/10.21037/atm.2018.12.10

Zhang, Y. (2021). The effect of science social capital on student science performance: Evidence from PISA 2015. *International Journal of Educational Research*, *109*, 101840. https://doi.org/10.1016/j.ijer.2021.101840

Zhao, Q. Y., Luo, J. C., Su, Y., Zhang, Y. J., Tu, G. W., & Luo, Z. (2021). Propensity score matching with R: conventional methods and new features. *Annals of translational medicine*, *9*(9). https://doi.org/10.21037/atm-20-3998

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4(2), 223-233. https://doi.org/10.1080/15434300701375832

Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Olvera Astivia, O. L., & Ark, T. K. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly*, *12*(1), 136-151. https://doi.org/10.1080/15434303.2014.972559