Journées de méthodologie statistique de l'Insee 2022

ÉTUDE DE L'EFFET DU MODE DE PASSATION D'ÉVALUATIONS STANDARDISÉES SUR LA PERFORMANCE DES ÉLÈVES :

L'EXEMPLE DE CEDRE MATHÉMATIQUES COLLÈGE

Reinaldo DOS SANTOS (*), Louis-Marie NINNIN (*), Vincent PAILLET (*), Franck SALLES (*)

(*) Ministère de l'Éducation Nationale, DEPP

reinaldodos@gmail.com

Mots-clés *(6 maximum)* : Modèles mixtes, multimode, psychométrie, séries temporelles

Domaine concerné : 4.4 Effets de mode ; 19.1 Évaluation des élèves

Résumé

La Direction de l'Évaluation, de la Prospective et de la Performance (DEPP) du Ministère de l'Éducation Nationale s'est engagée dans la transition d'enquêtes réalisées sur papier à des enquêtes au format numérique. Cette transition pose certaines questions d'ordre méthodologique, notamment en termes de comparabilité entre les cycles des séries temporelles disciplinaires (enquêtes CEDRE).

Afin de pouvoir observer l'impact du changement de mode d'attribution des évaluations, il faut pouvoir quantifier l'effet du mode sur l'item. Une fois cet « effet mode » pris en compte, il faut s'assurer que la performance des élèves demeure identique quel que soit le mode d'attribution. La théorie psychométrique permet de décrire les caractéristiques des items et la performance des élèves à une évaluation. Pour ce faire, il est nécessaire de créer une échelle abstraite, déterminée par la distribution de la performance des élèves.

Nous avons construit une évaluation composée d'items dans leur format papier et leur format numérique. Celle-ci a été présentée à un échantillon représentatif d'élèves de 3ème. A l'aide d'un plan en blocs incomplet équilibré, nous avons garanti qu'un élève ne croise pas deux fois le même item. L'utilisation de ce design expérimental a permis de calculer, pour chaque élève de notre étude, un score obtenu sur l'enquête « papier crayon » et un score obtenu sur l'enquête au format numérique. L'utilisation des modèles mixtes multi-niveaux permettent alors de décomposer les scores des élèves entre effets aléatoires à chacun des niveaux (intra-individuel, intra-établissement, intra-académique) et effets fixes, dont notamment le mode d'attribution de l'évaluation.

Cette étude démontre que la transition numérique ne remet pas en cause la qualité d'une série longitudinale en termes de comparaison dans le temps, ni en termes d'égalité de traitement entre les différents groupes d'élèves. Toutefois, il n'est pas possible de considérer que les items demeurent invariants quand on les adapte au format numérique. Il est donc nécessaire de maintenir un ancrage « papier crayon » avant de basculer l'ensemble de l'évaluation au format numérique.

Abstract

This study demonstrates that the switch to a digital format for mathematics surveys intended for 3rd grade students does not jeopardise the quality of a longitudinal series in terms of comparison over time, nor in terms of equality of treatment between the different groups of students. However, it is not possible to consider the items repeated from one assessment cycle to another as invariant when adapting them to the digital format. It is therefore necessary to first maintain a mixed mode of assessment (paper and digital) before switching the whole assessment to digital format.

1 Introduction

1.1 La DEPP au cœur de la transition vers des évaluations numériques

Le Cycle des Évaluations Disciplinaires Réalisées sur Échantillon (CEDRE) établit un bilan des acquis des élèves en fin d'école et fin de collège. Il couvre les compétences des élèves dans la plupart des domaines disciplinaires en référence aux programmes scolaires. Renouvelées tous les six ans (cinq ans depuis 2012), ces évaluations permettent de répondre à la question de l'évolution du niveau des élèves dans le temps. En faisant office de séries longitudinales, elles permettent de fournir une photographie du niveau général des élèves, et ainsi de disposer d'un suivi dans le temps des performances disciplinaires des élèves. La comparabilité du contenu de chaque nouveau cycle avec l'historique de la série est donc essentielle pour assurer la stabilité du construit observé[1].

De manière générale, la comparabilité du cycle N avec les cycles précédents de la série est assurée par l'utilisation d'un test d'ancrage, combiné avec de nouveaux items. Le test d'ancrage consiste en la reprise à l'identique d'items déjà utilisés lors des cycles précédents[2]. Son utilisation permet de placer sur la même échelle les performances des élèves de chaque cycle. En effet, la présence d'items communs avec les cycles précédents garantit à la fois la validité du construit observé et la fiabilité de la mesure réalisée. À l'opposé, l'injection de nouveaux items répond à un triple objectif : permettre tout d'abord d'éviter que le contenu du test soit connu par avance, et par conséquent le bachotage ou « teaching for the test »[3] ; permettre aussi de remplacer les items utilisés précédemment et ayant démontré une faible qualité ; et enfin, permettre de couvrir les évolutions potentielles des programmes disciplinaires, en mesurant de nouvelles compétences.

Cette pratique de combinaison d'items propres à un cycle de la série et d'items d'ancrage repris d'un cycle à l'autre est une méthode robuste d'alignement des scores de chaque population sur une échelle commune. Elle nécessite toutefois que certains critères soient réunis. Il est ainsi important que le test d'ancrage soit une reproduction fidèle du test présenté lors du cycle précédent[4]. Il peut s'agir d'un test strictement équivalent en termes de difficulté moyenne et de dispersion (minitest) ou simplement équivalent en termes de difficulté (miditest). Mais il est surtout fondamental que les caractéristiques psychométriques des items d'ancrage ne varient pas d'un cycle à l'autre[5]. Un item qui serait perçu comme plus difficile par une cohorte d'élèves pénaliserait le calcul du score de cette population, et par conséquent briserait l'équité de la série temporelle.

1.2 Comment maintenir une série temporelle en changeant de mode de passation ?

L'ancrage doit être envisagé avec un regard nouveau lorsque le mode de passation change entre deux cycles. En effet, jusqu'ici les évaluations CEDRE étaient réalisées au format papier-crayon. Mais celles-ci sont, depuis 2017, réalisées sur format numérique. Ce nouveau format varie en fonction du niveau étudié. Au collège, les évaluations sont réalisées en ligne, à l'aide d'un ordinateur d'un clavier et d'une souris. À l'école, elles sont réalisées sur support mobile, à l'aide de tablettes. Par conséquent, les items d'ancrage sont certes repris d'un cycle sur l'autre, mais on ne peut considérer qu'ils sont repris à l'identique, puisque le support sur lequel ils sont présentés à l'élève n'est pas le même.

Ce changement de mode de passation au cours d'une série temporelle peut être appréhendé de différentes façons, en fonction des hypothèses retenues. On peut tout d'abord considérer que les caractéristiques des items ne dépendent pas du mode de passation du test. Dans ce cas, les items du test d'ancrage peuvent être considérés comme invariants entre les cycles de l'évaluation. On peut également considérer que si la difficulté des items varie lors du changement de mode, cette variation est constante, et indépendante de caractéristiques secondaires (domaine didactique, format de réponse…) autre que le mode. Dans ce cas, les items communs aux cycles peuvent servir d'items d'ancrage en appliquant une transformation linéaire équivalente à « l'effet mode » mesuré. On peut autrement considérer que les items perdent leurs caractéristiques psychométriques lors du changement de mode de passation. Dans ce cas, il est nécessaire d'assurer l'ancrage à l'aide d'items repris du cycle précédent au format papier. Enfin, on peut également considérer qu'au-delà des paramètres des items, les performances des élèves ne sont pas indépendantes du mode de passation de l'évaluation, notamment au niveau des sous-populations (par exemple le genre). Dans ce cas, l'équité de la série temporelle n'est plus assurée[6].

1.3 Étude de cas : CEDRE mathématiques collège

Il est donc essentiel que l'impact de la transition des évaluations CEDRE d'un format papier-crayon vers un format numérique puisse être quantifié. Cet impact doit être mesuré, pour chaque couple élève-item, sur l'ensemble de l'espace composé par l'écart entre la difficulté de l'item et le niveau de performance de l'élève. Pour cela, il faut décomposer l'effet du mode en deux parties. Premièrement, l'effet du changement de mode sur l'item, c'est-à-dire la variation de difficulté entre sa version papier et sa version numérique. Puis l'effet du changement de mode sur la performance des élèves, c'est-à-dire l'écart de performance une fois pris en compte l'effet mode sur les items.

La première question de recherche concerne les items. Quel est l'impact de la transition numérique sur leurs paramètres psychométriques, et plus particulièrement sur leur difficulté ? Cet effet mode est-il constant, au sens qu'il ne dépend pas de la difficulté initiale de l'item ? Cet effet mode est-il indépendant, au sens qu'il ne varie pas en fonction de caractéristiques exogènes, telles que le domaine de compétences au sein de la discipline, ou encore le format papier de la réponse (QCM, tableau-série, question ouverte, etc.) ? La deuxième question de recherche concerne la performance des élèves. Celle-ci est-elle indépendante du mode de passation de l'évaluation ? Autrement dit, les élèves ont-ils globalement le même niveau de performance pour une évaluation papier et pour une évaluation numérique ? Cette équité de performance est-elle avérée également pour les sous-populations principales de notre échantillon ?

Cette étude, réalisée sur l'enquête CEDRE de mathématiques au collège, démontre que les items sont perçus par les élèves comme étant plus difficiles au format numérique qu'au format papier-crayon. Cet effet mode ne dépend pas du domaine de compétences auquel l'item appartient. En revanche, il varie en fonction du format de l'item. Une fois cet effet mode pris en compte, les performances des élèves sont équivalentes lorsqu'on les mesure sur le format papier-crayon ou numérique. Cette équivalence se retrouve également au niveau du retard éventuel de l'élève (redoublement), de son Indice de Position Sociale[7] ainsi que de la strate de l'établissement de scolarisation (public hors éducation prioritaire, éducation prioritaire, privé). Par contre, on constate que le passage au numérique favorise légèrement les performances des filles par rapport aux garçons.

Dans un premier temps, nous décrirons la méthodologie employée pour réaliser ces enquêtes. Les données, le plan de design et les techniques de collecte seront abordés, ainsi que les modèles utilisés pour analyser ces données. Dans un deuxième temps, les résultats obtenus seront présentés dans le détail. On distinguera l'étude de l'effet mode sur les items, puis l'étude de la performance comparée des élèves selon le mode. Enfin, les résultats seront commentés. Il s'agira de discuter de leur impact sur de futures recherches, des limitations de l'étude, ainsi que des recommandations qui peuvent être apportées au design des prochaines enquêtes CEDRE.

2 Méthodologie

2.1 Spécifications de l'enquête

2.1.1 Construit étudié

Le cadre d'évaluation de CEDRE mathématiques 2019 en fin de collège est organisé autour de deux dimensions :

- La dimension des contenus (savoirs), qui précise les domaines mathématiques évalués, en lien avec les programmes du cycle 4
- La dimension cognitive, qui précise les différents niveaux d'activité mathématique des élèves relatifs aux domaines mathématiques qui figurent dans les programmes en fin de cycle 4

Les domaines mathématiques évalués dans CEDRE en fin de collège correspondent aux thèmes et sous thèmes qui structurent le programme officiel de mathématiques de cycle 4 :

- Nombres et calculs (thème A)
  - Utiliser les nombres pour comparer, calculer et résoudre des problèmes (A1)
  - Utiliser le calcul littéral (A3),
- Organisation et gestion de données, fonctions (thème B)
  - Interpréter, représenter et traiter des données (B1)
  - Comprendre et utiliser des notions élémentaires de probabilités (B2)
  - Résoudre des problèmes de proportionnalité (B3)
  - Comprendre et utiliser la notion de fonction (B4)
- Grandeurs et mesures (thème C)
  - Calculer avec des grandeurs mesurables ; exprimer les résultats dans les unités adaptées (C1)
- Espace et géométrie (thème D)
  - Représenter l'espace (D1)
  - Utiliser les notions de géométrie plane pour démontrer (D2)
- Algorithme et programmation (thème E).
  - Écrire, mettre au point, exécuter un programme (E1).

Il convient de noter que le domaine « Algorithme et programmation » proposé en 2019, étant nouvellement inclus dans les programmes en 2015, n'était pas présent dans les cycles précédents (2008 et 2014). Nous n'avons donc que sa version numérique.

2.1.2 Échantillonnage

La méthode d'échantillonnage pour le collège est un tirage stratifié par grappes[8]. Ainsi, un certain nombre de classes sont tirées, et tous les élèves de ces classes sont sélectionnés. De plus, le tirage est équilibré selon certaines variables auxiliaires, permettant d'obtenir un échantillon représentatif de la population, dans son ensemble et au regard desdites variables auxiliaires. Celles-ci sont choisies comme étant liées à la performance des élèves : le genre de l'élève, son retard (un redoublement a-t-il eu lieu dans sa scolarité ?), et la strate de l'établissement de scolarisation (public ou privé, éducation prioritaire ou non).

Comme toute enquête réalisée par sondage, les évaluations des élèves sont exposées à la non-réponse. Bien que les taux de retour soient élevés, il est nécessaire de tenir compte de la non-réponse dans les estimations car celle-ci n'est pas purement aléatoire. Afin de la prendre en compte, un calage sur marges[9] est effectué. Cette méthode consiste à modifier les poids de sondage des répondants de manière à ce que l'échantillon ainsi repondéré soit représentatif de certaines variables

auxiliaires dont on connaît les totaux sur la population. C'est une méthode qui permet de corriger la non-réponse mais également de rendre cohérents les résultats observés sur l'échantillon pour ce qui concerne les informations connues sur l'ensemble de la population.

2.1.3 Design de l'enquête

Afin de pouvoir utiliser un nombre important d'items sans que la durée de la passation n'en devienne irréaliste, les évaluations CEDRE utilisent traditionnellement la méthode dite des cahiers tournants, qui est un design de la famille des plans en blocs. Un plan garantissant la même exposition à chacun des items est appelé un plan équilibré. Un plan où chaque bloc est présent dans tous les cahiers est appelé un plan complet. Les cahiers tournants sont donc un exemple de plan en blocs équilibré et incomplet ou BIBD (Balanced Incomplete Block Design)[10]. Dans notre cas, nous rajouterons comme contrainte qu'aucun cahier ne doit contenir un item sous ses deux formes. En effet, présenter deux fois le même item à un élève, quand bien même le format changerait, ne protège pas celui-ci de l'effet d'entraînement.

Au regard de notre banque d'items d'ancrage, nous nous retrouvons donc avec un plan composé de deux cahiers composés de deux blocs d'items, un au format papier et un au format numérique. Soit :

- Un cahier A, composé du bloc 1 au format papier et du bloc 2 au format numérique
- Un cahier B, composé du bloc 1 au format numérique et du bloc 2 au format papier

Malheureusement, les contraintes supplémentaires sur les items engendrent une situation où nous avons deux sous-échantillons qui ne possèdent aucun item commun. Afin d'assurer l'équilibrage du plan de design, mais également afin de réduire le biais d'échantillonnage, l'allocation du cahier A ou B à chaque élève sera réalisé de manière aléatoire dans chaque classe. Le fait de tirer deux sous-échantillons dans chaque classe entraîne un biais de sondage supplémentaire, lié à la variabilité intra-classe. En effet, il nous faut nous assurer que pour chaque classe, les deux sous-groupes soient également représentatifs de la population. C'est pourquoi la consigne stricte est d'assigner aléatoirement les élèves à l'un ou l'autre des deux groupes. Cette méthode présente deux avantages. Le tirage aléatoire simple minimise le biais de sélection pour chacun des sous-groupes, qui ne dépend plus que de la taille de chaque sous-groupe. De plus, il limite les effets aléatoires aux individus, rendant ainsi fixes les effets portés par les variables de niveau supérieur (classe, établissement, …) [11].

2.2 Méthodologie

2.2.1 Estimation des paramètres d'items à l'aide de modèles bifactoriels

L'analyse des données s'appuie sur la théorie psychométrique, et plus particulièrement sur les modèles de réponse à l'item ou Item Response Theory (IRT)[12]. Celle-ci repose sur l'idée que la probabilité d'une réponse à un item est une fonction mathématique des paramètres de la personne et de l'item. Le paramètre de la personne est interprété comme un trait ou une dimension latente unique. Il s'agit, ici, de la performance générale de l'élève dans la discipline observée. Les paramètres sur lesquels les items sont caractérisés comprennent leur difficulté, ainsi que leur discrimination, représentant la pente à laquelle le taux de réussite des individus varie en fonction de leur capacité.

Selon le modèle à deux paramètres dit de Birnbaum[13], la probabilité de réussite de l'élève à l'item X suit une fonction ogive normale de type :

$$P(X = 1 \mid \Theta) = \Phi(a^* (\Theta - b))$$

avec

- Θ le niveau de performance de l'élève sur le trait étudié
- b la difficulté de l'item X
- a la discrimination de l'item X

Le modèle décrit ci-dessus induit le principe d'unidimensionnalité : la performance de l'élève ne dépend que d'un seul trait cognitif. Afin de contrôler l'éventuelle multidimensionnalité générée par le

caractère hybride des modes de passation, on utilisera un modèle bifactoriel. Les modèles bifactoriels[14] sont des modèles multidimensionnels imbriqués. On considère que la performance d'un élève à un jeu d'items dépend de plusieurs traits latents, mais que ceux-ci sont fortement corrélés. On peut alors décomposer la variance portée par l'item selon deux facteurs. Le premier, appelé facteur général, est commun à l'ensemble des traits latents présents dans le test. Le second, appelé facteur spécifique, contient la variance propre au trait porté par l'item. Ces modèles permettent de « nettoyer » la dimension principale du bruit potentiel engendré par le caractère composite de l'évaluation. Les facteurs spécifiques, en agissant comme des variables de contrôle dans la fonction logistique de Birnbaum, permettent de « piéger » la variance résiduelle qui n'est pas attribuable au trait latent que nous cherchons à estimer, représenté par le facteur général[15].

De plus, le caractère incomplet du plan de design génère par construction deux sous-échantillons sans item commun, et donc deux échelles de score non comparable a priori. Afin de pouvoir constituer une échelle unique, nous utilisons la méthode de calibration concurrente[5]. Soit trois cohortes d'élèves d'intérêt : la cohorte précédente (CEDRE 2014), qui n'a été évaluée que sur papier. La cohorte d'intérêt, qui s'est vue présenter notre évaluation hybride. Et la cohorte suivante (CEDRE 2019), qui a été interrogée exclusivement sur format numérique. La calibration concurrente est une méthode de calage qui consiste à considérer que les paramètres des items communs à plusieurs cohortes sont invariants, et donc identiques pour chaque groupe. Parmi les nombreuses méthodes de calibration existantes, la calibration concurrente est considérée comme particulièrement robuste[16]. Elle permet ainsi de déterminer des paramètres d'item, et par extension une échelle de score commune aux différentes cohortes, dont les deux sous-échantillons de la « bridge study ».

2.2.2 Estimation des scores individuel des élèves

Une fois les paramètres d'items calculés, nous allons chercher à évaluer la performance de chaque élève pour chacune des modalités de passation.

Tout d'abord, nous allons réduire la performance des élèves au facteur général du modèle bifactoriel. En effet, et bien que l'étude complète du construit peut présenter un intérêt certain, notre objectif est de s'assurer que la performance générale de l'élève en mathématiques n'est pas sujette à variation, une fois l'effet mode pris en compte. Les facteurs spécifiques du modèle bifactoriel n'ont ici d'autre fonction que celle de capter la variance résiduelle des données.

Dans le cadre de la certification CEDRE, le score des élèves sur le facteur général est estimé à l'aide d'un estimateur dit de Warm ou Weighted Maximum Likelihood Estimator (WMLE)[17]. Toutefois, cet estimateur s'appuie sur l'hypothèse d'une distribution asymptotiquement normale de la population scorée. Si cette hypothèse est acceptable lorsqu'on score l'ensemble de l'espace généré par la matrice personnes-items, elle est plus contestable quand on réduit cet espace à une sous-population (uniquement la cohorte « bridge »), à un sous-test (uniquement les items papier de l'évaluation hybride) et à une seule dimension (le facteur général). Le Maximum A Posteriori (MAP) est un estimateur bayésien très proche de l'estimateur par maximum de vraisemblance (MLE), mais en diffère car il ne nécessite pas d'hypothèse de distribution normale. À la place, il autorise la possibilité de définir une distribution a priori des scores, et ainsi de « pondérer » la fonction de vraisemblance selon l'échelle de score.[18]

La méthode du Maximum A Posteriori consiste à trouver la valeur de Θ qui maximise la grandeur $L(\theta)*p(\theta)$, où $L(\theta)$ est la vraisemblance et $p(\theta)$ la distribution a priori des paramètres θ.

Ainsi, l'estimateur du maximum de vraisemblance est l'estimateur MAP pour une distribution a priori uniforme.

2.2.3 Modèles mixtes multi-niveaux sur les scores des élèves

Une fois obtenu pour chaque élève un score « papier » et un score « numérique », le score de l'élève va être utilisé comme variable d'intérêt dans un modèle mixte. Il s'agit d'une famille de modèles statistiques contenant à la fois des effets fixes et des effets aléatoires[19]. Ces modèles sont particulièrement utiles dans les cas où des mesures répétées sont effectuées sur les mêmes unités statistiques (étude longitudinale), ainsi qu'en présence de valeurs manquantes. Dans notre cas, les

scores des élèves vont être traités comme une série longitudinale, le format (papier ou numérique) associé au score étant traité comme la variable temporelle. En attribuant un effet aléatoire à l'élève, on peut ainsi distinguer l'effet fixe du format de l'effet aléatoire de l'élève, soit l'inégalité de performance entre élèves.

Afin de prendre en compte la complexité du plan de sondage, les variables de contexte (sexe, retard, strate de l'établissement…) seront ajoutées au modèle comme effets fixes. Toutefois, le caractère stratifié du modèle n'est pas pris en compte à ce stade. De plus, il est reconnu que les différentes strates du système scolaire français (établissement, circonscription, département, académie…) sont des facteurs déterminants de la performance des élèves. Les modèles à plusieurs niveaux[20] - également appelés modèles linéaires hiérarchiques, modèles à données imbriquées ou modèles à parcelles divisées - sont des modèles statistiques de paramètres qui varient à plus d'un niveau. Ils sont particulièrement adaptés aux plans de recherche où les données relatives aux participants sont organisées à plus d'un niveau. Les unités d'analyse sont généralement des individus (à un niveau inférieur) qui sont emboîtés dans des unités contextuelles/agrégées (à un niveau supérieur). Si le niveau de données le plus bas dans les modèles à plusieurs niveaux est généralement un individu, des mesures répétées d'individus peuvent également être examinées.

3 Résultats

3.1 Effet mode par item

En accord avec la validité attendue, nous avons choisi de réaliser un modèle bifactoriel à partir des 9 sous-dimensions croisant le mode de passation et les domaines de mathématiques , à savoir :

- Nombres et calculs
  - Passation « papier-crayon »
  - Passation numérique
- Organisation et gestion de données, fonctions
  - Passation « papier-crayon »
  - Passation numérique
- Grandeurs et mesures
  - Passation « papier-crayon »
  - Passation numérique
- Espace et géométrie
  - Passation « papier-crayon »
  - Passation numérique
- Algorithme et programmation
  - Passation numérique

Afin de déterminer l'effet porté par les items, nous avons fait le choix de nous concentrer sur le facteur général du modèle bifactoriel que nous avons construit. En effet, ce facteur est censé représenter l'information commune aux différents domaines mathématiques présents dans le test, que ce soit au format numérique ou « papier-crayon ».

Dans un premier temps, nous allons comparer, pour chaque item, la contribution au facteur général selon le mode de passation de l'item. La contribution (ou loading) d'un facteur à un item représente l'amplitude de l'influence du facteur sur un item. En d'autres termes, on peut l'utiliser comme indice de la qualité de l'item à décrire le facteur[21].

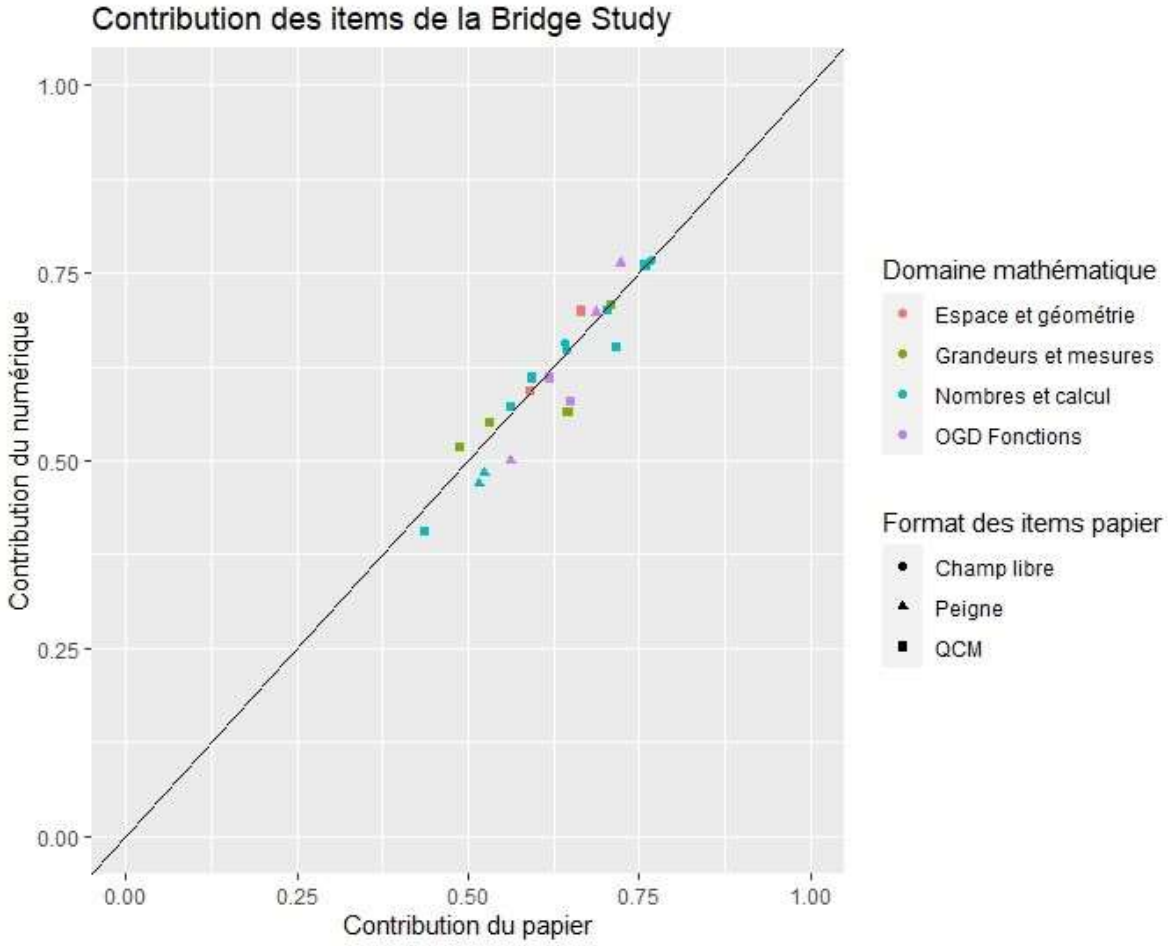
Dans un second temps, nous allons nous concentrer sur le paramètre de difficulté de chaque item, en fonction du mode de passation de l'item. Dans le cadre d'un modèle de réponse à l'item, la probabilité pour un élève i de répondre correctement à l'item j suit la fonction de lien suivante :

$$P_i(j = 1 \mid \Theta_i) = \frac{1}{e^{-1,7 * a_j * (\Theta_i - b_j)}}$$

Avec

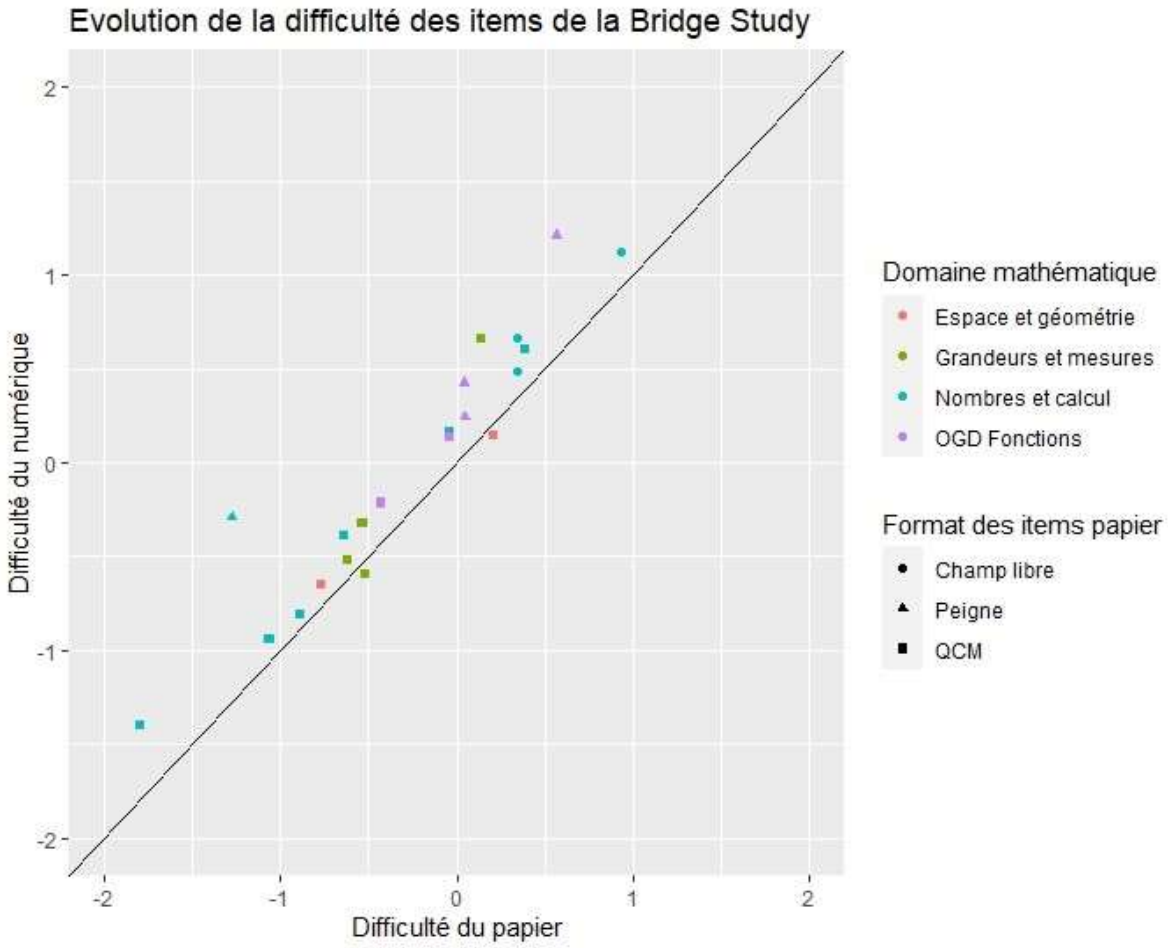
- Θ_i le niveau de l'élève i
- a_j la discrimination de l'item j
- b_j la difficulté de l'item j.

En ce qui concerne les contributions comparées de chaque mode de passation, on peut les représenter sur la figure suivante:



Nous constatons donc une forte stabilité des contributions des items entre leurs deux modes de passation. Néanmoins, on peut distinguer deux catégories d'items : ceux dont la contribution est d'une stabilité remarquable et ceux qui voient leur contribution légèrement diminuer lors du passage au mode numérique.

En ce qui concerne l'effet sur la difficulté des items, on peut les représenter sur la figure suivante. Les caractéristiques du graphique précédent sont reprises à l'identique.



Nous constatons, de manière générale, une élévation moyenne de la difficulté des items lors du passage au format numérique. Si cette augmentation est relativement homogène, on distingue deux items qui semblent subir un effet mode plus important que les autres. Ces deux items sont au format « peigne » » en papier, on peut supposer que ce format se transpose plus difficilement en numérique que les QCM.

Le format « peigne » est un format (papier) consistant en un champ pré-casé, permettant d'inscrire un caractère par case. Sa transposition au format numérique n'est pas évidente, et le choix a été fait de le transposer par un champ libre, mais avec limitation du nombre de caractères autorisé. Les formats ne sont pas en eux-mêmes particulièrement plus faciles ou difficiles que d'autres formats, mais la « distance » entre le « peigne » et le champ libre avec limitation du nombre de caractères est plus forte qu'entre un QCM papier et un QCM numérique (par exemple). On peut donc interpréter l'effet mode supérieur par une distance de transposition plus importante.

3.2 Effet mode sur la population

Afin de modéliser l'effet mode sur la population, les données ont été réduites à la cohorte concernée (2018), et le score des élèves a été décrit selon le modèle multi-niveaux suivant :

*theta ~ 1 + (IPS + Strate + Retard + Sexe + DOM) * Mode + (1 | taker) + (1 | UAI)*

Les effets fixes conservés dans le modèle correspondent aux variables explicatives habituelles, à savoir le sexe de l'élève, l'existence d'un redoublement dans son parcours, la strate de l'établissement de scolarisation, le lieu de scolarisation (en France métropolitaine ou dans les DROM) et l'indice de position sociale (IPS) des élèves. L'IPS est un outil de mesure quantitatif de la situation sociale des élèves face aux apprentissages dans les établissements scolaires français. Cet indice est construit à partir des professions et catégories socioprofessionnelles (PCS) des représentants légaux des élèves. Plus l'indice est élevé, plus l'élève évolue dans un contexte familial favorable aux apprentissages[7].

À ces variables usuelles s'ajoute la variable « Mode », qui correspond au mode de passation de l'évaluation. Enfin, afin de détecter de potentiels effets mode sur des sous-populations, les interactions entre la variable support et les autres variables explicatives ont été incorporées.

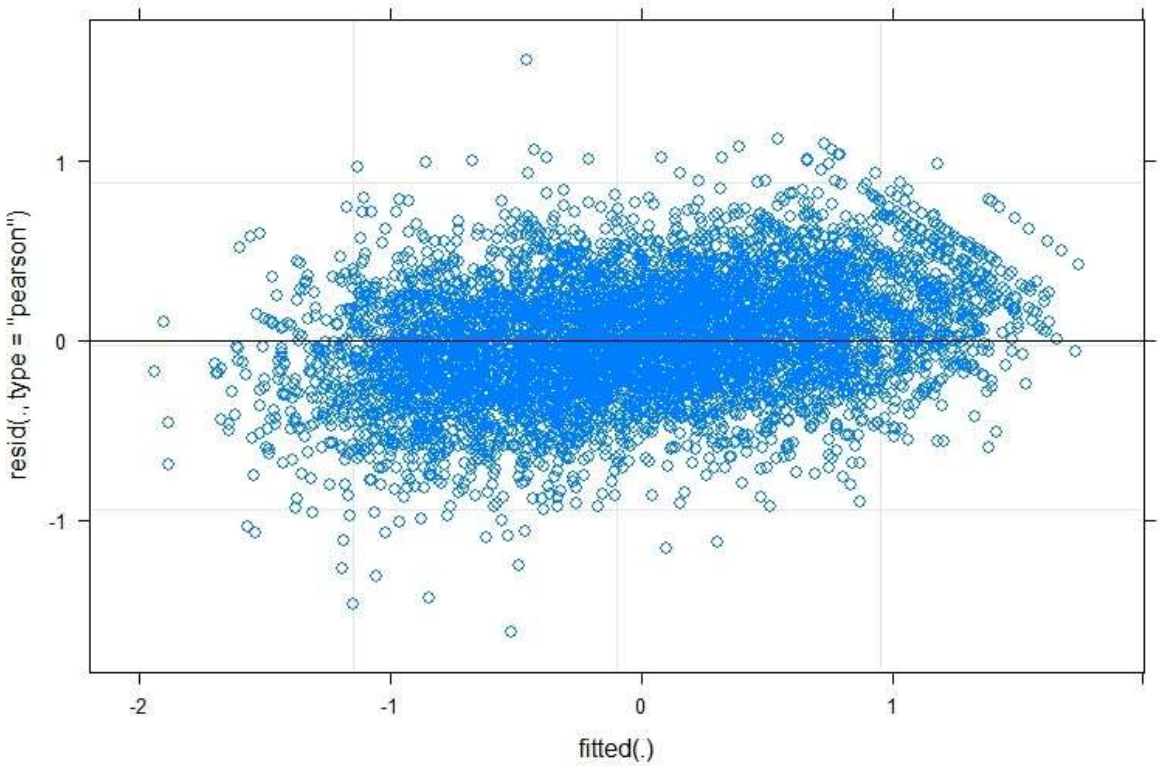
En ce qui concerne les effets aléatoires, nous avons conservé: le niveau individu (taker) tout d'abord, afin d'isoler l'effet mode des variations intra-individuelles ; l'effet établissement (UAI), afin de capter les variations inter-établissements ; enfin, le niveau académique a été dans un premier temps conservé, puis écarté une fois sa faiblesse constatée. Les effets de pentes aléatoires ont été également écartés, pour des raisons soit de non-significativité, soit de singularité.

Pour estimer la qualité du modèle, trois indicateurs sont disponibles et informatifs. Le coefficient de corrélation intra-classe indique la pertinence des regroupements réalisés dans le cadre d'un modèle hiérarchique. Le R^2 marginal indique la part de variance expliquée par les effets fixes[22]. Enfin, le R^2 conditionnel indique la part de variance expliquée par les effets fixes et les effets aléatoires.

Les résultats sont les suivants :

	Modèle complet	**Modèle réduit**
σ^2	0.18	0.18
τ_{00} individus	0.32	0.32
τ_{00} établissements	0.05	0.05
ICC	0.68	0.68
Établissements	139	139
Individus	3048	3048
Observations	6096	6096
R^2 Marginal	0.184	0.184
R^2 Conditionnel	0.738	0.738

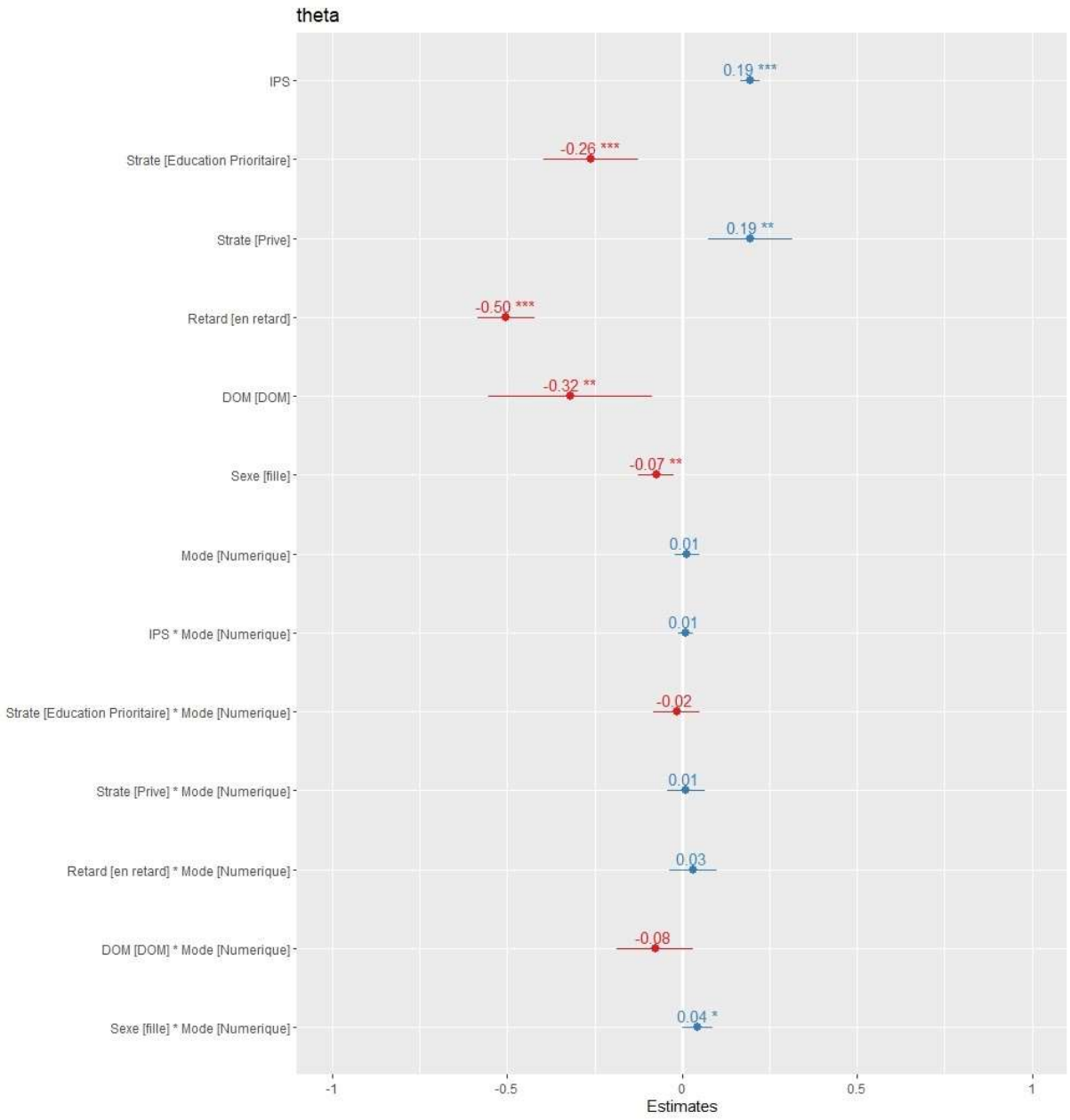
Le coefficient ICC est de 0.68, ce qui peut être interprété comme la part de la variance expliquée par le regroupement par niveau. Un coefficient aussi élevé démontre la pertinence du choix du modèle par rapport à un modèle de régression linéaire simple. Le R^2 marginal de 0.184. Il décrit ici la proportion de la performance de l'élève qui est expliquée par les effets fixes. Enfin, un R^2 conditionnel à 0.738 est élevé, sans être absolu, comme le démontre la figure suivante.



Cette figure permet de détecter des valeurs extrêmes (*outliers*), des effets de non-linéarité ou l'incomplétude du modèle. Ici, nous voyons bien une relation linéaire entre valeurs prédites et résidus. Cela confirme le R^2 conditionnel et implique que le modèle « manque » quelque chose. Dans notre cas, cela a peu d'importance, puisque nous ne cherchons pas à créer un modèle prédictif mais à capter l'importance de la relation entre les variables explicatives et la variable d'intérêt. Toutefois, nous serons prudents sur l'interprétation d'éventuels effets, et nous sommes conscients du risque de présence de facteurs de confusion.

Sans pour autant rejeter l'intérêt d'étudier les effets aléatoires en tant que tels, nous concentrons notre étude sur l'observation des effets fixes. Nous les représentons sur la figure suivante. Ce graphique représente le coefficient de régression de chacune des variables conservées comme effets fixes dans le modèle. La barre horizontale représente, pour chacun, l'intervalle de confiance à 95%. Les effets négatifs sont en rouge, les effets positifs sont en bleu. Enfin, les étoiles représentent la significativité statistique de ces effets, selon la nomenclature suivante :

- p-value < 0.05 *
- p-value < 0.01 **
- p-value < 0.001 ***



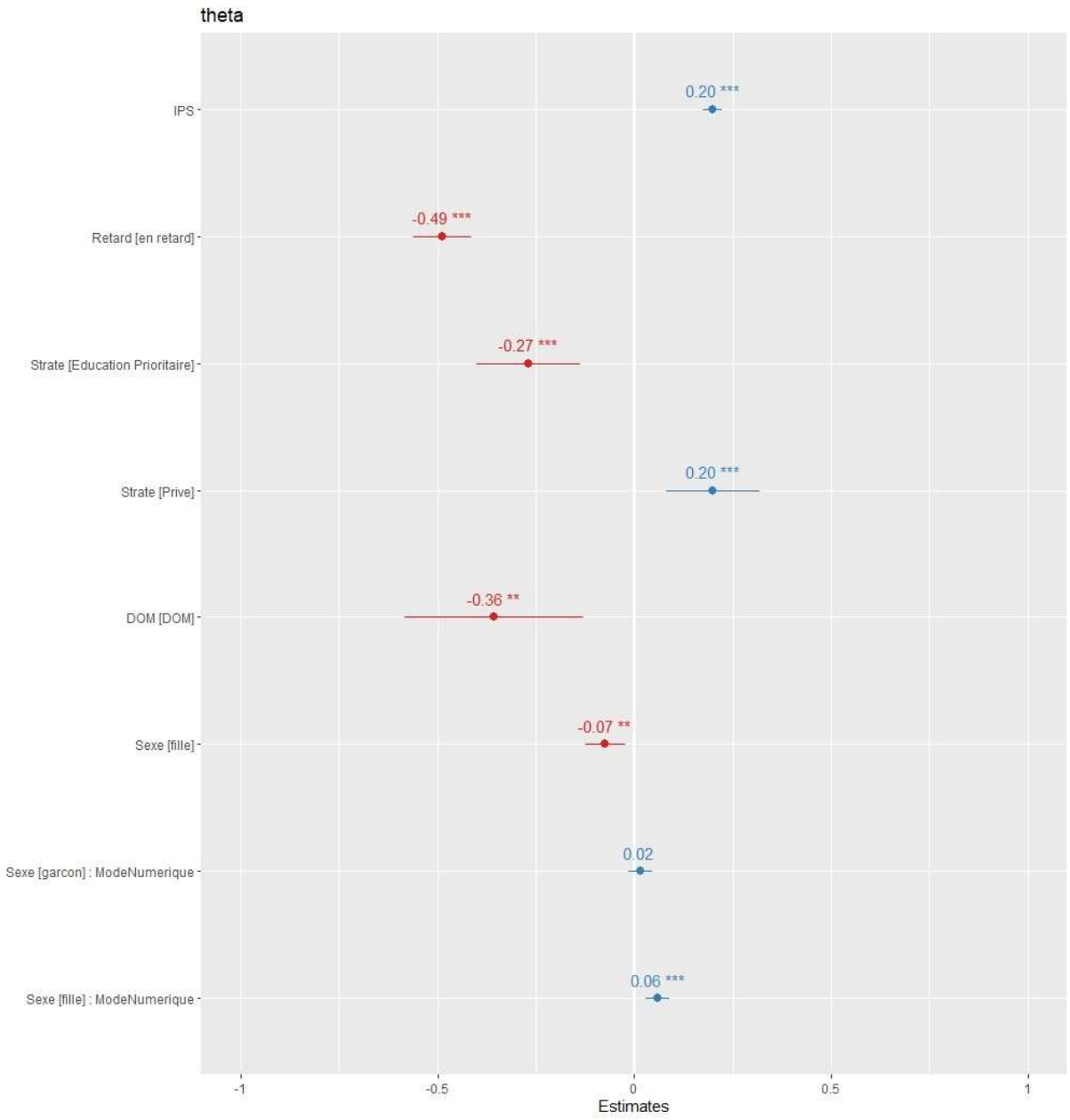
Ainsi, nous constatons que les variables explicatives socio-démographiques usuelles présentent toutes, à l'exception du sexe, un lien avec la performance de l'élève. Ainsi, un IPS élevé, la scolarisation dans le secteur privé sont positivement corrélés avec la performance de l'élève. A contrario, le fait d'avoir redoublé, la scolarisation dans un établissement relevant de l'éducation prioritaire et la scolarisation outre-mer sont négativement corrélés avec la performance de l'élève. Enfin, les filles, toutes choses égales par ailleurs, sous-performent par rapport aux garçons. Tous ces éléments sont en cohérence avec les résultats obtenus sur les différentes cohortes de CEDRE Mathématiques en fin de collège[23].

À l'inverse des variables socio-démographiques, l'effet du mode de passation de l'évaluation sur la performance des élèves n'est pas significatif. Toutefois, on constate un léger effet d'interaction entre le mode de passation et le sexe des élèves, au bénéfice des filles. Pour confirmer cela, nous réduisons le modèle aux variables signifiantes :

theta ~ 1 + IPS + Strate + Retard+ DOM + Sexe + Sexe:Mode + (1 | taker) + (1 | UAI)

Les indicateurs du modèle restent identiques, les deux modèles n'étant pas significativement différents.

Les effets fixes de ce modèle réduit sont décrits sur la figure suivante. Les caractéristiques du graphique précédent sont reprises à l'identique.



On voit ici apparaître nettement l'effet d'interaction précédemment cité : pour les garçons, la transition au format numérique est sans effet significatif. Par contre, une passation numérique est légèrement favorable à la performance des filles par rapport à une passation « papier- crayon », et ce de manière significative.

4 Discussion

4.1 Le construit est maintenu

Dès 1993, Mead et Dragow[24] insistent sur le prérequis à toute étude de comparabilité : le construit doit être maintenu. Sans quoi, les deux modes de passation du test ne sauraient être considérés comme comparables[25]. Pour cela, différentes méthodes sont utilisables, dans le champ de la Théorie de la Réponse à l'Item[26] comme dans celui des équations structurelles[27]. Le choix d'un modèle bifactoriel permet d'allier les deux cadres théoriques. La caractérisation des items selon une analyse factorielle confirmatoire, par le biais d'un modèle bifactoriel, permet de déterminer les contributions de chaque item pour chacun des modes, pour le facteur général du modèle ainsi que pour les facteurs spécifiques. Le fait que cette « qualité » de l'item, au sens de sa contribution au construit, reste globalement inchangée selon le mode de passation, nous permet de considérer que le construit étudié est équivalent quel que soit le mode.

4.2 Les passations numériques sont perçues comme plus difficiles

Dans de nombreux cas pratiques[28][29][30], on constate que le passage au numérique augmente la difficulté perçue des items. Cet effet mode peut varier selon la discipline[31], mais aussi selon le support numérique vers lequel la transition est réalisée[32]. Dans notre situation précise, on constate que la difficulté des items est bien augmentée lors du passage au numérique. Les items présentent un effet mode moyen équivalent à 0.16 écart-type, en faveur du mode « papier-crayon ». Cet effet mode est relativement élevé si on le compare à d'autres cas pratiques, mais tout dépend du construit étudié (ici en l'occurrence le programme de mathématiques de fin de cycle 4).

4.3 L'effet mode dépend des caractéristiques des items

Si l'on peut calculer un effet mode moyen, il ne faut pas nécessairement en déduire que l'effet mode est constant pour l'ensemble des items. Ainsi, il a été démontré que l'effet mode porté par les items était d'autant plus important que la tâche imposée par l'item était complexe[33]. Ainsi, du questionnaire à choix multiples jusqu'à la question rédactionnelle ouverte, l'amplitude de l'augmentation de la difficulté de l'item peut être variable[34]. Dans notre situation, si les effets modes semblent varier faiblement, on constate malgré tout que les items au format « peigne » semblent se distinguer comme étant sujets à un effet mode plus fort que les QCM. A contrario, les questions ouvertes ne semblent pas se distinguer.

4.4 La performance des élèves est globalement préservée

Traditionnellement, la variabilité des performances entre sous-populations est estimée par une analyse de potentiels fonctionnements différentiels d'item[35]. Le choix d'un modèle multi-niveaux en aval de la prise en compte des effets modes potentiels présente un pouvoir explicatif nettement supérieur, mais écarte l'étude de potentiel effets non linéaires. Pour ce qui est de notre étude, s'il n'y a pas de modification générale de la performance des élèves entre le format traditionnel et le format numérique, on constate une amélioration significative (bien que faible) de la performance des filles quand la transition a lieu au format numérique.

5 Conclusion

Ces résultats permettent de dessiner une image assez précise des effets de la transition numérique. Tout d'abord, le construit étudié est préservé. C'est la condition sine qua none du maintien de la série temporelle : en effet, il n'est pas pertinent de maintenir une comparaison dans le temps si le construit varie entre les différents points de mesure. Toutefois, si le construit est préservé, ce n'est pas le cas des caractéristiques psychométriques des items. Ceux-ci étant globalement perçus comme plus difficiles au format numérique, il est essentiel de prendre en compte cet « effet mode » dans le calcul des scores des élèves, au risque de percevoir une baisse artificielle de la performance globale. Enfin, après intégration de la variation de la difficulté des items selon le mode, on constate un léger effet mode selon le sexe des élèves. Sa faiblesse ne remet pas en question la série temporelle dans son ensemble, mais comme souvent lors de changement méthodologique en cours de série longitudinale, il est important de préciser l'existence d'un léger biais entre le dernier point « papier-crayon » et le premier point « numérique », en ce qui concerne l'écart entre filles et garçons.

Dans un premier temps, nous pouvons observer un certain nombre de limites au design que nous avions choisi. Nous citerons trois éléments que nous pensons améliorables dans l'optique d'une reproduction future de cette étude. Tout d'abord, le nombre d'items que nous avons choisi de conserver dans le cadre de cette étude de comparabilité est assez faible. Seuls 34 items ont été considérés comme suffisamment robustes, au sens où ils contribuaient suffisamment au facteur général du modèle bifactoriel choisi. Un seuil de contribution minimum à 0.4 a été choisi, afin de réduire le bruit lié aux items de faible qualité. Le choix d'un seuil plus tolérant aurait été envisageable. Concernant les items toujours, parmi ces 34 items, certains n'ont atteint le seuil de contribution de 0.4 que pour une seule de leurs modalités (papier ou numérique). Ainsi, avec pour objectif de pouvoir comparer l'effet mode sur un item, il était nécessaire que les deux modalités soient présentes. Ce qui a encore réduit le nombre d'items à 22. Enfin, avec si peu d'items, il était difficile d'envisager des analyses complémentaires, notamment quand il s'est agi de comparer l'ampleur de l'effet mode en fonction de variables secondaires, telles que le format de l'item, ou son domaine de compétences mathématiques.

Cette forte attrition du nombre d'items est une limite de l'étude, mais elle ouvre également la porte à des sujets de recherche supplémentaires. En effet, l'étude des items écartés par l'analyse psychométrique est une source d'hypothèses intéressante. Tout d'abord, dans quelle mesure le choix d'un seuil de contribution plus bas permettrait de maintenir les résultats observés ? Le choix d'un seuil à 0.4 peut être considéré comme un choix assez fort, on peut trouver dans la littérature des études avec un seuil inférieur, comme 0.3 [36]. Dans quelle mesure la conservation d'items supplémentaires (à la contribution située entre 0.3 et 0.4) confirme ou infirme les hypothèses précédentes ? De façon plus générale, alors que nous constatons une forte stabilité de la contribution des items selon le mode de passation, un certain nombre d'entre eux voient leur contribution suffisamment varier pour qu'une de leurs modalités disparaisse. Ces items présentent-ils une particularité didactique, ou s'agit-il simplement d'un effet de seuil ? Enfin, cette étude permet de mettre en lumière l'effet de la transition du « papier-crayon » vers une passation en ligne sur ordinateur, en mathématiques, pour des élèves de 3e. Dans quelle mesure retrouve-t-on des résultats similaires pour une autre discipline, pour un autre format numérique (des tablettes mobiles, par exemple), et pour un autre niveau scolaire ? Le cycle des évaluations CEDRE est un bon cadre pour reproduire cette étude sous d'autres paramètres.

6 Annexes

6.1 Tableau des contributions des items

item	Domaine	Format	Contribution papier	Contribution numérique
2	Grandeurs et mesures	QCM	0,71	0,71
12	Nombres et calcul	Champ libre	0,64	0,66
13	Nombres et calcul	QCM	0,44	0,40
14	Nombres et calcul	QCM	0,72	0,65
19	Nombres et calcul	QCM	0,56	0,57
21	OGD Fonctions	QCM	0,62	0,61
29	OGD Fonctions	QCM	0,65	0,58
30	Nombres et calcul	QCM	0,76	0,76
36	OGD Fonctions	Peigne	0,72	0,76
37	OGD Fonctions	Peigne	0,69	0,70
38	OGD Fonctions	Peigne	0,56	0,50
42	Grandeurs et mesures	QCM	0,53	0,55
44	Grandeurs et mesures	QCM	0,65	0,57
47	Nombres et calcul	Peigne	0,52	0,47
48	Nombres et calcul	Peigne	0,52	0,48
50	Grandeurs et mesures	QCM	0,49	0,52
55	Nombres et calcul	Champ libre	0,64	0,65
63	Espace et géométrie	QCM	0,67	0,70
68	Nombres et calcul	Champ libre	0,77	0,77
73	Nombres et calcul	QCM	0,59	0,61
74	Espace et géométrie	QCM	0,59	0,59
75	Nombres et calcul	QCM	0,70	0,70

6.2 Tableau des difficultés des items

item	Domaine	Format	Difficulté papier	Difficulté numérique
2	Grandeurs et mesures	QCM	-0,53	-0,32
12	Nombres et calcul	Champ libre	0,34	0,67
13	Nombres et calcul	QCM	0,39	0,60
14	Nombres et calcul	QCM	-1,06	-0,94
19	Nombres et calcul	QCM	-0,04	0,17
21	OGD Fonctions	QCM	-0,43	-0,22
29	OGD Fonctions	QCM	-0,04	0,14
30	Nombres et calcul	QCM	-0,89	-0,81
36	OGD Fonctions	Peigne	0,05	0,24

37	OGD Fonctions	Peigne	0,04	0,43
38	OGD Fonctions	Peigne	0,57	1,21
42	Grandeurs et mesures	QCM	0,14	0,66
44	Grandeurs et mesures	QCM	-0,52	-0,59
47	Nombres et calcul	Peigne	-3,08	-2,86
48	Nombres et calcul	Peigne	-1,28	-0,29
50	Grandeurs et mesures	QCM	-0,62	-0,52
55	Nombres et calcul	Champ libre	0,34	0,49
63	Espace et géométrie	QCM	-0,77	-0,65
68	Nombres et calcul	Champ libre	0,94	1,13
73	Nombres et calcul	QCM	-1,80	-1,40
74	Espace et géométrie	QCM	0,21	0,15
75	Nombres et calcul	QCM	-0,64	-0,39

Bibliographie

[1] Rocher, T. (2015). Quelles méthodes pour l'évaluation standardisée des compétences des élèves? Statistique et Société, 3(2), 59-66.

[2] Angoff, W. H. (1971). The College Board Admissions Testing Program: A technical report on research and development activities relating to the Scholastic Aptitude Test and Achievement Tests.

[3] Phelps, R. P. (2016). Teaching to the test: A very large red herring. Nonpartisan Education Review/Essays, 12(1).

[4] Sinharay, S., & Holland, P. (2006). The correlation between the scores of a test and an anchor test. ETS Research Report Series, 2006(1), i-28.

[5] Von Davier, M., & Von Davier, A. A. (2004). A unified approach to IRT scale linking and scale transformations. ETS Research Report Series, 2004(1), i-21.

[6] Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. Journal of Educational Measurement, 41(1), 43-68.

[7] Rocher, T. (2016). Construction d'un indice de position sociale des élèves. Éducation et formations, (90), 5-27.

[8] Thionet, P. (1961). La méthode des sondages. Revue de Statistique Appliquée, 9(1), 7-52.

[9] Sautory, O. (1991). Redressements d'échantillons d'enquêtes auprès des ménages par calage sur marges. Proceedings, Journée de Méthodologie Statistique, 299-326.

[10] Van der Linden, W. J., Veldkamp, B. P., & Carlson, J. E. (2004). Optimizing balanced incomplete block designs for educational assessments. Applied psychological measurement, 28(5), 317-331.

[11] Tillé, Y. (1998). Estimation in surveys using conditional inclusion probabilities: Simple random sampling. International Statistical Review, 66(3), 303-322.

[12] Van der Linden, W. J., & Hambleton, R. K. (Eds.). (1996). Handbook of Modern Item Response Theory. Springer Science & Business Media.

[13] Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 28(4), 989-1020.

[14] Reise, S. P. (2012). The rediscovery of bifactor measurement models. Multivariate behavioral research, 47(5), 667-696.

[15] Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. Journal of personality assessment, 92(6), 544-559.

[16] Tian, F. (2011). A comparison of equating/linking using the Stocking-Lord method and concurrent calibration with mixed-format tests in the non-equivalent groups common-item design under IRT (Doctoral dissertation, Boston College. Lynch School of Education).

[17] Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. Psychometrika, 54(3), 427-450.

[18] Kim, J. K., & Nicewander, W. A. (1993). Ability estimation for conventional tests. Psychometrika, 58(4), 587-599.

[19] Givord, P., & Guillerm, M. (2016). Les modèles multiniveaux (No. m2016-05). Institut National de la Statistique et des Etudes Economiques.

[20] Bressoux, P., Coustère, P., & Leroy-Audouin, C. (1997). Les modèles multiniveaux dans l'analyse écologique: le cas de la recherche en éducation. Revue française de sociologie, 67-96.

[21] Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω H: Their relations with each other and two alternative conceptualizations of reliability. Psychometrika, 70(1), 123-133.

[22] Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods in ecology and evolution, 4(2), 133-142.

[23] Ninnin, L. M., & Salles, F. (2020). Cedre 2008-2014-2019. Mathématiques en fin de collège: des résultats en baisse.

[24] Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. Psychological bulletin, 114(3), 449.

[25] Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. Educational Measurement: Issues and Practice, 20(3), 16-25.

[26] Buerger, S., Kroehne, U., & Goldhammer, F. (2016). The transition to computer-based testing in large-scale assessments: Investigating (partial) measurement invariance between modes. Psychological Test and Assessment Modeling, 58(4), 597.

[27] Schroeders, U., & Wilhelm, O. (2011). Equivalence of reading and listening comprehension across test media. Educational and Psychological Measurement, 71(5), 849-869.

[28] Jerrim, J., Micklewright, J., Heine, J. H., Salzer, C., & McKeown, C. (2018). PISA 2015: how big is the 'mode effect' and what has been done about it? Oxford Review of Education, 44(4), 476-493.

[29] Khoshsima, H., & Hashemi Toroujeni, S. M. (2017). Comparability of Computer-Based Testing and Paper-Based Testing: Testing mode effect, testing mode order, computer attitudes and testing mode preference. International Journal of Computer (IJC), 24(1), 80-99.

[30] Fishbein, B., Martin, M. O., Mullis, I. V., & Foy, P. (2018). The TIMSS 2019 item equivalence study: examining mode effects for computer-based assessment and implications for measuring trends. Large-scale Assessments in Education, 6(1), 11.

[31] Backes, B., & Cowan, J. (2018). Is the Pen Mightier than the Keyboard? The Effect of Online Testing on Measured Student Achievement. Working Paper 190. National Center for Analysis of Longitudinal Data in Education Research (CALDER).

[32] Schroeders, U., & Wilhelm, O. (2010). Testing reasoning ability with handheld computers, notebooks, and paper and pencil. European Journal of Psychological Assessment, 26(4), 284.

[33] Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. Handbook of test development, 329-347.

[34] Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. The Journal of Technology, Learning and Assessment, 6(9).

[35] Feskens, R., Fox, J. P., & Zwitser, R. (2019). Differential item functioning in PISA due to mode effects. In Theoretical and Practical Advances in Computer-based Educational Measurement (pp. 231-247). Springer, Cham.

[36] DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. Practical Assessment, Research, and Evaluation, 14(1), 20.