Royaume du Maroc

Conseil Supérieur de l'Enseignement

Programme National d'Evaluation des Acquis

PNEA 2008

RAPPORT méthodologique

Mai 2009

RAPPORT METHODOLOGIQUE

PNEA 2008

TABLE DES MATIERES

INTRODUCTION

En vertu de la loi de sa création, l'Instance Nationale d'Evaluation auprès du Conseil Supérieur de l'Enseignement est chargée:

- D'apprécier, de manière globale, les aptitudes, les connaissances et les compétences acquises par les apprenants au cours des cycles de formation et les modalités de leur contrôle ;
- D'évaluer les avantages que tire la collectivité nationale du système d'éducation et de formation eu égard à l'effort financier déployé;
- D'apprécier le développement des performances internes et externes du système d'éducation et de formation et l'amélioration de la qualité des services fournis aux élèves et étudiants ;
- De développer tous les instruments d'évaluation qui concourent au bon exercice de ses fonctions, et soutenir la recherche scientifique dans ce domaine.

S'inscrivant dans cette perspective, le Programme National d'Evaluation des Acquis (PNEA) a été piloté par l'Instance Nationale d'Evaluation du Système d'Education et de Formation auprès du Conseil Supérieur de l'Enseignement (CSE), en collaboration avec le Centre National des Examens et d'Evaluation relevant du Ministère de l'Education Nationale.

Ce programme vise à instaurer un référentiel national d'évaluation à même d'éclairer la prise de décision en matière de politiques éducatives relatives aux programmes, aux curricula et aux apprentissages. Le présent rapport présente de manière détaillée le cadre conceptuel et méthodologique qui a présidé à la mise en place du dispositif et les différentes étapes d'analyse qui s'en sont suivies. A cet effet, ce rapport qui se veut un complément méthodologique du rapport thématique 2009, explique précisément les fondements théoriques et empiriques ayant permis la production des fascicules descriptifs et le rapport analytique du PNEA-2008.

I. CADRE METHODOLOGIQUE DU PNEA-2008

Traiter de l'évaluation des apprentissages dans un contexte de transition d'une pédagogie par objectifs à une pédagogie par compétences est un exercice très délicat. Ainsi, l'approche par compétences structurant les nouveaux programmes devrait, en principe, s'accompagner d'une façon différente d'évaluer. En effet, le modèle pédagogique préconisé par la Charte Nationale d'Education et de Formation et censé être véhiculé par les nouveaux programmes et manuels scolaires fonde l'apprentissage sur le développement des compétences et considère les connaissances en tant que ressources.

En conséquence, à un enseignement basé sur l'acquisition et l'accumulation du savoir dispensé par l'enseignant devrait se substituer un enseignement focalisé sur l'implication active de l'élève dans la construction de ses connaissances et le développement de ses compétences. C'est dire qu'on est en présence de deux paradigmes opposés ; l'un considère l'apprentissage comme simple transmission des connaissances par un acteur externe tandis que l'autre l'envisage comme un processus qui se construit par le sujet apprenant lui-même. Ce changement de paradigme doit logiquement se traduire par un changement de contexte d'apprentissage et de pratiques pédagogiques et partant, un changement de la manière d'évaluer les apprentissages.

En fait, la révision des programmes inspirée par la pédagogie par compétences n'est pas accompagnée d'un changement des pratiques pédagogiques classiques qui continuent à prévaloir dans notre système scolaire. Faut-il donc respecter l'esprit de la réforme et procéder à une évaluation des compétences ou plutôt faire une évaluation par objectifs et ce conformément aux pratiques pédagogiques en vigueur ?

Pour résoudre ce dilemme, une évaluation du rendement scolaire basée sur les contenus des programmes scolaires tels qu'ils sont prescrits a été adoptée.

1.1. Développement du dispositif d'évaluation

D'une manière générale, il s'agit de mettre en œuvre un dispositif d'évaluation permettant de mesurer et d'expliquer le niveau d'apprentissage effectif des élèves de 4ème et 6ème années primaire ainsi que celui des élèves de 2ème et 3ème années secondaire collégial en Arabe, en Français, en Mathématiques et en Sciences à un instant donné (juin 2008).

Plus spécifiquement, il s'agit d'une part de concevoir des tests disciplinaires (Arabe, Français, Mathématiques et Sciences) et de les expérimenter auprès d'un échantillon pré-expérimental constitué d'élèves choisis judicieusement, avant leur administration à un échantillon représentatif de la population cible et, d'autre part, pour une analyse plus poussée des données recueillies, d'élaborer des questionnaires "Elève", "Ecole", "Enseignant" et "Parents" qui seront administrés simultanément avec les tests.

1.1.1. Elaboration des tests

La réalisation de cette tâche a été confiée à une équipe de 24 enseignants et inspecteurs pédagogiques sélectionnés en fonction de :

- Leur expérience en matière d'évaluation des apprentissages ;
- Leurs connaissances dans l'élaboration des curricula et programmes scolaires ;
- Leurs contributions et participations aux études similaires.

Cette équipe a été organisée en neuf comités répartis en trois pôles, à savoir «Pôle des Langues», «Pôle des Mathématiques» et «Pôle des Sciences».

D'ailleurs, le Comité de pilotage du Programme National d'Evaluation des Acquis a assuré l'encadrement des travaux selon une méthodologie de travail bien définie.

Puisque le diagnostic des apprentissages porte sur les programmes scolaires officiels, on a d'abord collecté tous les documents pédagogiques qui s'y réfèrent, notamment le livre blanc, la Charte Nationale d'Education et de Formation, les circulaires officielles, les guides d'enseignants ainsi que les livres et les manuels scolaires. Un inventaire des objectifs finaux et des compétences visées a été opéré et ce, pour chaque matière et chaque niveau concernés par l'évaluation des apprentissages.

En outre, des cadres de références par niveau et par matière ont été élaborés. Ces cadres définissent avec précision et d'une manière concise les modèles d'apprentissages auxquels on doit se référer lors de l'élaboration des tests de rendement scolaire. La structure de ces modèles se présente ainsi :

- Définition opérationnelle des domaines principaux d'apprentissages ;
- Détermination des objectifs finaux/compétences selon les domaines d'apprentissage ;
- Précision du degré d'importance des objectifs finaux/compétences ;
- Elaboration des tableaux de spécification pour chaque test ;
- Répartition des items du test selon les domaines d'apprentissages et les niveaux d'habiletés.

Tableau de spécifications

Le tableau de spécifications fait le pont entre les objectifs/compétences visés par les programmes scolaires d'une part, et les items du test, d'autre part, compte tenu de leur représentativité en nombre et leur importance relative. En effet, il est pratiquement impossible de poser toutes les questions/problèmes susceptibles de mesurer les objectifs/compétences d'un programme scolaire annuel et, partant, le tableau de spécifications s'avère un outil incontournable pour avoir un échantillon représentatif de l'univers objet de l'évaluation des apprentissages. D'un point de vue plus technique, le tableau de spécifications est assimilé à un plan d'échantillonnage des questions/problèmes compte tenu du mode de codage des réponses.

1.1.2. Elaboration des tests des Mathématiques

Un diagnostic approfondi du livre blanc et des programmes et manuels scolaires a permis d'inventorier :

- 23 objectifs finaux pour la 4ème année primaire ;
- 18 objectifs finaux pour la 6ème année primaire ;
- 18 objectifs finaux pour la 2ème année du secondaire collégial ;
- 19 objectifs finaux pour la 3ème année du secondaire collégial.

Ces objectifs finaux/compétences sont répartis selon trois domaines d'apprentissages principaux à savoir les activités numériques, les activités géométriques et les activités de mesure. La plupart de ces objectifs (entre 80% et 90% selon le niveau scolaire) concerne les ressources d'apprentissages, alors que le reste est formulé sous forme de situations complexes nécessitant le recours à la combinaison de plusieurs ressources spécifiques de chacun des trois domaines d'apprentissages.

Tous les objectifs finaux/compétences ont été considérés de même importance, en l'occurrence «très important», et donc sont équivalents du point de vue du degré d'importance. Une telle manière de procéder s'avère tout à fait normale puisque seuls les objectifs/compétences principaux du programme scolaire prescrit ont été retenus.

Quant aux niveaux d'habiletés de chaque objectif/compétence, ils ont été définis conformément à la nomenclature internationale des mathématiques en vigueur. En réalité, cette tâche a été rendue facile par le fait que la majorité des objectifs concerne les aspects connaissances

et applications tandis qu'une partie minime seulement d'entre eux porte sur la capacité de résoudre les problèmes (objectifs/compétences relatifs aux situations d'intégration).

En plus de son expérience et de ses perceptions en la matière, l'équipe pédagogique chargée de l'élaboration des tests a procédé à une analyse approfondie des charges horaires et des contenus des programmes et manuels scolaires afin d'estimer l'importance relative des domaines d'apprentissages selon les objectifs/compétences.

Ainsi, a-t'on construit les tableaux de spécifications suivants :

Tableau de spécifications des Mathématiques selon les objectifs (Primaire)

Domaine / objectifs	4ème année primaire				6ème année primaire		
	Connaissance	Application	Problème	Total	Application	Problème	Total
Activités numériques	04%	32%	05%	41%	33%	06%	39%
Activités géométriques	04%	18%	05%	27%	22%	06%	28%
Activités de mesure	04%	23%	05%	32%	27%	06%	33%
Total	**12%**	**73%**	**15%**	**100%**	**82%**	**18%**	**100%**

Tableau de spécifications des Mathématiques selon les objectifs (Secondaire collégial)

Domaine/objectif	2ème année secondaire collégial				3ème année secondaire collégial			
	Connaissance	Application	Problème	Total	Connaissance	Application	Problème	Total
Activités numériques	12%	31%	05%	48%	11%	18%	07%	35%
Activités géométriques	11%	27%	04%	42%	15%	25%	10%	50%
Activités de mesure	03%	06%	01%	10%	05%	08%	03%	15%
Total	**26%**	**64%**	**10%**	**100%**	**30%**	**50%**	**20%**	**100%**

Pour des raisons de coût, de précision et d'objectivité, (correction, saisie, traitement..), les tests sont composés essentiellement (environ 80% des items) de questions fermées : questions de type «Vrai/Faux», questions à choix multiples ou questions à réponse courte. Pour ces items, on a opté pour un codage dichotomique en accordant un point à la bonne réponse et 0 à la mauvaise réponse. La résolution des problèmes ne représente qu'à peine 20% des items et le codage adopté consiste à affecter 2 aux réponses complètes, 1 aux réponses partielles et 0 aux réponses fausses. La répartition des items selon le type de question se présente ainsi :

Répartition des items des Mathématiques selon le type de questions (Primaire)

Type de question	4ème année primaire		6ème année primaire	
QCM	15	65%	13	72%
QRC	03	13%	02	11%
QL	01	04%	-	-
QO	04	18%	03	17%
Total	**23**	**100%**	**18**	**100%**

QCM : Questions à choix multiples
QRC : Questions à réponses courtes,
QL : Questions de liaison
QO : Questions ouvertes

Répartition des items des Mathématiques selon le type de questions (Secondaire collégial)

Type de question	2ème année secondaire collégial		3ème année secondaire collégial	
QCM	14	78%	11	58%
QRC	02	11%	04	21%
QL	00	00%	00	00%
QO	02	11%	04	21%
Total	**18**	**100%**	**19**	**100%**

QCM : Questions à Choix Multiples, QRC : Questions à Réponses Courtes, QL : Questions de Liaison, QO : Questions Ouvertes

1.1.3 Elaboration des tests des sciences

- *Eveil Scientifique*

Les thèmes à caractère physique ou biologique prédominent dans le programme de l'Eveil Scientifique au primaire alors que les thèmes géologiques y sont peu présents. Ainsi, ce programme qui traite des concepts fondamentaux, notamment la matière, la vie, le temps, la causalité, se divise en deux domaines principaux : la Physique-Chimie et les Sciences de la Vie et de la Terre. La Physique-Chimie traite quatre sous-domaines, à savoir les Gaz, la Température, les Caractères de l'Etat et l'Electricité. Quant aux Sciences de la Vie et la Terre au primaire, elles portent sur les sous-domaines de la nutrition, du mouvement, du cycle de la vie, des animaux vertébrés, des plantes, de l'eau et de la nature.

Une analyse minutieuse des curricula, des programmes et manuels scolaires a permis de déterminer l'importance relative de chaque sous-domaine d'apprentissage.

A chaque sous-domaine est attribué l'ensemble de ressources correspondant dans le curricula et ces ressources sont réparties selon deux niveaux de ressources cognitives, à savoir le niveau de maîtrise et le niveau de mobilisation.

Importance relative des ressources cognitives par sous-domaine d'apprentissages

Sous-domaine d'apprentissages	Degré d'importance	Niveaux de ressources cognitives	
		Maîtrise	Mobilisation
Gaz	08%	75%	25%
Température	12%	66%	34%
Caractéristiques de l'état	08%	100%	00%
Electricité	16%	25%	75%
Nutrition	12%	75%	25%
Mouvement	06%	100%	00%
Cycle de vie	08%	100%	00%
Classification des animaux vertébrés	06%	50%	50%
Plantes	08%	66%	34%
Eau et Nature	18%	80%	20%

De ce tableau, on déduit le tableau de spécifications suivant :

Tableau de spécifications de l'Eveil Scientifique

Domaine d'apprentissages	Sous-domaines	Niveaux de ressources cognitives	
		Maitrise	Mobilisation
Physique-Chimie	Gaz	06%	02%
	Température	8%	04%
	Caractéristiques de l'état	08%	00%
	Electricité 04%	12%	
Sciences de la Vie et la Terre	Nutrition	09%	03%
	Mouvement	06%	00%
	Cycle de vie	08%	00%
	Classification des animaux vertébraux	03%	03%
	Plantes	05%	03%
	Eau et Nature	14%%	04%

- *Physique-chimie*

Cette matière concerne seulement l'enseignement secondaire collégial. Ainsi, l'élaboration des cadres de référence s'est basée sur les programmes et les manuels scolaires de la physique-chimie de 2ème et 3ème années du secondaire collégial, en plus du livre blanc et des circulaires officielles. Le diagnostic effectué a permis :

- En 2ème année du secondaire collégial, d'inventorier trois domaines d'apprentissage principaux, à savoir la matière, la lumière et l'électricité. Ces domaines principaux sont à leur tour subdivisés en 13 sous-domaines opérationnels ;
- En 3ème année du secondaire collégial, d'inventorier quatre domaines principaux, à savoir l'électricité, la lumière, le mouvement et l'inertie ainsi que les matières. Neuf sous-domaines opérationnels ont été avancés et chacun d'entre eux fait l'objet d'une ou plusieurs questions.

L'importance relative des domaines d'apprentissage est déterminée en fonction de la part de la charge horaire de chaque domaine dans l'enveloppe horaire globale prescrite à la matière. En conséquence, les tableaux de spécifications par domaine d'apprentissages se présentent ainsi :

Tableau de spécifications de Physique-Chimie selon le domaine d'apprentissages

Domaines principaux	2ème année secondaire collégial	3ème année secondaire collégial
La matière/Matières	50%	50%
La lumière	21%	17%
L'électricité	29%	09%
Mouvement et Inertie	-	24%
Total	**100%**	**100%**

Ces domaines principaux d'apprentissage sont eux même subdivisés en sous-domaines opérationnels qui ont fait par la suite l'objet d'une ou plusieurs questions.

Quant aux niveaux d'habiletés, leur importance relative a été définie en tenant compte de :

- La progressivité des apprentissages d'un niveau à l'autre ;
- L'analyse des contenus des programmes et manuels scolaires.

En outre, trois niveaux d'évaluation ont été proposés, à savoir :

- Niveau 1 : Restitution et exploitation des ressources ;
- Niveau 2 : Mobilisation des ressources pour résoudre des situations peu intégrées ;
- Niveau 3 : Mobilisation des ressources pour résoudre des situations complexes.

Ainsi, a-t-on obtenu les tableaux de spécifications suivants :

Tableau de spécifications de Physique-Chimie selon les niveaux d'habiletés

Niveaux taxonomiques/niveau	Connaissances		Habiletés	
	Nombre	Pourcentage	Nombre	Pourcentage
2ème année secondaire collégial	40	60%	26	40%
3ème année secondaire collégial	40	50%	40	50%

Tableau de spécifications de Physique-Chimie selon les niveaux d'évaluation

Niveaux d'évaluation/niveau	Niveau 1	Niveau 2	Niveau 3
2ème année secondaire collégial	60%	30%	10%
3ème année secondaire collégial	50%	30%	20%

D'ailleurs, les items du test sont composés de questions fermées : questions de type « Vrai/Faux », de questions à choix multiples ou de questions à réponse écrite.

1.1.4. Elaboration du test d'Arabe

L'enseignement de la langue « arabe » au primaire et au secondaire collégial est articulé autour de trois composantes, à savoir la lecture, la grammaire ainsi que l'expression écrite et orale.

Puisque le curriculum relatif à la langue « arabe » est focalisé sur l'acquisition des ressources de base préalables au développement des compétences des élèves, le choix s'est porté sur l'évaluation du degré d'acquisition des ressources. Concernant les compétences, le curriculum officiel ne détaille ni la nature et le degré de maîtrise spécifique à chaque niveau scolaire, ni la progressivité de maîtrise en articulation avec le cursus scolaire.

Par conséquent, dans cette étude, l'évaluation a porté sur le degré d'acquisition des ressources considérées comme objectifs d'apprentissage. Un tel choix est motivé par le fait que les programmes et manuels scolaires, même s'ils mentionnent le concept de « compétences », sont surtout axés sur les objectifs finaux d'apprentissage.

En d'autres termes, le dispositif d'évaluation élaboré ne se fixe pas pour objectif de mesurer le degré de maîtrise des compétences, mais plutôt de s'assurer de l'acquisition par les élèves des ressources nécessaires à la maîtrise des compétences telles qu'elles sont tracées dans le curriculum et les programmes officiels.

Pour des raisons de représentativité et de couverture du contenu des programmes officiels, une analyse minutieuse de tous les documents relatifs à l'enseignement de la langue « Arabe » au primaire et au secondaire collégial a permis de déterminer les aspects opérationnels.

1.1.5. Elaboration du test du Français

Les membres de la commission d'élaboration des tests d'évaluation des acquis en Français ont procédé à l'analyse et à l'étude des documents officiels, relatifs à l'enseignement/apprentissage du Français en tenant compte des pratiques pédagogiques en vigueur en vue d'arrêter un nombre de compétences dans les domaines suivants :

- Compréhension de l'écrit ;
- Activités réflexives sur la langue ;
- Langue et communication ;
- Production de l'écrit.

1.2. Pré-test et analyse psychométrique

1.2.1. Analyse psychométrique des items

- *Validité et fidélité des tests*

La validité d'un test réfère au degré avec lequel celui-ci mesure effectivement ce qu'il a pour but de mesurer. C'est le point le plus crucial de l'évaluation d'un instrument. Ainsi, la validité de contenu porte sur la représentativité des items retenus dans le test pour évaluer les acquis scolaires. En effet, les tableaux de spécifications élaborés sont censés représenter les domaines des objectifs/compétences à évaluer. En d'autres termes, les experts pédagogiques sont amenés à respecter les deux points suivants :

- Chaque item doit appartenir à l'univers du construit défini ;
- L'ensemble des items représente tous les aspects de cet univers.

En fait, la validité du contenu n'est pas basée sur les réponses des sujets au test mais plutôt sur la perception des experts ayant élaboré les cadres de référence.

Quant au degré de validité du construit, il est approché par le coefficient d'homogénéité ou de consistance interne basé sur les corrélations entre les items et le test. Ainsi, il est logique de penser que nous somme en présence d'un construit bien défini lorsque tous les items d'un test corrèlent bien entre eux et avec le total : tous les items devraient contribuer à appréhender un(e) même objectif/compétence.

Dans cette étude, l'homogénéité de l'instrument a été considérée comme satisfaisante lorsque la valeur du coefficient alpha de Cronbach était au moins égale à 0,80, seuil généralement utilisé dans ce type de test.

D'ailleurs, on calcule les alphas-i afin de simuler l'impact de chaque item sur la fidélité du test et partant, on écarte tous les items ayant un impact négatif sur la fidélité du test.

Coefficient d'homogénéité ou (de consistance interne) ou Alpha de Cronbach :

L'homogénéité d'un test réfère au degré de cohérence entre les réponses fournies aux différents items, i.e. à quel point chacun des items est une mesure de ce que le test, dans son ensemble, mesure effectivement.

Coefficient d'homogénéité est un indice statistique variant entre 0 et 1 qui permet d'évaluer l'homogénéité (la consistance ou cohérence interne) d'un instrument d'évaluation ou de mesure composé par un ensemble d'items qui, tous, devraient contribuer à appréhender une même entité «sous-jacente»: le niveau de connaissance ou de compétence sur un thème donné; le niveau d'aptitude, d'attitude, de motivation, d'intérêt dans tel domaine ou par rapport à tel objet, etc.

Cet indice traduit un degré d'homogénéité (une consistance interne) d'autant plus élevé(e) que sa valeur est proche de 1. Dans la pratique, on considère généralement que l'homogénéité de l'instrument est satisfaisante lorsque la valeur du coefficient à 0.80.

Le coefficient Alpha se calcule en appliquant l'une des formules suivantes, avec j est le nombre total d'items qui composent l'instrument, S^2_T est la variance de l'instrument dans son ensemble, S^2_i la variance de l'item générique i et r_m est la corrélation moyenne entre tous les couples d'items (pour j items on aura $(j^2 - j) / 2$ coefficients de corrélation) :

$$\alpha = \frac{j}{j - 1}\left[1 - \frac{\sum_i S^2_i}{S^2_T}\right]$$

$$\alpha = \frac{j \times r_m}{1 + (j-1) \times r_m}$$

- **Niveau de difficulté des items**

L'indice de difficulté indique la contribution de l'item au score total. Ainsi, en calculant le pourcentage des élèves qui réussissent l'item, on peut sélectionner les items dont la difficulté est appropriée à la situation. En effet, un item présentant un taux de réussite très bas ou très élevé est peu informatif.

En général, on élimine tous les items dont le taux de réussite est supérieur à 85% ou inférieur à 10% et ce, pour avoir un maximum d'informations sur le niveau des élèves.

- **Pouvoir discriminant des items**

Un item n'est utile pour évaluer les performances des élèves que s'il est significativement discriminatif, c'est-à-dire sensible aux différences d'apprentissage entre les élèves testés. Cette sensibilité est mesurée par l'indice de discrimination.

L'indice de discrimination a été approché par le coefficient de corrélation linéaire entre le score à l'item et le score total. Il est donc normal de ne garder dans le test que les items ayant un pouvoir discriminant relativement élevé.

En pratique, on élimine des tests tous les items dont l'indice de discrimination est inférieur à 0,20.

Indice de discrimination :

Lorsqu'une démarche d'évaluation a pour but de distinguer des individus ou des objets en fonction d'un critère donné (leur niveau de compétence par exemple), on a recours de préférence à des items qui possèdent un pouvoir de discrimination élevé.

Pour calculer l'indice de discrimination, on répartit l'ensemble des élèves en trois groupes selon leur niveau de réussite sur l'ensemble de l'épreuve: les 27 % dont les résultats sont les plus élevés (E), les 27 % dont les résultats sont les plus faibles (F) et les 46 % ayant des résultats intermédiaires.

On considère ensuite les deux premiers groupes seulement (E et F) et on calcule pour chacun d'entre eux la proportion de réussites à l'item: proportion désignée respectivement par R_E et par R_F (dans chaque groupe, rapport entre le nombre d'individus qui réussissent l'item et le nombre total d'individus).

L'indice de discrimination est alors calculé de la manière suivante:

$$D = R_E - R_E$$

Cet indice (qui varie théoriquement entre –1 et +1) indique dans quelle mesure l'item considéré est apte à discriminer les élèves de la même manière que le fait l'ensemble de l'épreuve. On considère donc que le pouvoir de l'item est d'autant plus élevé que la valeur de D est proche de +1.

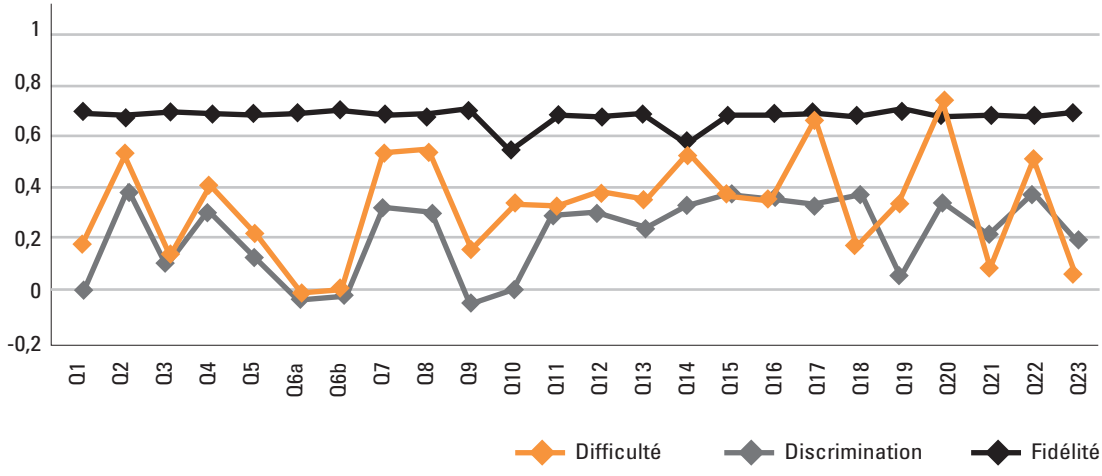
D'autre part, une valeur négative de cet indice révèlerait l'existence d'une sorte d'anomalie, car l'item serait globalement mieux réussi par les élèves ayant les résultats les plus faibles sur l'ensemble du test que par les élèves qui présentent des résultats plus élevés.

1.2.2. Analyse psychométrique du test des mathématiques

En fait, ces indices calculés pour les items retenus dans les tests de mathématiques sur la base des données du pré-test montrent que :

Pour la quatrième année primaire

Figure 1. Indices des items du test des mathématiques (4° année primaire)

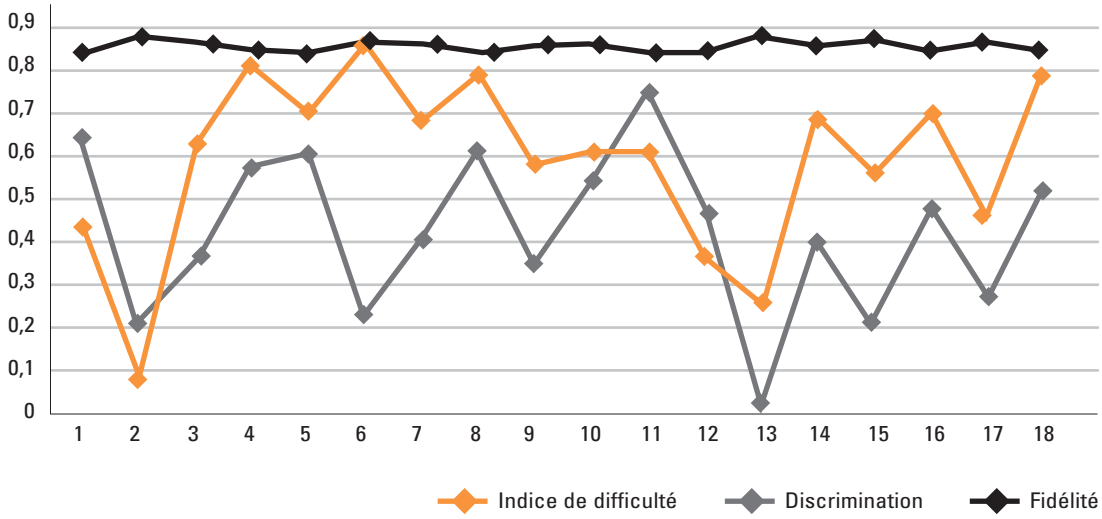


- Tous les indices de difficulté sont inférieurs à 0,85 mais quatre items d'entre eux sont inférieurs à 0,1 et partant sont très difficiles ;
- Neuf items parmi 23 items ne sont pas discriminants car leurs indices de discrimination sont tous inférieurs à 0,20 ;

- Tous les items ont une fidélité faible : leurs indices de fidélité sont inférieurs à 0,80.

Pour la sixième année primaire

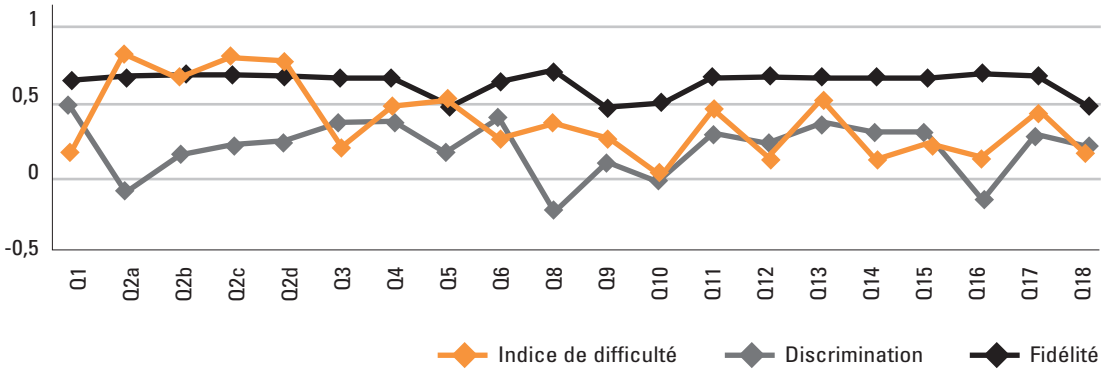
Figure 2 : Indices des items du test des mathématiques (6° année primaire)



- A part l'item 1 dont l'indice de difficulté est inférieur à 0,10, tous les autres items ont des indices de difficulté respectant les normes requises ;
- Tous les items ont des indices de fidélité acceptables car ils s'alignent tous sur le seuil de 0,80 ;
- Quatre items parmi 18 items ne sont pas discriminants car leurs indices de discrimination sont inférieurs à 0,20.

Pour la deuxième année collégiale

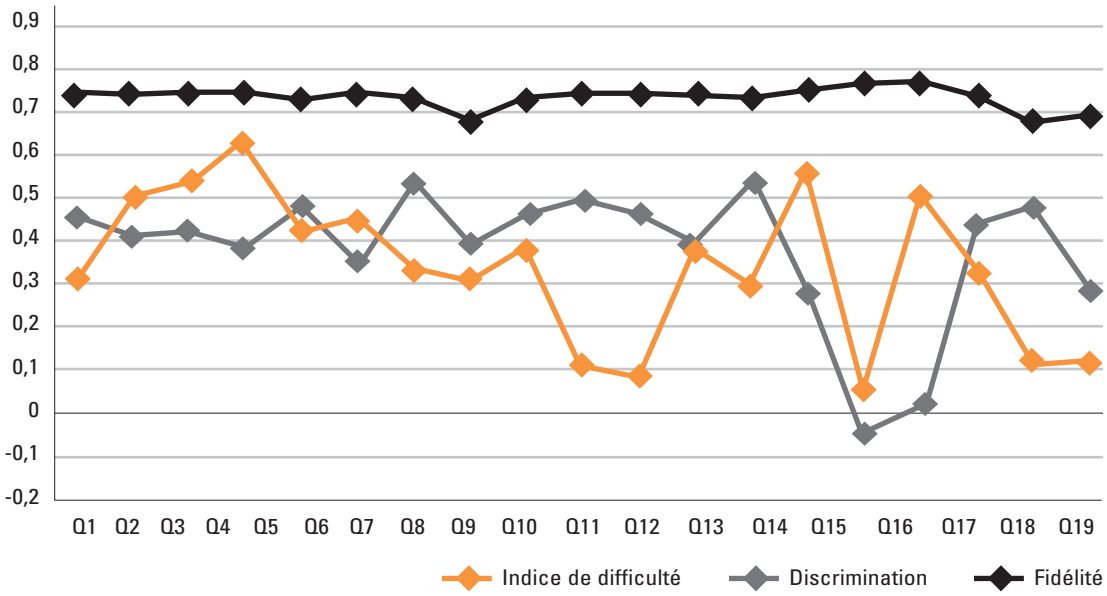
Figure 3 : Indices des items du test des mathématiques (2° année collégiale)



- Les indices de difficulté respectent les normes requises : ils sont tous supérieurs à 0,10 et inférieurs à 0,80 ;
- Sept items ne sont pas discriminants car leurs indices de discrimination sont inférieurs à 0,20 ;
- Tous les items sont de faible fidélité car leurs indices de fidélité sont inférieurs à 0,80.

Pour la troisième année collégiale

Figure 4 : Indices des items du test des mathématiques (3° année collégiale)



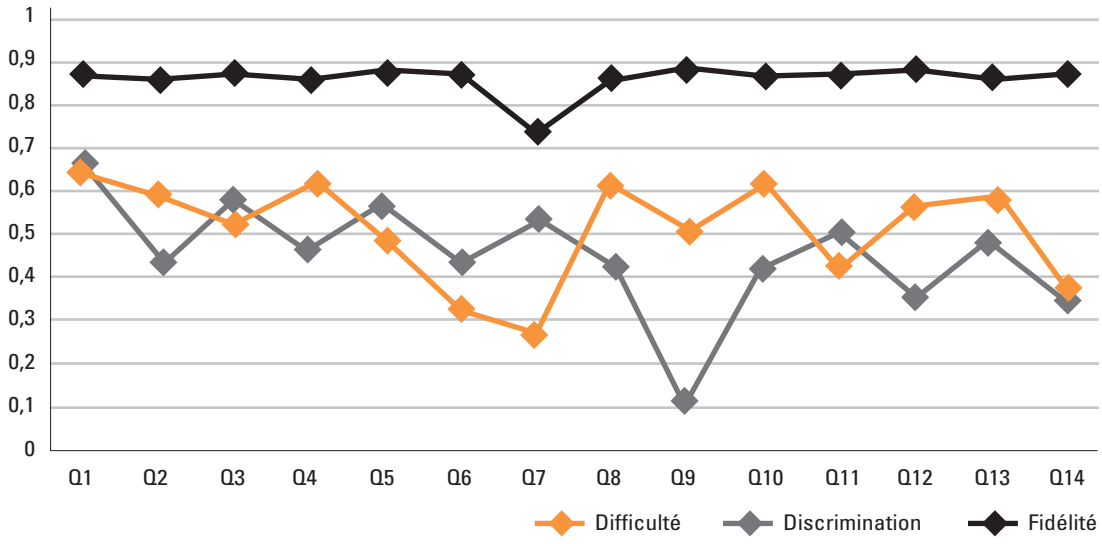
- Les indices de difficulté respectent les normes requises : ils sont tous supérieurs à 0,10 et inférieurs à 0,85 ;
- Seuls deux items ont des indices de discrimination inférieurs à 0,20 et par conséquent ne sont pas discriminants ;
- A part trois items dont les indices de fidélité sont inférieurs à 0,70, tous les autres items ont une fidélité qui dépasse ce niveau sans toutefois atteindre le seuil 0,80.

1.2.3. Analyse psychométrique du test de Français

Les indices de difficulté, de discrimination et de fidélité calculés pour les items retenus dans les tests de français sur la base des données du pré-test montrent que :

Pour la quatrième année primaire

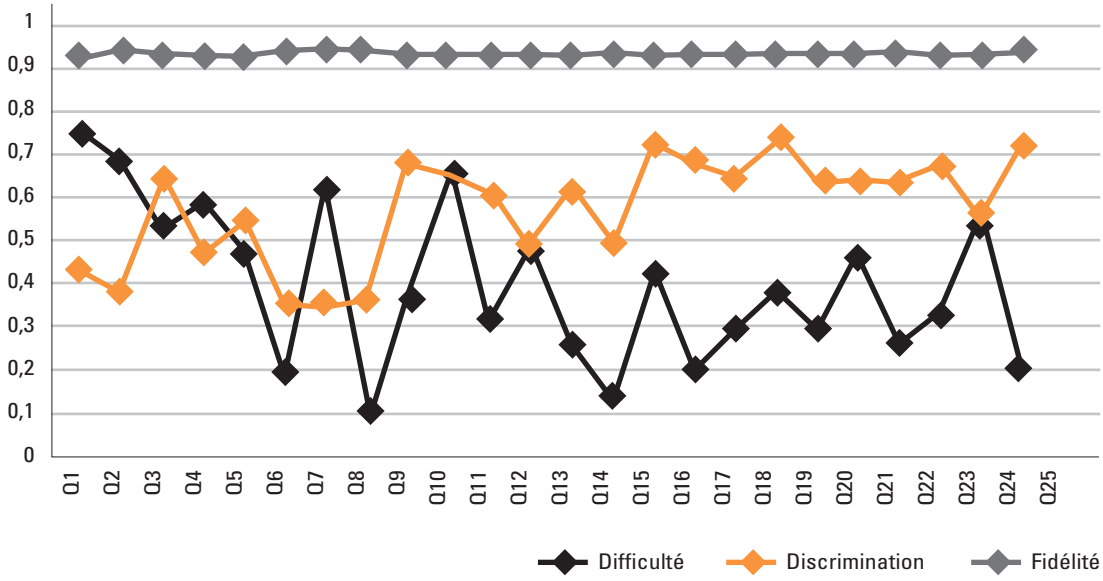
Figure 1 : Indices des items du test de français (4° année primaire)



- Tous les indices de difficulté sont inférieurs à 0,85 et supérieurs à 0,10 et partant, tous les items ne sont ni très faciles ni très difficiles ;
- Tous les items sont discriminants car leurs indices de discrimination sont supérieurs à 0,20 ;
- A part l'item Q7, tous les items ont une fidélité forte : leurs indices de fidélité sont supérieurs à 0,80.

Pour la sixième année primaire

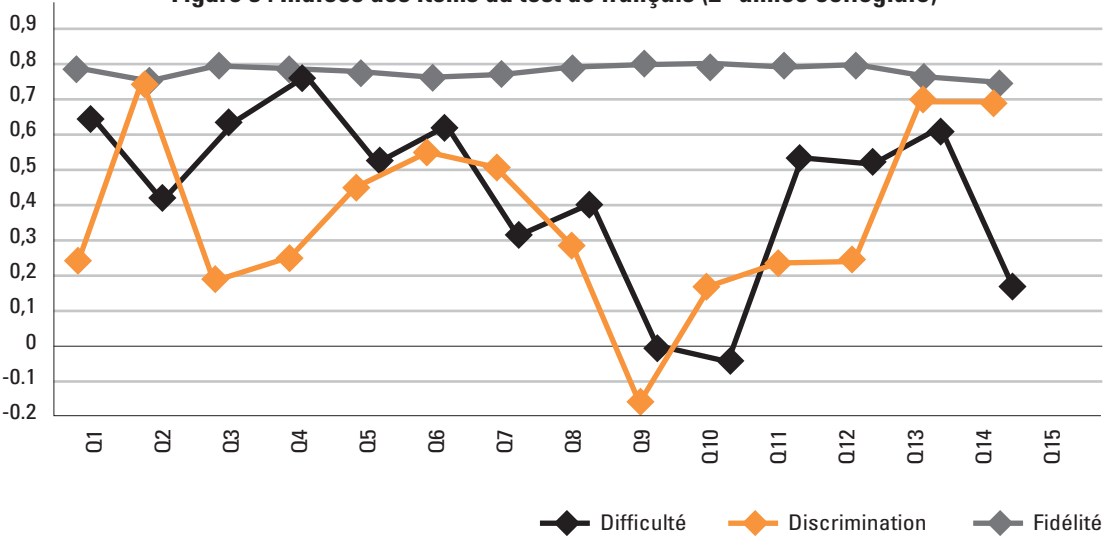
Figure 2 : Indices des items du test de français (6° année primaire)



- A part l'item Q8 jugé très difficile car son indice de difficulté est inférieur à 0,10, tous les autres items ont des indices de difficulté respectant les normes requises ;
- Tous les items ont des indices de fidélité très élevés car ils dépassent le seuil de 0,80 ;
- Tous les items sont discriminants car leurs indices de discrimination sont supérieurs à 0,20.

Pour la deuxième année collégiale

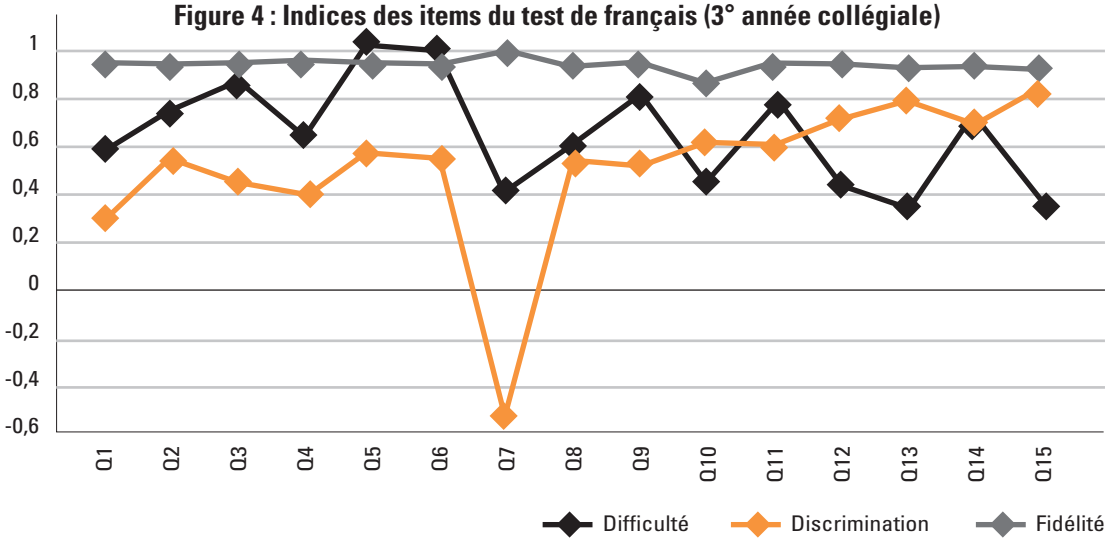
Figure 3 : Indices des items du test de français (2° année collégiale)



- Seuls les items Q9 et Q10 sont jugés très difficiles car leurs indices de difficulté sont inférieurs à 0,10 ;
- A part l'item Q9, tous les autres items sont discriminants car leurs indices de discrimination sont supérieurs à 0,20 ;
- Tous les items ont une fidélité acceptable car même si leurs indices de fidélité sont inférieurs ou égaux à 0,80, ils restent supérieurs à 0,70.

Pour la troisième année collégiale

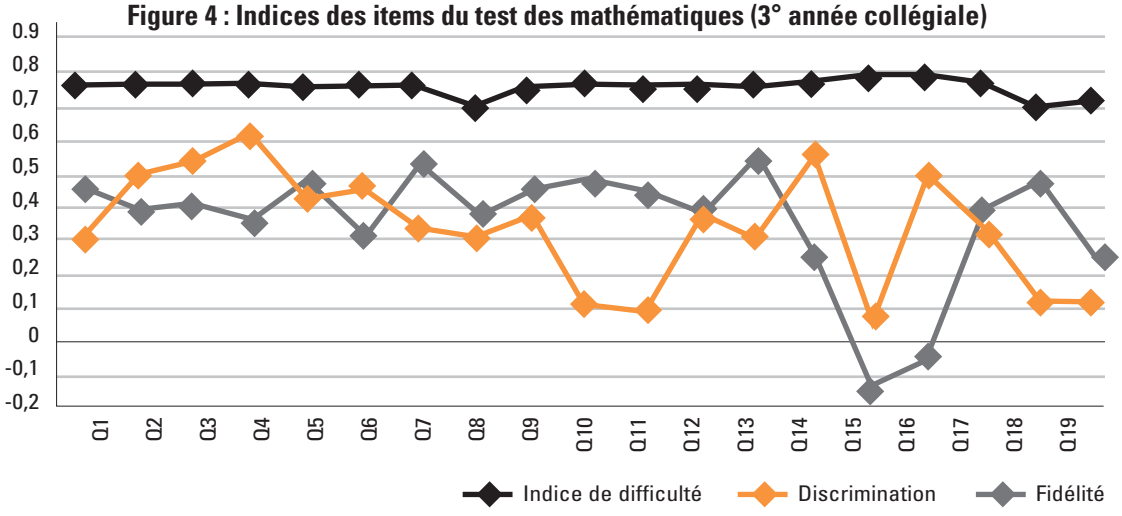
Figure 4 : Indices des items du test de français (3° année collégiale)



- Les indices de difficulté respectent les normes requises : ils sont tous supérieurs à 0,10 et inférieurs à 0,85 ;
- Seuls l'item Q7 est jugé non discriminant puisque son indice de discrimination est inférieur à 0,20 ;
- Le test est d'une fidélité acceptable puisque tous les indices de difficultés sont alignés sur 0,80.

Pour la troisième année collégiale

Figure 4 : Indices des items du test des mathématiques (3° année collégiale)



- Les indices de difficulté respectent les normes requises : ils sont tous supérieurs à 0,10 et inférieurs à 0,85 ;
- Seuls deux items ont des indices de discrimination inférieurs à 0,20 et par conséquent, ne sont pas discriminants ;
- A part trois items dont les indices de fidélité sont inférieurs à 0,70, tous les autres items ont une fidélité qui dépasse ce niveau sans toutefois atteindre le seuil de 0,80.

1.2.4. Analyse psychométrique du test de l'arabe

La même démarche entamée dans les matières précédentes a été poursuivie pour la discipline arabe et a débouché sur des items dûment validés et prêts pour l'expérimentation principale.

1.2.5. Analyse psychométrique du test au niveau des sciences

Au niveau des sciences (Eveil scientifique et sciences de la vie et de la terre), le Centre National des Examens n'a pas procédé à des tests d'analyse psychométriques.

1.3. Elaboration des questionnaires

Les performances des élèves dépendent de plusieurs facteurs, notamment le type et les ressources de l'établissement, les pratiques pédagogiques, la formation des enseignants, la motivation des élèves, le soutien familial, la langue parlée au domicile, les caractéristiques économiques et sociodémographiques des parents. Les questionnaires de contexte jouent donc un rôle central dans l'analyse des performances des élèves.

Afin de recueillir des informations contextuelles permettant d'analyser les facteurs susceptibles d'influencer les performances des élèves, 4 questionnaires ont été développés et validés par un groupe de 22 experts nationaux et sont destinés aux élèves, aux enseignants, aux directeurs d'établissement et aux parents d'élèves. Chaque questionnaire est composé d'axes répartis en pôles de variables qui font elles-mêmes l'objet de questions.

- Le questionnaire «Elève» porte sur les caractéristiques personnelles et familiales de l'élève, les variables relatives aux conditions de scolarité et les perceptions de l'élève sur l'école ;
- Le questionnaire «Enseignant» se focalise sur les caractéristiques personnelles et professionnelles de l'enseignant, ses perceptions du métier d'enseignant, les conditions de travail, les programmes scolaires, etc. ;
- Le questionnaire «Ecole» cerne les caractéristiques personnelles et professionnelles du directeur de l'établissement, les équipements de l'école, la vie scolaire, l'environnement de l'école, l'approche de gestion, etc. ;
- Le questionnaire «Parents» appréhende les variables socio-culturelles, le travail des enfants, les perceptions de l'école, etc.

Ces informations contextuelles permettront d'établir les corrélations éventuelles entre le niveau des performances des élèves et certains facteurs déterminants de la réussite scolaire.

II. PLAN D'ÉCHANTILLONNAGE

L'échantillonnage est la procédure par laquelle les échantillons sont prélevés dans une population. De nombreuses méthodes de tirage sont possibles, chacune ayant ses avantages et ses inconvénients. D'une manière générale, on distingue les méthodes d'échantillonnage probabilistes et non-probabilistes. Les échantillons aléatoires ont pour avantage essentiel de se prêter à une évaluation rigoureuse alors qu'il est impossible de juger la précision et la fiabilité des résultats obtenus par les échantillons non-aléatoires.

Le plan d'échantillonnage PNEA 2008 est plutôt un plan d'échantillonnage mixte et partant la précision et la fiabilité des résultats sont loin d'être évalués. Ainsi, pour sélectionner les échantillons de l'étude, le Centre National des Examens et d'Evaluation a procédé de la manière suivante :

2.1. Base de sondage

Au primaire, d'un effectif de 19233 établissements scolaires publics éligibles (disposant d'au moins un niveau de 4ème année), on a intégré dans la base de sondage 1974 établissements mères ou autonomes à structure complète ayant un ratio élèves /classe élevé.

L'exclusion des écoles satellites et des établissements à faible ratio élèves/classe est un choix justifié par le souci de collecter une information sur les facettes les plus importantes et significatives de la réalité éducative marocaine sans entrer dans le détail de certaines spécificités.

La base de sondage pour l'enseignement collégial public est constituée de 1554 collèges publics à structure complète.

Quant à la base de sondage de la strate "enseignement privé", elle comprend les établissements scolaires privés à structure complète.

2.2. Taille des échantillons

L'échantillon retenu se compose de :

- 230 établissements scolaires au primaire, soit 6900 élèves/niveau à tester ;
- 212 collèges au secondaire collégial, soit 6360 élèves/niveau à tester.

2.3. Stratification

La stratification explicite est effectuée selon la région et l'enseignement privé. Ainsi, a-t-on obtenu 17 strates explicites au primaire (les 16 régions du royaume et une strate pour l'enseignement primaire privé) et 9 strates explicites au secondaire collégial (8 régions groupées et une strate pour l'enseignement collégial privé). Les strates explicites ont été stratifiées à leur tour par la variable implicite « milieu ». En écartant l'enseignement privé en milieu rural, on obtient 33 strates implicites au primaire et 17 strates implicites au secondaire collégial.

2.4. Echantillonnage de premier niveau

Enseignement public

Les bases de sondage adoptées ont été triées par région et par milieu et les tailles des échantillons des régions ont été obtenues par la répartition entre les régions des échantillons nationaux au prorata du nombre d'établissements par région.

Ensuite, on a procédé au tirage d'un échantillon par région.

Enseignement privé

La taille de l'échantillon est obtenue en appliquant le poids des effectifs des établissements privés à la taille de l'échantillon national des établissements scolaires. Ainsi, on a échantillonné 15 établissements scolaires privés au primaire et 10 collèges au secondaire collégial.

Nombre d'écoles par régions

Nombre d'écoles/Région	Nombre d'écoles		
	Public		Privé
	Urbain	Rural	Urbain
Chaouia-Ouerdigha	7	3	0
Doukkala-Abda	6	4	0
Fés-Boulmane	12	3	2
Gharb-Chrarda-Beni Hssain	6	4	1
Grand Casablanca	26	4	6
Guelmim-Essmara	4	1	0
Oriental	11	5	1
Laâyoune-Boujdour-Sakia Al Hamra-Guelmim Essmara & Oued Eddahab - Lagouira	5	0	0
Marrakech-Tensift-Al Haouz	12	4	2
Meknès-Tafilalt	13	4	0
Oued Eddahab-Lagouira	3	0	0
Rabat-Salé-Zemmour-Zaer	14	6	2
Souss-Massa-Draa	10	10	0
Tadla-Azilal	6	4	0
Tanger-Tetouan	14	6	1
Taza-Taounate-Al Houceima	5	3	0
National	**154**	**61**	**15**

2.5. Echantillonnage de deuxième niveau

De chaque établissement scolaire échantillonné, une classe par niveau a été tirée d'une façon aléatoire (tirage aléatoire simple) .

2.6. Echantillonnage de troisième niveau

Tous les élèves des classes échantillonnées font partie de la population à tester.

En principe, on doit échantillonner 30 élèves par classe. Donc, les effectifs de l'échantillon par région, par milieu et par type d'établissement devraient se présenter ainsi :

Effectif par région

Effectifs de l'échantillon/Région	Effectifs de l'échantillon (estimés)		
	Public		Privé
	Urbain	Rural	Urbain
Chaouia-Ouerdigha	210	90	0
Doukkala-Abda	180	120	0
Fés-Boulmane	360	90	60
Gharb-Chrarda-Beni Hssain	180	120	30
Grand Casablanca	780	120	180
Guelmim-Essmara	120	30	0
Oriental	330	150	30
Laâyoune-Boujdour-Sakia Al Hamra-Guelmim Essmara & Oued Eddahab - Lagouira	150	0	0
Marrakech-Tensift-Al Haouz	360	120	60
Meknès-Tafilalt	390	120	0
Oued Eddahab-Lagouira	90	0	0
Rabat-Salé-Zemmour-Zaer	420	180	60
Souss-Massa-Draa	300	300	0
Tadla-Azilal	180	120	0
Tanger-Tetouan	420	180	30
Taza-Taounate-Al Houceima	150	90	0
National	**4620**	**1830**	**450**

III. MÉTHODOLOGIE DE L'ANALYSE CONTEXTUELLE

Le présent chapitre met en évidence la méthodologie d'analyse utilisée dans le rapport analytique, et tente de présenter les fondements théoriques permettant l'analyse de l'incidence des caractéristiques individuelles et contextuelles des élèves, mesurées dans le cadre du PNEA, sur leur rendement par niveau et discipline testés.

Comme souligné dans le rapport analytique, deux démarches complémentaires ont été adoptées. La première est une analyse bi-variée d'hypothèse (ou analyse bi-variée) et la seconde faisant usage de la modélisation multiniveaux. Si la première analyse cherche à examiner les effets absolus des caractéristiques individuelles et contextuelles sur le rendement par le recours aux relations bi-variées, la deuxième analyse ces caractéristiques dans un cadre multidimensionnel qui tient compte des rapports entre les variables. Cette analyse des effets permet de distinguer les variables liées le plus étroitement au rendement.

L'examen du rôle des variables, prises séparément, à travers des tests de chi2 ne tiendra pas compte de la structure hiérarchique de la population étudiée. Il renseignera sur l'existence d'une relation fonctionnelle ou non entre une variable explicative et le rendement scolaire, sans chercher la nature de cette relation.

Le recours aux analyses multiniveaux a été spécifiquement utilisé pour étudier les populations comportant plusieurs niveaux emboîtés. Ces analyses sont appropriées pour les systèmes scolaires, dans lesquels les élèves constituent le premier niveau, les classes le deuxième niveau et l'école le troisième niveau. Pour des considérations d'échantillonnage, seuls deux niveaux seront retenus pour étudier les effets élève et les effets établissement.

3.1. Fondements de l'analyse bi-variée

La présente section traite de l'analyse bi-variée. Elle donne un aperçu théorique sur l'approche utilisée dans la description du rendement des élèves au PNEA-2008, en fonction de certaines variables récurrentes dans les différentes revues de littérature portant intérêt sur les tentatives d'explication du rendement scolaire.

Dans ce rapport technique, l'accent sera mis sur l'idée sous jacente derrière l'utilisation des tests non paramétriques de chi2 ayant servi pour apprécier l'existence d'une relation entre les performances des élèves et certaines variables prises séparément.

Ensuite, seront rappellées de manière précise la formulation théorique de ce test non paramétrique, ainsi que son application dans le cadre de la recherche de déterminants potentiels des acquis scolaires.

Enfin, le choix des variables retenues sera explicité pour l'examen de ces relations à travers cette analyse bi-variée et également la classification des scores.

3.1.1. Utilisation de l'analyse bi-variée

Avant d'aborder la justification du choix de l'analyse bi-variée et des tests non paramétriques de chi2 ayant servi pour la description des résultats des élèves dans le PNEA-2008, il y a lieu de souligner qu'une analyse descriptive des variables contextuelles a été effectuée, et qu'elle a permis d'avoir une idée sur les profils des élèves, leurs parents, leurs enseignants et leurs établissements.

Par la suite, il a été procédé, séparément, à la fusion du contenu de chaque questionnaire avec les fichiers correspondants, contenant les résultats des élèves, par niveau et discipline, pour pouvoir procéder à l'analyse bi-variée du rendement des élèves avec les variables du contexte. En utilisant les mêmes variables, l'examen des relations pouvant exister entre les performances des élèves avec ces variables a été effectué. Pour des besoins d'analyse, ont été considérés l'ensemble des élèves, par niveau scolaire étudié, pour lesquels l'information relative à tous les tests et chaque questionnaire pris à part, est complète.

L'objectif que sous-tend cette analyse bi-variée est d'examiner l'existence d'une relation entre la variable expliquée, le rendement scolaire et les variables potentiellement explicatives prises séparément. Plus précisément, on cherche la présence des effets absolus des caractéristiques individuelles, familiales et scolaires sur le rendement des élèves. L'effet absolu d'une variable s'exerce bien entendu en l'absence d'autres variables. L'effet absolu mesure en fait l'apport indépendant de la variable au rendement des élèves. Une variable peut avoir de l'importance en soi, mais peut ne pas en avoir lorsque d'autres variables sont prises en considération.

Pour y parvenir, les méthodes d'inférence statistique offrent deux possibilités :

- Utiliser des techniques paramétriques, par des méthodes qui imposent des restrictions sur la forme de la relation, généralement linéaire en l'occurrence ;
- Ou bien procéder par des méthodes non paramétriques qui ne font pas allusion à la forme de la relation.

A ce stade d'analyse, ces dernières méthodes s'avèrent préférables et sont souvent utilisées lorsque l'on cherche à établir des relations générales ou fonctionnelles. De plus, il arrive souvent dans les applications statistiques en sciences sociales en général, et en sciences de l'éducation en particulier, que l'on ait recours à des méthodes non paramétriques pour ce genre d'analyse.

Les méthodes de statistiques non paramétriques offrent une panoplie de tests, applicables dans des situations et des contextes bien déterminés. Les tests d'indépendance de chi2

répondent le mieux à notre situation. En effet, ces tests permettent d'évaluer si la répartition des effectifs dans un tableau de croisé ou de contingence est significativement différente de celle de la table calculée sous l'hypothèse d'indépendance des deux variables croisées. Autrement dit, l'objectif du test du chi2 est de déterminer si les lignes et les colonnes d'un tableau croisé (c'est à dire les deux variables étudiées) sont indépendantes. Par indépendances, on entend :

- le fait d'appartenir à une modalité de la première variable n'a pas d'influence sur la modalité d'appartenance de la deuxième variable ;
- les pourcentages des lignes du tableau croisé sont les mêmes pour toutes les lignes;
- les pourcentages des colonnes du tableau croisé sont les mêmes pour toutes les colonnes.

C'est pour ces raisons que ce genre de test a été choisi, par opposition aux analyses de régression car nous ne nous sommes pas placés dans une perspective de détecter l'existence d'une relation causale mais plutôt d'une liaison fonctionnelle. Ceci va de pair avec l'objectif visé à cet égard, qui consiste à examiner la relation pouvant exister entre le rendement scolaire et les différentes variables explicatives, prises séparément, et de tester cette relation. Ce qui correspond ici à un tableau croisé ou de contingence et répond aux conditions d'utilisation des tests non paramétriques d'indépendance de chi2.

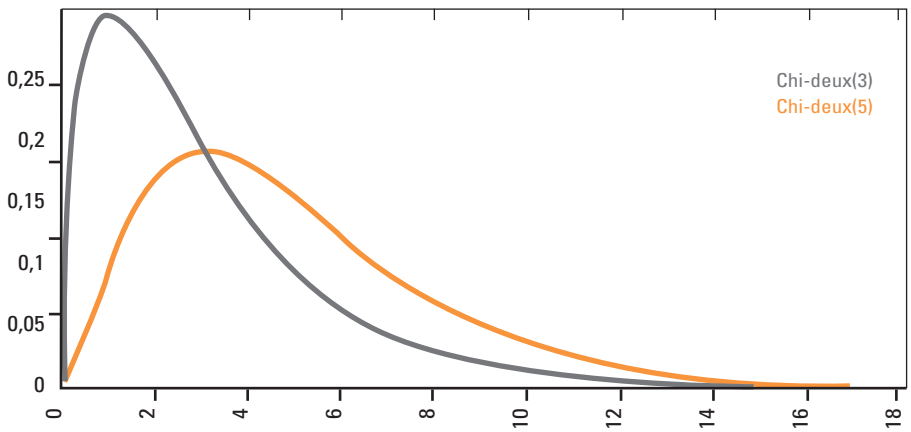
Pour ce faire, le degré de signification de chacune de ces relations a également été évalué en ayant recours au test statistique de chi2. Par ailleurs, nous avons considéré que toute relation ayant une statistique, issue de ce test, dont la probabilité (p-value) inférieure à un seuil de 10%, a été identifié comme significative.

3.1.2. Formulation théorique des tests de chi2

Le test de Chi2 permet de mesurer la significativité de la relation entre deux caractères. L'idée générale de ce test est de calculer la statistique de Chi2 (observée) qui est la somme des déviations entre effectifs observés et effectifs théoriques, présents à l'intérieur d'un tableau de contingence, et de la comparer à la valeur théorique d'une loi de probabilité Chi2 de degré de liberté k et un risque (le risque de rejeter l'hypothèse nulle alors qu'elle est vraie).

La courbe de Chi2 est représentée par la figure suivante, pour différents degrés de liberté :

Courbe de la loi de chi2



Ainsi pour tester s'il y a une relation significative entre deux caractères (par exemple le score des élèves et le genre) :

(1) On pose l'hypothèse H0 : "Il n'y a pas de relation entre les deux caractères Y".

(2) On détermine la valeur Chi-2 Observé du tableau étudié (I x J) et qui est égale à:

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Avec :

- Oij la valeur observée
- Eij la valeur attendue sous l'hypothèse d'indépendance

Et on a : $E_{ij} = \frac{O_{i.} \times O_{.j}}{N}$ Où $O_{i.} = \sum_{j=1}^{J} O_{ij}$ et $O_{.j} = \sum_{i=1}^{I} O_{ij}$

(3) On détermine le nombre de degrés de liberté k = (I-1)*(J-1) où I et J représentent le nombre de lignes et colonnes du tableau.

(4) On fixe le risque d'erreur a de rejeter H0 à tort (ex. a=10%).

(5) On détermine la valeur Chi-2 (k, a). Cette valeur est lue dans une table du test du Chi-2.

(6) On procède au test :

H0 est vraie si : La statistique de Chi-2 Observée est inférieure ou égale à la valeur Chi-2 (k, a) théorique.

Le test de Chi-2 vise à tester l'hypothèse d'indépendance des lignes et des colonnes d'un tableau croisé.

Le test de Chi-2 se base sur la valeur du Chi-2 du tableau, qui est une mesure de l'écart entre le tableau observé et le tableau qu'on aurait obtenu si les variables étaient parfaitement indépendantes, et sur le nombre de degrés de liberté du tableau, qui dépend du nombre de lignes et de colonnes.

A partir de ces deux données, le test donne une valeur *p*, qui est la probabilité que les variables soient indépendantes compte tenu du tableau observé, ou encore le nombre de chances de se tromper si on dit que les deux variables ne sont pas indépendantes. Le seuil de significativité pour le *p* est par convention fixé à 5 %, ou 0,05, ou 5 chances sur cent. Si le *p* est supérieur à ce seuil, autrement dit si on a plus de 5 chances sur 100 de se tromper en disant l'inverse, alors on considère que les deux variables sont indépendantes. Sinon, on considère qu'il y a un lien entre les deux.

Dans cette étude, trois seuils de significativité ou de l'ampleur de l'effet ont été fixés : 1%, 5% et 10%. Ce dernier étant largement acceptable dans le contexte d'études relatives aux sciences de l'éducation.

Les tableaux inclus dans le rapport analytique se fondent sur le Programme National d'Evaluation des Acquis des élèves de 2008. Ils indiquent pour chaque niveau et discipline, le degré de significativité du test de Chi-2, en croisant les scores des élèves avec les variables contextuelles.

3.1.3. Choix des variables

Les questionnaires utilisés et administrés lors du PNEA ont eu pour but de saisir les principales caractéristiques des élèves et des enseignants des classes et des établissements scolaires échantillonnées et de recueillir toute information pertinente susceptible d'être mise en relation avec le rendement scolaire en vue d'une tentative d'explication de ce dernier.

Devant la richesse du contenu des questionnaires et la multiplicité des variables qu'ils contiennent, des choix ont du être faits pour éviter la lourdeur de l'analyse. Il fallait se limiter aux variables les plus pertinentes aidant à la simplification et à la robustesse de l'analyse et des estimations. Pour ce faire, l'étude s'est basée sur certains critères qui favorisent les facteurs les plus récurrents dans les écrits les plus récents, qui évitent les redondances[1] par un recoupement des données entre différents questionnaires et qui présentent une qualité minimale requise pour l'analyse.

L'analyse des facteurs de la réussite scolaire liés à l'élève renvoie souvent à l'âge, le sexe, le retard scolaire, les relations sociales, notamment avec les pairs. Les facteurs psychologiques traditionnels attribués aux élèves sont également cités. Il s'agit de la motivation, la perception ou l'estime de soi, les attentes et les aspirations professionnelles des élèves. Les difficultés extra scolaires du contexte, telles que le rapprochement de l'école des agglomérations, ou encore, des facteurs liés au vécu scolaire des élèves mesuré par leurs antécédents scolaires, sont également présents et constituent des facteurs centraux de la réussite académique.

Selon les chercheurs en matière de rendement interne de l'éducation, la famille est également l'un des facteurs ayant le plus d'impact sur la réussite scolaire. Le traitement de ce facteur est varié, avec des données souvent originales. D'une part, on porte intérêt aux relations parent-enfant, ou parent-école. D'autre part, on privilégie le statut socioéconomique de la famille comme explication de la réussite scolaire. on insiste alors sur l'analyse des facteurs dont plusieurs appréciés par la littérature et qui sont d'actualité. Il s'agit notamment du statut socioéconomique des familles : le nombre d'enfants et la structure des ménages, le niveau d'instruction des parents et de leur profession, les conditions et le milieu de vie, la participation des enfants dans les travaux domestiques et le soutien pédagogique familial par l'aide et l'assistance apportées par les parents à leurs enfants dans leurs devoirs et également les attentes des parents de l'école.

En raison de l'abondance de la littérature sur les facteurs liés à l'enseignant, nous nous limiterons à mentionner ici la compétence du personnel enseignant en insistant sur l'expérience accumulée en nombre d'années d'exercice du métier et en adjoignant les facteurs touchant à la formation initiale et continue. Des facteurs liés aux perceptions, aux attentes et aux attributions des enseignants envers l'école sont également de taille, sans oublier les rapports entretenus avec le milieu scolaire, dont on peut signaler à juste titre les relations enseignant-administration de l'école, enseignant-élève. Des questions telles que la satisfaction et la stabilité ont été également retenues dans cette étude.

Si les facteurs individuels ne sont pas maîtrisables par les pouvoirs éducatifs, le contexte scolaire constitue le champ d'action des établissements scolaires. De nombreux auteurs ont souligné l'importance de l'impact de ces facteurs sur les acquisitions scolaires. Cet impact comprend deux formes : la gestion de l'établissement scolaire sur le plan matériel (ressources pédagogiques, etc.), et organisationnel (ambiance scolaire, civisme), d'une part, et le rôle pédagogique de la direction de l'école (ancienneté, niveau d'instruction, etc.), d'autre part.

En plus des facteurs sus mentionnés, d'autres variables, parfois significatives, pouvaient être traitées à ce niveau. Les informations qu'elles renseignent sont sujettes à un recoupement avec les informations contenues dans les différents questionnaires. Elles se prêtent à titre d'exemple, aux relations parent-élève, parent-établissement scolaire, ou parent-enseignant, ou encore enseignant-établissement scolaire et élève-enseignant. Dans une optique de simplification et afin d'assurer la pertinence de l'analyse, elles ont été traitées soit parmi les facteurs liés à l'élève ou à sa famille, soit dans l'appréciation des caractéristiques des établissements scolaires et des enseignants respectivement.

[1] Certaines variables ont été appréhendées dans deux ou trois questionnaires, qui concernent l'enfant dans sa relation avec sa famille par exemple, ou relations parents-établissements, etc.

Le traitement des données relatives à ces variables a révélé l'existence de certains problèmes liés à la qualité des données recueillis. Lesquels sont traduits dans les éléments suivants :

- Le premier, c'est que ces données n'ont pas la qualité requise pour l'exploitation, autrement dit, sont mal renseignées
- Le second, c'est le taux de non réponses, traduit par la prédominance des données manquantes.

C'est pour cette raison que seules ont été retenues les données qui présentent les qualités requises pour l'exploitation et l'analyse.

3.1.4. Classification des scores

Les résultats des élèves dans le programme national de l'évaluation des acquis sont généralement faibles. La proportion des élèves ayant réussi au moins 50% de bonnes réponses aux tests s'élève dans le meilleur des cas à 46% pour les élèves de la sixième année en sciences. Ainsi défini dans l'Atlas du système d'éducation et de formation (cf. indicateurs 3.22 et 3.23), le taux de connaissances de base apprécie l'effectivité et la qualité de l'éducation à partir de l'évaluation des acquis[2]. Ces résultats obtenus du PNEA-2008, prédisent dans l'ensemble un niveau faible des performances et une qualité encore loin des niveaux souhaités et des objectifs prescrits.

Taux de connaissances de base

	Enseignement primaire		Enseignement collègial	
	4ème année	6ème année	2ème année	3ème année
Arabe	13	24	36	37
Français	29	20	19	17
Mathématiques	23	44	8	16
Sciences	28	46	-	-
SVT	-	-	2	10
Global	**25**	**38**	**24**	**24**

Cependant, il y a lieu de s'interroger, dans quelle mesure, les résultats des élèves au PNEA rendent compte de leur véritable rendement. A priori, il n'y a pas d'éléments suffisants pour dire que les curricula prescrits sont ceux pratiqués. Autrement dit, a t-il été procédé à une vraie mesure de la performance des élèves ou à une mesure partielle ?

En s'appuyant sur ces deux éléments, les performances des élèves ont alors été catégorisées, pour des fins de description selon les caractéristiques individuelles et contextuelles, selon les scores très bas, moyennement bas et bas. Vu ces résultats, il serait pratiquement difficile d'envisager une typologie classique des performances des élèves : ceux qui affichent des faibles performances, des performances moyennes et des bonnes performances. Dans l'analyse bi-variée des rendements des élèves, une répartition des scores en trois tranches a été retenue, en ayant à l'esprit le taux de connaissance de base. Ainsi, ces tranches se présentent comme suit :

- Rendement très faible, correspondant aux scores les plus bas, soit ceux en dessous du fractile de 40% de la distribution des scores ;
- Rendement moyennement faible, correspondant aux scores intermédiaires, soit ceux compris entre les fractiles de 40% et de 75% de la distribution des scores ;
- Rendement faible, correspondant aux scores relatifs aux meilleurs rendements les plus bas, soit ceux en dessus du fractile de 75% de la distribution des scores.

[2] Le taux de connaissance de base se fixe un taux de 40% de bonnes réponses comme seuil, nous avons retenu dans le cadre de l'analyse des résultats du PNEA, un seuil de 50%.

3.2. Les fondements de l'analyse multiniveaux

En éducation comme en sciences sociales, les champs d'observation présentent souvent une structure hiérarchique induisant des effets inter et intra-groupe. Les données afférentes aux acquis des élèves forment une configuration qui renferme plusieurs niveaux emboités les uns dans les autres. Ainsi, les élèves sont sélectionnés dans des classes, qui, elles-mêmes sont nichées dans des écoles. Ces dernières sont sujettes aux décisions des académies régionales, qui traduisent les règlements ministériels ayant trait à la politique éducative mise en vigueur.

L'utilisation d'une telle structure de données remet en question les techniques habituellement utilisées dans la modélisation des relations entre les variables en question, donnant ainsi lieu à une famille de modèles désignés par le vocable « les modèles multiniveaux » ou encore « les modèles hiérarchiques ».

En dépit des diverses discussions et controverses sur les conditions de l'utilisation de ces modèles, la modélisation multiniveaux apporte toutefois des solutions précises permettant de combler les insuffisances qui peuvent avoir lieu suite à l'utilisation des méthodes économétriques classiques telles que les moindres carrées ordinaires ou la régression logistique. Son originalité méthodologique réside dans la représentation plus ou moins intelligible de la réalité sociale. Celle-ci étant toujours en dépendance hiérarchique, le recours aux modèles multiniveaux reste toutefois de mise dans les contextes sociaux, en particulier dans le domaine de l'évaluation en éducation.

Les premières recherches fondatrices des techniques d'analyses multiniveaux remontent aux années 60, avec la genèse des travaux de la bio-statistique sur des expériences répétées, ainsi que des études empiriques sur les données socioéconomiques longitudinales[3]. En réalité, ces dernières ont commencé bien avant avec les essais de Paul Douglas portant sur l'estimation d'une fonction de production dynamique de type Cobb-Douglass aux Etats-Unis[4].

Mais à partir du milieu des années 80, les recherches vont connaître leur plein essor avec les travaux de : Aitkin & Longford ; Raudenbush & Bryk ; Goldstein et Mason et Wong & Entwise. Ces mêmes auteurs continuent toujours à étendre leurs travaux pour couvrir des aspects complexes de l'analyse multiniveaux (les effets non-linéaires, problèmes de l'endogéneité des facteurs contextuels, etc.). Pour de plus amples détails, Searle, Casella et McCulloch (1992) ont prévu un historique intéressant dans leur ouvrage intitulé « Variance Components ».

Dans le cadre de ce rapport technique, il sera question de la modélisation multiniveaux ayant servi à l'explication des principaux déterminants des acquisitions scolaires des élèves dans le cadre du PNEA-2008. Pour ce faire, une justification du choix de tels modèles sera faite à l'aide d'une comparaison entre les modèles hiérarchiques linéaires et les modèles utilisant les moindres carrées ordinaires. Cette comparaison est souvent désignée sous le nom "biais".

Ensuite, la spécification théorique du modèle, ainsi que celle adoptée pour la modélisation des principaux déterminants des acquisitions scolaires des différents cycles concernés par le dispositif d'évaluation seront examinées de manière plus ou moins détaillée.

Enfin, puisque nous avons utilisé des modèles de régressions linéaires avec correction robuste de l'hétéroscédasticité pour tenter d'apprécier les effets-maître et les effets-établissement agrégés, nous présenterons, sommairement, en quoi consiste les estimations par les moindres carrées robustes.

A la différence de la section précédente, la modélisation repose sur l'appariement intégral de tous les fichiers de données du dispositif d'évaluation, c'est-à-dire une fusion par matière et niveau scolaire des questionnaires contextuels et des résultats des élèves aux tests du PNEA. Soit 18 fichiers d'analyse répartis sur les cycles d'enseignement considérés selon les disciplines.

[3] L'on citera à titre d'exemple :
- Elston, R. C. et Grizzle, J. E. [1962]. Estimation of time-response curves and their confidence bands.
- Robinson, W. S. (1950) "Ecological Correlations and the Behavior of Individuals"

[4] Voir aussi les travaux de Chenery, Minhas et Solow en 1961, et ceux de Y.Mundlack dans la même période.

En effet, les différents questionnaires contextuels, comme évoqué, relatent des informations personnelles et professionnelles des acteurs agissant sur les apprentissages des élèves. A ce titre, les données recueillies sont le résultat de déclarations des personnes enquêtées, abstraction faite de la justesse ou de la conformité des attitudes et des perceptions de la réalité étudiée. Il reste entendu qu'un nombre d'opérations de nettoyage et de redressement ont été effectuées pour la cohérence et l'exploitation d'un maximum de données disponibles.

L'utilisation des modèles hiérarchiques linéaires fait appel à des techniques statistiques. Pour cela, le lecteur de ce rapport devrait disposer d'un certain nombre de pré-requis statistiques en matière de régression et d'analyse de données lui permettant de mettre à profit les résultats des modèles et comprendre aisément l'interprétation des estimations.

A la fin de ce rapport, figure une bibliographie accessible des références de base permettant de donner plus de détails techniques et d'approfondir les dimensions diverses relatives à l'emploi des modèles multiniveaux.

3.2.1. Pourquoi une analyse multiniveaux ?

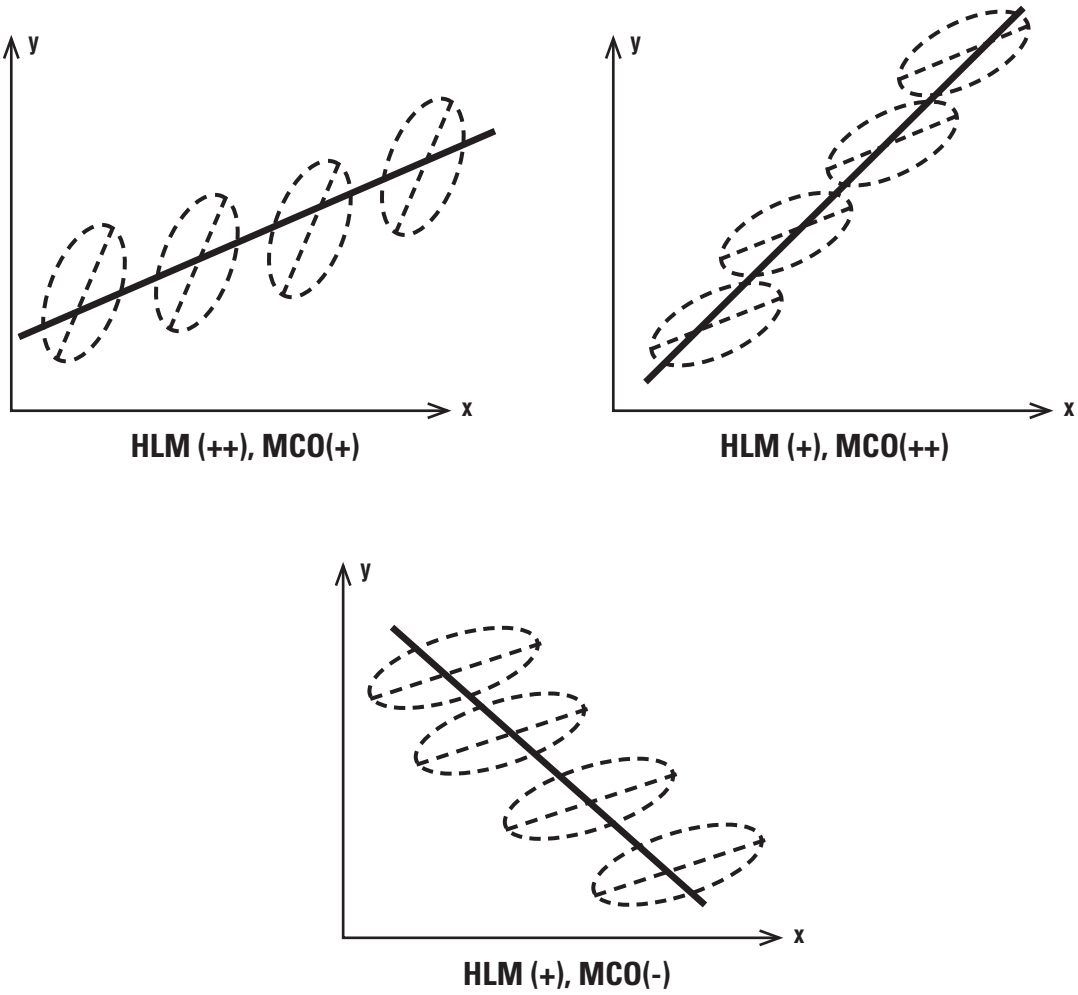
Les travaux d'investigation menés sur les déterminants des apprentissages figurent en bonne place parmi les débats sur les méthodes statistiques idoines pour comprendre les principaux facteurs agissant sur les performances scolaires. Le rapport Coleman[5], le premier travail séminal ayant donné le coup d'envoi à cet exercice, a fait l'objet de nombreuses critiques remettant en question l'utilisation des méthodes statistiques fondées sur les corrélations et la régression directe. A rappeler que Coleman conclut dans son rapport que l'établissement scolaire n'a aucun effet significatif expliquant la variabilité des acquisitions scolaires des élèves et que seuls ces derniers, de par leurs caractéristiques individuelles, sont responsables des différences entre les performances.

Par la suite, plusieurs études empiriques portant sur les données de l'évaluation se sont succédées pour expliquer la qualité des apprentissages des élèves par l'entremise des différentes techniques d'estimation. C'est dire que l'utilisation des modèles multiniveaux est loin de faire l'unanimité entre les chercheurs. Certains chercheurs apprécient les effets des établissements en se basant sur les coefficients de détermination R^2 après contrôle de certaines variables. Cependant, les adeptes des modèles hiérarchiques insistent sur l'existence d'un biais majeur pouvant entraîner des estimations erronées des effets fixes, ce qui pourrait résulter en une interprétation erronée des résultats qui va à l'encontre de la réalité étudiée. Ce biais est le plus souvent désigné sous le vocable Biais écologiques (ecological fallacy)[6]. Le biais écologique résulte, en effet, des études sur des données à plusieurs niveaux emboités. Il peut entraîner une erreur d'estimation du degré d'association entre la variable explicative et la variable expliquée. Ce biais est en général une implication d'un biais d'agrégation qui fait que les facteurs agissants sur les apprentissages des élèves impactent de la même manière l'ensemble des élèves, abstraction faite des spécificités des écoles auxquelles ils appartiennent. Ainsi, les moindres carrées ordinaires considèrent les niveaux d'analyse comme étant un seul niveau. Sans procéder aux tests d'homogénéité ou de spécification, il serait réducteur d'estimer directement les paramètres du modèle. Il devient donc nécessaire d'utiliser une représentation statistique qui prend en considération la configuration hiérarchique des données. Ainsi, les élèves sont contenus dans des classes qui sont à leur tour sélectionnées dans des écoles, lesquelles sont nichées dans des régions.

Le schéma qui suit montre clairement l'ampleur et la nature du biais entre l'utilisation d'un modèle hiérarchique linéaire (HLM) et les moindres carrées ordinaires (MCO). L'erreur est d'autant plus importante que les MCO estiment les paramètres avec un signe inversé comparativement aux HLM. Aussi, dans plusieurs cas, les MCO surestiment ou sous-estiment les paramètres agrégés comme est indiqué dans le schéma suivant :

[5] Rapport de Coleman en 1966 intitulé : « equality of educational opportunity research ».
[6] Kreft et Leeuw [2004] ; Bryk et Raudenbush [1992]

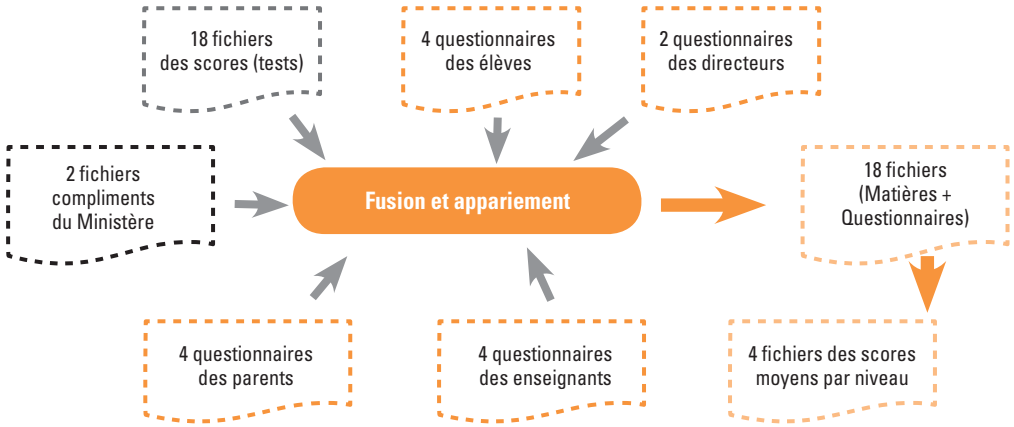
Différences entre les modèles hiérarchique et les modèles à un seul niveau



Préparation des données servant à l'analyse multiniveaux

Les fichiers d'analyse de données ont été constitués à partir d'un certain nombre d'appariements de fichiers de base. D'abord, les 18 fichiers des 7 tests ayant servi au calcul des scores des matières ont été fusionnés avec les huit fichiers correspondant aux caractéristiques sociodémographiques et scolaires de l'élève et de ses parents. Ces fichiers de données ont été ensuite appariés avec les 4 fichiers relatifs aux caractéristiques socioprofessionnelles et pédagogiques des enseignants des quatre niveaux scolaires. Une dernière fusion a été effectuée pour incorporer les deux questionnaires adressés aux directeurs des établissements scolaires du primaire et du secondaire collégial afin de rendre compte des informations sur les caractéristiques personnelles du directeur, les ressources et le fonctionnement de ces établissements. Ainsi, 18 fichiers de données on été reconstitués pour former la plateforme permettant la modélisation par niveau et par matière. Enfin, une dernière fusion a eu lieu et a permis de calculer le score moyen, par niveau, des matières considérées. Soit au total, quatre fichiers de données.

Le schéma ci-après illustre les différentes étapes poursuivies dans le processus d'appariement et de fusion des fichiers.



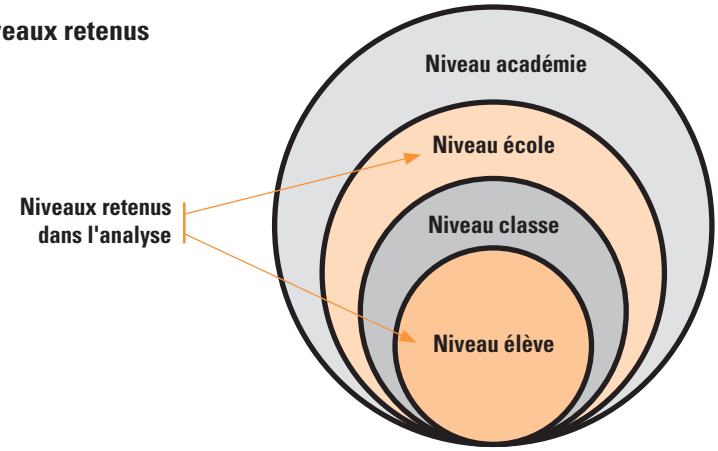
Une fois ces données reconstituées, des opérations d'apurement systématiques des variables d'intérêt ont eu lieu afin de les rendre le plus exploitable possible. On a ainsi, effectué un traitement des aberrations au niveau des variables dont les données connaissent un certain nombre d'erreurs. Certaines de ces erreurs ont été simplement corrigées en ayant recours à une série de données provenant du Ministère, en particulier, celles relatives au soutien social, taille de l'école, nombre de classes, effectif des élèves et des enseignants et bien d'autres informations complémentaires sur l'établissement scolaire.

Aussi, les non-réponses ont été traitées en se fixant un seuil nécessaire pour l'estimation qui est de l'ordre de 10% comme taux de non réponse. En effet, plusieurs techniques permettent de procéder à l'estimation des points manquants, deux d'entre elles ont souvent été utilisées, il s'agit de l'imputation multiple et de l'interpolation linéaire[7].

Compte tenu du plan d'échantillonnage qui a été adopté, seuls deux niveaux hiérarchiques à analyser ont été retenus, il s'agit du niveau école (niveau 2) et du niveau élève (niveau 1). Ceci-étant, le niveau classe se trouve contenu dans le deuxième niveau et par conséquent, l'estimation des effets-établissement peut être augmentée des effets-classe[8].

Le schéma suivant reprend les niveaux possibles et les niveaux retenus dans le cadre de la modélisation des déterminants des acquis des élèves.

Schéma des niveaux retenus



[7] Une observation manquante est imputée lorsque celle-ci est remplacée par une observation valide au moyen des relations entre la variable en question et les autres variables. L'interpolation linéaire établit des bornes pour l'observation manquante à partir desquelles on estime une valeur valide, pas nécessairement exacte pour la non réponse. D'autres moyens empiriques ou par inférence peuvent être utilisés pour l'estimation des observations manquantes.

3.2.2. Formulation du modèle hiérarchique linéaire

Les modèles hiérarchiques, appelés aussi les modèles mixtes ou modèles à coefficients aléatoires, ou encore modèles à variance composée, sont des modèles contenant à la fois des effets fixes et des effets aléatoires. Ils sont une généralisation de la régression linéaire permettant d'inclure des effets aléatoires (inter et intra), autres que ceux associés à l'ensemble du terme d'erreur. En notation matricielle, le modèle s'écrit en général comme suit :

$$Y = X\beta + Zu + \epsilon \quad (1) \quad \text{Avec} \quad E[Y|u] = X\beta + Zu \quad (2)$$

Où y est vecteur n×1 de réponses, X est matrice (n×p) de covariables pour les effets fixes, β et Z est matrice (n×p) covariables pour les effets aléatoires u de moyenne nulle et de variance Ω. Le vecteur n×1 d'erreurs, est supposée suivre une normale multivariée de moyenne nulle et de matrice de variance-covariance $\sigma_\epsilon^2 + I_n$.

La partie fixe du modèle (1), $X\beta$, est analogue à celle d'une régression linéaire multiple avec les moindres carrées ordinaires, appliquées à l'estimation des coefficients du vecteur β. En ce qui concerne la partie aléatoire, nous supposons a priori que le vecteur aléatoire u de matrice à de variance-covariance Ω est orthogonal au vecteur d'erreurs ϵ associées au niveau 1.

De sorte que : $\text{Var}[(u\ \epsilon)'] = \begin{pmatrix} \Omega & 0 \\ 0 & \sigma^2_\epsilon \times I_n \end{pmatrix}$ et $\text{Var}[\epsilon] = R$

Le vecteur de réponse Y suivra alors une loi normale multivariée qui s'écrit :

$$Y \sim N(X\beta, Z\Omega Z' + R)$$

Nous remarquons que les effets fixes interviennent uniquement dans la détermination de la moyenne de Y alors que les effets aléatoires figurent en bonne place dans la variance totale de la variable explicative.

On peut éventuellement procéder à la violation de cette hypothèse en stipulant que les effets aléatoires sont parfaitement corrélés. Le nombre de paramètres à estimer se verra alors augmenter et l'estimation du modèle (1) devra faire l'objet d'équations supplémentaires pour pallier au problème de l'identification.

A noter que le vecteur des effets aléatoires n'est pas directement estimé. On le perçoit à partir des composantes de la matrice Ω connue sous le nom de "variance components" et qui sont estimés avec les erreurs résiduelles σ_ϵ^2.

Les formes générales de la conception des matrices X et Z permettent d'incorporer un large éventail de modèles linéaires selon la réalité étudiée. L'on citera à titre d'exemple: les plans expérimentaux hiérarchisés, les split-plot, les courbes de croissance, les modèles multiniveaux ou hiérarchiques et les modèles de valeur ajoutée. Le modèle (1) est également une autre représentation permettant d'estimer les modèles économétriques sur des données de panel. La matrice Ω permet en effet une grande souplesse selon les objectifs de l'étude, la nature des données, le plan de sondage adopté et la taille des unités dans chaque niveau d'analyse. On pourrait alors opérer des restrictions multiples sur les paramètres dont la qualité d'ajustement, l'ampleur des effets et les éventuelles possibilités de convergence des estimations pourraient entrer en ligne de compte pour juger de la pertinence de telles restrictions. C'est ainsi que l'on peut décider de faire varier la ou/et les pentes du modèle et, le cas échéant, en choisissant les effets à corréler dans la matrice Ω .

Pour la non réponse. D'autres moyens empiriques ou par inférence peuvent être utilisés pour l'estimation des observations manquantes

[8] Il serait possible de considérer le niveau 2 comme étant la classe, mais cela ne changera en rien les estimations, puisque l'on a retenu une classe par école. C'est la raison pour laquelle on a choisi de parler de l'effet établissement plutôt que de l'effet classe vu que ce dernier est inclu dans le précédent.

Pour rappeler les hypothèses essentielles sur lesquelles repose l'analyse des données hiérarchisées, notons surtout celles qui sont les plus associée à ce genre d'analyse, soit :

H 1. L'hypothèse d'additivité prévoit la linéarité de la relation entre la variable expliquée et celle explicative.

H 2. La variable expliquée est censée être continue, non bornée mesurée sans erreur.

H 3. Absence de colinéarité entre les variables explicatives[9].

H 4. Les erreurs du niveau 1 sont identiquement indépendantes et normalement distribuées autour d'une moyenne nulle et une variance constante.

H 5. Les erreurs du niveau 2 possèdent une structure similaire à la matrice ø.

H 6. Les erreurs des deux niveaux 1 et 2 sont indépendantes.

H 7. Absence d'endogénéité impliquant la dépendance entre les variables explicatives et les termes d'erreurs du niveau 1.

H 8. Absence d'endogénéité impliquant la dépendance entre les variables explicatives et les termes d'erreurs du niveau 2.

Plusieurs méthodes statistiques permettent l'estimation des effets fixes et les composantes de la variance du modèle. L'idée sous-jacente d'une éventuelle estimation est initialement attribuée à Henderson(1953) et repose sur l'idée de l'estimation des composantes d'un modèle ANOVA à plan d'expérience non équilibré. Les groupes hiérarchiques sont introduits comme étant des variables muettes dans le modèle de régressions et l'inférence porte sur la comparaison des effets fixes moyennant le test d'hypothèses basé sur l'égalité des paramètres associés aux groupes.

Cependant la méthode ANOVA, qui est une mise en perspective des effets fixes, comporte un certain nombre de faiblesses, entre autres :

- L'estimation des paramètres et des tests y associés devient difficilement réalisable en présence d'un nombre important de groupes au sein de chaque niveau.
- Les estimations ANOVA sur des groupes de tailles réduites se verront déséquilibrées et peuvent conduire à des conclusions erronées.
- Si les groupes découlent des plans de sondage aléatoires, le modèle ANOVA n'est pas le moyen approprié pour comparer les effets étudiés.
- Les effets associés aux groupes ne peuvent pas être séparés si l'on veut étudier également des variables spécifiquement communes à ces groupes[10].

Par ailleurs, les techniques les plus utilisées dans les modèles à effets mixtes sont le maximum de vraisemblance (ML), le maximum de vraisemblance à information limitée(REML), les moindres carrées généralisées (GLS) et les moindres carrées généralisées restrictives(RGLS). Tandis que la vraisemblance repose sur les hypothèses du modèle relatives à la distribution probabiliste des observations, les moindres carrées généralisées, quant à elles, font face à l'hétéroscédasticité due à la variabilité de la variance en fonction des niveaux retenus. La restriction, lorsqu'elle est associée à telle ou telle technique, veut simplement dire que l'on peut imposer des contraintes linéaires qui ne dépendent guère des effets fixes, mais plutôt des différentes composantes de la matrice de variance-covariance. Par exemple, l'on peut appliquer les algorithmes d'estimation compte tenu de ces contraintes linéaires (Thompson 1962). En effet, il existe plusieurs algorithmes servant à l'estimation des paramètres reposant le plus souvent sur une approximation par itérations successives en utilisant soit la méthode Raphson-Newton, ou Expectation-Maximisation (EM algorithme). Le lecteur pourra trouver

[9] Sinon une estimation moyennant les estimateurs de Ridge peuvent pallier à ce problème.

[10] Searle, Casella, and McCulloch 1992.

des approfondissements techniques détaillés sur l'algorithme(EM) dans l'article de Dempster et al.(1977) intitulé « Maximum likelihood from incomplete data via the EM algorithm ».

Explicitement, si l'on choisit d'estimer le modèle en maximisant la vraisemblance, des développements statistiques permettent d'estimer les paramètres du modèle à variance (V) inconnue comme suit :

- Au niveau des effets fixes : $ML(X\beta) = X\hat{\beta} = X(X'\hat{V}^{-1}X)^{-1}\ X'\hat{V}^{-1}Y$
- Au niveau des effets aléatoires : $\hat{u} = \hat{D}^{-1}Z'\hat{V}^{-1}(y - X\hat{\beta})$

  Ici, il est important d'insister sur le fait que le vecteur aléatoire « u » ne constitue pas un paramètre à estimer au sens statistique du terme, seuls les effets aléatoires reflétés dans les composantes de la matrice ø sont à estimer.

Habituellement, on peut étudier les propriétés statistiques des estimateurs et éventuellement des intervalles de confiance permettant les inférences nécessaires. Cependant, et pour des raisons de commodité et pour ne pas surcharger ce rapport de formulations souvent lourdes, nous nous limitons à indiquer l'estimation des paramètres tels que prévu dans ce genre de modèles[11].

Pour simplification, l'écriture scalaire du modèle s'articule autour de l'équation suivante :

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \cdots + \beta_{k-1j}x_{ij} + \epsilon_{ij}$$

$$(2) \qquad \begin{aligned} \beta_{0j} &= \beta_0 + u_{0j} \\ \beta_{1j} &= \beta_1 + u_{1j} \\ &\vdots \quad \vdots \quad \vdots \\ \beta_{k-1j} &= \beta_{k-1} + u_{k-1j} \end{aligned}$$

$$\text{Avec} \quad \epsilon_{ij} \overset{iid}{\rightarrow} N(0, \sigma^2{}_{\epsilon}) \ et \ u_{kj} \sim N(0, \Omega)$$

Ainsi les paramètres à estimer dans le système (2) sont les paramètres du vecteur β_0, soit les $\sigma^2{}_{uk}, \sigma_{u_{kl}}$, ainsi que l'erreur résiduelle $\sigma^2{}_{\epsilon}$ du niveau 1.

Plus on ajoute de paramètres et de niveaux, plus le nombre d'observations doit être suffisant pour permettre la convergence des estimations. En effet, plusieurs auteurs s'accordent sur la difficulté empirique de faire varier à la fois les pentes et la constante du modèle tel qu'indiqué en (2). D'après les travaux de Kreft et Raudenbush(1997), les modèles les plus courants adoptent une configuration où généralement une constante et une pente avec ou sans corrélation sont introduites comme fonction du niveau hiérarchique supérieur. Cette configuration se présente comme suit :

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \cdots + \beta_{k-1j}x_{ij} + \epsilon_{ij}$$

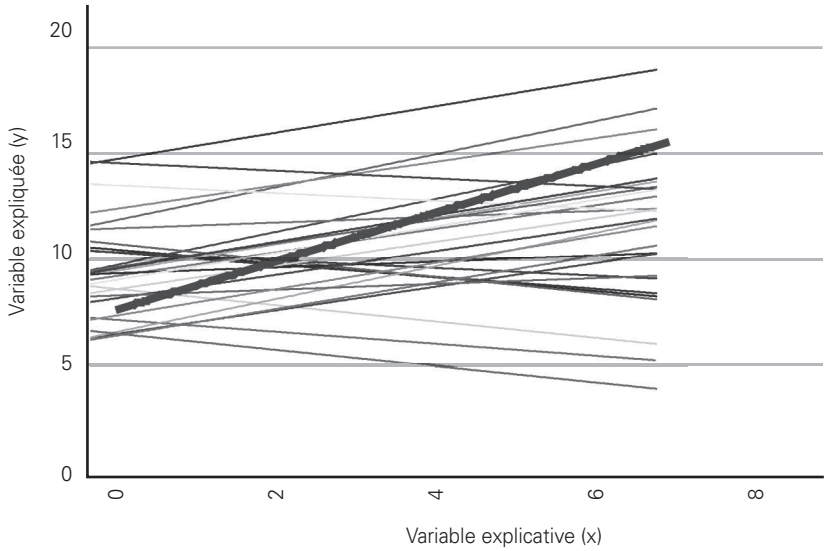
$$(3) \qquad \begin{aligned} \beta_{0j} &= \beta_0 + u_{0j} \\ \beta_{1j} &= \beta_1 + u_{1j} \\ \beta_{2j} &= \beta_2 \\ &\vdots \quad \vdots \quad \vdots \\ \beta_{k-1j} &= \beta_{k-1} \end{aligned}$$

$$\text{Avec } \epsilon_{ij} \overset{iid}{\rightarrow} N(0, \sigma^2{}_{\epsilon}) \ , u_{rj} \sim N(0, \sigma^2{}_{ur}) \forall r \epsilon \{0,1\} \ et \ cov(u_{0j}, u_{1j}) \neq 0$$

[11] De Leeuw, J., & Meijer, E.. (2008). Handbook of Multilevel Analysis.

Cette spécification paraît clairement dans la figure suivante :

Figure 1. : Modélisation multiniveaux d'une variable expliquée Y en fonction d'une variable explicative X à pente aléatoire.



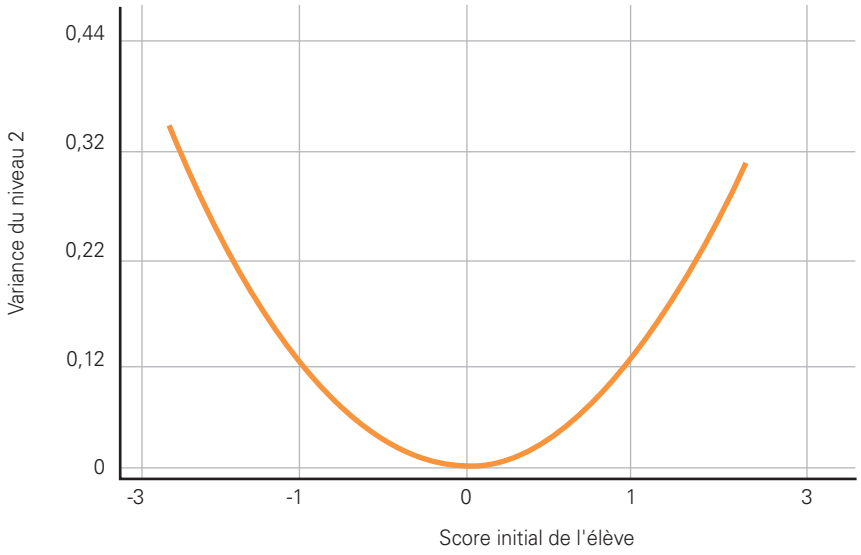
Les modèles à pente aléatoire permettent un autre type de modélisation qu'est la décomposition de la variance totale en fonction, notamment, de l'interaction entre les niveaux et les variables explicatives. Si l'on prend, à titre d'exemple, la spécification (3), la variance totale du modèle peut être décomposée sous la forme suivante :

$$Var(y_{ij}|x_{ij}) = \sigma^2{}_{\epsilon} + \sigma^2{}_{u0} + \sigma^2{}_{u1} * x^2{}_{ij} + 2 * \sigma_{u01} * x_{ij}$$

$$\text{où} \qquad \sigma_{u01} = cov(u_{0j}, u_{1j})$$

Il est ainsi clair, de par cette spécification, que la variance du niveau 2 s'explique comme étant une relation quadratique de x. Cette modélisation peut être représentée dans ce graphique :

Figure 2. : Modélisation de la variance du niveau 2 en fonction du score initial de l'élève



La spécification qui a été envisagée dans le cadre du rapport analytique est celle dont seule la constante est aléatoire. Plusieurs essais permettant de tester la pertinence des pentes aléatoires ont été réalisés, soit 18 modèles correspondant aux matières étudiées. En général, seule la pente associée aux conditions socioéconomiques des élèves a donné des résultats concluants et ce, dans certaines disciplines enseignées. Pour cette raison et pour permettre la simplification de l'interprétation et l'uniformisation des modèles relatifs aux différentes disciplines couvertes par le PNEA-2008, nous nous sommes limités à un seul effet aléatoire associé à la constante pour indiquer l'ampleur des effets du niveau 2, soit les effets-établissement.

Cette spécification se présente comme suit :

$$(4)\qquad \begin{aligned} y_{ij} &= \beta_{0j} + \beta_{1j}x_{ij} + \cdots + \beta_{k-1j}x_{ij} + \epsilon_{ij} \\ \beta_{0j} &= \beta_0 + u_{0j} \\ \beta_{1j} &= \beta_1 \\ &\vdots \quad \vdots \quad \vdots \\ \beta_{k-1j} &= \beta_{k-1} \end{aligned}$$

Avec $\epsilon_{ij} \overset{iid}{\rightarrow} N(0,\sigma^2{}_\epsilon), u_{0j} \overset{iid}{\rightarrow} N(0,\sigma^2{}_{u0})$ et $cov(u_{0j},\epsilon_{ij}) = 0$

Par voie de conséquence $y_{ij}|u_{0j} \overset{iid}{\rightarrow} N(\beta_0 + u_{0j}, \sigma^2{}_\epsilon)$

Le modèle vide :

Le modèle vide, dit aussi inconditionnel, revêt une importance particulière dans la modélisation multiniveaux. D'ailleurs, c'est la première étape à parcourir avant de procéder à une telle modélisation. Il permet ainsi de justifier la pertinence du choix d'une modélisation hiérarchique plutôt qu'agrégée.

Le modèle vide correspond à une écriture sans variables explicatives, il s'agit en fait d'une représentation ANOVA permettant de révéler le niveau moyen des acquis ainsi que l'existence des différences significatives entre les établissements scolaires. Ce modèle s'écrit :

$$(5)\qquad \begin{aligned} y_{ij} &= \beta_{0j} + \epsilon_{ij} \\ \beta_{0j} &= \beta_0 + u_{0j} \end{aligned}$$

Avec $\epsilon_{ij} \overset{iid}{\rightarrow} N(0,\sigma^2{}_\epsilon), u_{0j} \overset{iid}{\rightarrow} N\left(0,\sigma^2{}_{u_0}\right)$ et $cov(u_{0j},\epsilon_{ij}) = 0$

β_0 représente la moyenne générale de la population des élèves, ϵ_{ij} est l'écart intra-niveau et u_{0j} l'écart inter-niveaux, c'est-à-dire par rapport à la moyenne générale β_0 .

Après avoir estimé les paramètres, le modèle permet de calculer plusieurs coefficients tels que par exemple : Le coefficient intraclasse, le coefficient de fidélité du modèle, les pseudos R^2, le coefficient de rétrécissement,etc.

Le coefficient intraclasse est d'une importance capitale dans les modèles multiniveaux. Après avoir testé la significativité des effets aléatoires, il permet de mesurer l'ampleur de ces effets et partant, d'apprécier la variabilité des performances des élèves d'un établissement à l'autre. Ce coefficient se calcul comme suit :

$$\rho(y_{i,j}, y_{ij}) = \frac{\sigma^2{}_{u0}}{\sigma^2{}_{u0} + \sigma^2{}_\epsilon}$$

$$= \frac{\text{Variance interniveaux}}{\text{Variance totale}}$$

Une autre manière d'interprétation de ce coefficient serait de le considérer comme étant le degré de dépendance entre deux unités d'observation (élèves) aléatoirement choisies au sein d'un même groupe (établissement). C'est aussi le pourcentage de la variabilité des acquis qui peut être attribuée à l'établissement. Notons que ñ est par construction positif et varie à 0% et 100%.

Ainsi, le modèle vide donne un profilage appréciant de prime abord l'étendue des effets-établissement et des effets-élève sur les acquisitions scolaires. Le modèle vide ne permet pas d'analyser le détail de ces effets. Pour ce faire, il faudra procéder à leur décomposition selon les variables caractérisant les élèves et le cadre organisationnel et pédagogique ainsi que la composition scolaire et sociale des établissements dite en terme anglais « school mix » (Thrupp, 1999).

Le modèle complet : à la recherche d'une « causalité »

Une fois l'estimation et les tests associés ont été faits sur le modèle vide, la deuxième étape consiste simplement à ajouter les variables explicatives à la spécification à laquelle on s'intéresse.

A ce niveau se pose évidemment la question de causalité entre la variable explicative et la variable expliquée. Contrairement aux sciences expérimentales, cette question soulève toujours des controverses méthodologiques redoutables entre les chercheurs en sciences sociales et surtout entre les sociologues. Il s'agit précisément de l'attribut que l'on donne à un exercice de modélisation visant toute schématisation rationnelle des faits sociaux non facilement réductibles à des mises en calcul. Comme évoqué par plusieurs sociologues, s'agit-il d'une explication de fait ou encore d'une quête de « causalité probable » (au sens de Bourdieu)?[12]. D'autres chercheurs vont jusqu'à des essais d'assimilation ayant pour but de se rapprocher le plus de la logique expérimentale fondée sur le contrôle des conditions de l'observation, la répétition de l'expérience et la comparaison des résultats. Cette démarche est à juste titre l'apanage de tout un courant économétrique fondé sur l'économétrie des données expérimentales[13].

En tout état de cause, l'objectif ultime recherché dans notre cas n'est pas celui de relever les causalités au sens poppérien du terme, mais de simplement dégager l'ensemble des facteurs qui concourraient à l'explication des acquis scolaires des élèves. Pour ce faire, nous n'avons cherché à introduire que les variables[14] dont l'importance a fait l'objet de beaucoup d'attention dans la littérature, au vu de l'information cruciale qu'elle véhicule par rapport à la variable étudiée. Cette recherche documentaire a été amplement discutée dans le rapport analytique. C'est la raison pour laquelle nous n'y revenons pas dans le cadre de ce rapport qui se veut en premier lieu méthodologique.

Les variables explicatives utilisées dans l'investigation des déterminants des acquis scolaires forment trois blocs, à savoir les caractéristiques de l'élève, de l'enseignant et celles de l'établissement. La modélisation poursuivie repose sur l'intégration progressive de chaque bloc. Ainsi, une fois le modèle vide estimé, on intègre le bloc des variables individuelles, auquel s'ajoutera, dans une seconde étape, le bloc des variables enseignantes, et enfin les variables du troisième bloc seront introduites dans le modèle final contenant ainsi tous les blocs des variables explicatives.

[12] Notamment dans les travaux de Raymond Boudon et de Pierre Bourdieu.

[13] Pour des récits plus détaillés, voir notamment Morgan (1990) et Pirotte(2004)

[14] Il est de coutume d'incorporer autant de variables explicatives que nécessaire afin de maximiser le pouvoir prédictif du modèle, ce qui pourrait entraver la convergence des estimations dans le cadre de données hiérarchisées ou encore empêcher l'utilisation des variables supplémentaires à cause de problèmes de colinéarité.

Le pouvoir explicatif du modèle

Le pouvoir explicatif, dans le cadre des modèles hiérarchiques, est légèrement différent de celui que l'on rencontre dans le cas des modèles MCO. Néanmoins, il s'agit du même principe impliquant l'utilisation d'un indicateur qui se rapproche du coefficient de détermination R^2. Cet indicateur est le plus souvent appelé pseudo R^2. Il est ainsi calculé pour chaque niveau hiérarchique en prenant simplement la différence de la variance du modèle vide et la variance résiduelle rapportée à la variance du modèle vide. Autrement dit :

$$\text{Pseudo } R^2 = \frac{\text{Variance du modèle vide - Variance résiduelle}}{\text{Variance du modèle vide}}$$

Ce rapport permet d'apprécier le gain de variance expliquée suite à l'introduction des variables explicatives ou supplémentaires.

Par ailleurs, le pseudo R^2 peut s'avérer problématique dans certains cas, en raison du fait que la variance résiduelle pourrait augmenter par rapport au modèle vide. Les recherches économétriques en matière de données hiérarchiques explorent toujours les pistes possibles permettant de pallier une telle insuffisance[15].

Dans le cas des modèles à pentes aléatoires, le calcul du pouvoir explicatif devient de plus en plus complexe et la probabilité d'augmentation de la variance résiduelle s'accentue. L'une des solutions proposées consiste simplement en la modélisation de la variance de chaque niveau en fonction des variables explicatives dont la perte est typiquement aléatoire.

Au bout du compte, il faut noter que le pouvoir explicatif devient de plus en plus compliqué à mesure que l'on remonte dans les niveaux hiérarchiques et que les pentes aléatoires (corrélées ou non) sont nombreuses dans le modèle. C'est aussi l'une des raisons pour lesquelles nous avons adopté une spécification simple avec une constante aléatoire uniquement.

Les tests d'hypothèses dans le cadre des analyses multiniveaux

Les tests portant sur les paramètres :

Comme tout autre modèle statistique, les paramètres estimés doivent être sujets à des tests de significativité. En effet, les paramètres des modèles multiniveaux sont de deux sortes : ceux des effets fixes et ceux associés aux effets aléatoires. En ce qui concerne les effets fixes, le test se fait pratiquement comme en régression classique par les MCO. Moyennant l'hypothèse H4 qui prévoit la normalité des observations, le rapport des coefficients estimés et les erreurs standards (b/s) sont comparés au fractile de la loi de Student au seuil de 5%, soit 1,96. Si b/s est supérieur à 1.96, on rejette l'hypothèse de la nullité des paramètres.

En ce qui concerne les effets aléatoires reflétés dans les variances associées aux différents niveaux, on procède généralement de la même manière, qui consiste à rapporter le coefficient à son erreur standard et on le compare de nouveau avec la valeur 1,96. Si la différence est positive, le test sera alors significatif au seuil de 5%. A noter que ce procédé bénéficie, non pas d'une statistique de test déterminée analytiquement, mais plutôt du théorème central limite qui fait que le quotient converge asymptotiquement vers une loi normale lorsque la taille de l'échantillon du niveau supérieur est suffisamment grande. Cependant, cette proposition figure également parmi les discussions portant sur sa pertinence statistique.

Les tests portant sur l'ensemble des paramètres ou tests de significativité globale :

Dans ce qui a précédé, nous avons présenté une logique permettant de tester chaque paramètre séparément, sans prendre en considération les autres paramètres pris en commun.

[15] Voir Snijders et Bosker (1994) cité par Pascal Bressoux.

A cet effet, les statisticiens ont développé une mesure similaire à la statistique de Fisher utilisée dans les régressions par les MCO. Cette mesure appelée « la déviance» permet d'assurer la pertinence de l'introduction des variables supplémentaires, ce qui est à même de comparer les estimations entre différents modèles.

Analytiquement, la déviance se définit comme étant la différence entre la vraisemblance de deux modèles. Si L1 et L2 sont respectivement la vraisemblance des modèles 1 et 2, avec L2, le modèle contenant des variables supplémentaires, la déviance D est définie comme suit :

$$D = -2 * \log(L1) - 2 * \log(L2)$$

On prouve que cette quantité suit une loi de chi2 avec m degré de liberté égal au nombre de variables supplémentaires. En comparant la valeur D avec le fractile de la loi de chi2(m), on rejette ou non l'hypothèse de la pertinence de rajouter des variables supplémentaires par rapport au modèle initial.

D'autres statistiques alternatives à la déviance on été mises en place par les statisticiens. Il s'agit de Akaike information criterion(AIC) et Bayesian criterion information(BIC). Ces deux critères, qui utilisent aussi la vraisemblance, renferment l'avantage de prise en compte à la fois de la taille de l'échantillon et du nombre de paramètres à estimer.

La quasi totalité des logiciels statistiques incorporant un module d'estimation des données hiérarchisées permettent de calculer la déviance, le AIC et le BIC. Ils donnent aussi la statistique de Student, la p-value et l'intervalle de confiance, notamment AML, LISREL, SAS, SPSS et STATA. Le logiciel MTWIN fait l'exception puisqu'il ne fournit que les paramètres estimés et leurs erreurs standards, ce qui revient à forcer l'utilisateur à construire ses propres tests et intervalles de confiance.

Est-il nécessaire de faire usage de l'analyse multiniveaux ?

L'analyse multiniveaux est de plus en plus utilisée dans les domaines présentant une structure hiérarchique. Ainsi, les économistes, les démographes, les sociologues, etc font appel à un large éventail de catégories de modèles hiérarchiques (modèles longitudinaux, cross section data, modèles croisés, modèle de valeur ajoutée, etc.) leur permettant d'analyser en profondeur les effets imbriqués, contrairement à une démarche, somme toute, fondée sur des comportements agrégés et identiques quelque soient les groupes étudiés.

Cependant, les chercheurs ont longuement débattu sur les critères qu'il faut prendre en considération pour entreprendre une bonne analyse multiniveaux. Il est question aussi d'identifier la démarche à adopter, si les résultats de ces modèles sont opposés à ceux estimés par les moindres carrées ordinaires. Les réponses que l'on avance généralement à ces préoccupations portent essentiellement sur le coefficient intraclasse et le nombre d'unités d'observation figurant à chaque niveau hiérarchique. En effet, il serait judicieux d'abandonner l'usage des modèles multiniveaux dans les cas suivants :

- Lorsque le coefficient intraclasse est très faible en raison de différences non significatives entre les groupes.
- En présence d'un plan de sondage compliqué (voire déformé) et ne présentant pas clairement la structure hiérarchique.
- Lorsque le nombre de groupes est faible et que celui des observations est important.
- Lorsque l'information contenue dans les niveaux hiérarchiques supérieurs est jugée peu pertinente et que l'on s'intéresse surtout à des effets fixes du modèle. Dans ce cas, l'estimation agrégée par MCO pourrait être préférable à une régression multiniveaux.

Il faut noter que le recours à l'estimation par les moindres carrées ordinaires est peu solide en présence de l'hétéroscédasticité due à l'inconstance de la variance, ce qui est de nature à produire des intervalles de confiance plus larges en sous-estimant les erreurs types des estimateurs. Il existe plusieurs procédures statistiques permettant de pallier au manque de robustesse des estimations et d'effectuer des tests permettant d'obtenir des erreurs-type corrigées. L'une des techniques utilisées repose sur la reconfiguration de la matrice de variance-covariance en introduisant la matrice robuste de White[16].

La formulation du modèle linéaire multiple se présente comme suit : $Y = X\beta + \epsilon$

Une fois que l'on a estimé le vecteur de paramètres par les moindres carrées ordinaires, on procède par la suite au calcul de la matrice robuste de White, donnée par la formule suivante :

$$\widehat{Var(\hat{\beta})} = N(X'X)^{-1}S(X'X)^{-1} \quad où \quad S = \frac{1}{T}\sum_{i=1}^{N} \widehat{e^2}_i x_i x'_i$$

La majorité des logiciels statistiques produisent des options permettant de procéder à des calculs robustes des intervalles de confiance en présence de l'hétéroscédasticité. Voulant apprécier l'ampleur des effets fixes, notamment dans la modélisation de l'impact des caractéristiques de l'enseignant et de l'établissement sur les apprentissages des élèves, nous avons procédé dans le rapport analytique à l'estimation de certains modèles par l'entremise des moindres carrées robustes. Cela a permis également de voir dans quelle mesure le signe et la significativité des estimations sont divergents par rapport aux estimations du modèle multiniveaux.

[16] Pour un récit détaillé voir l'article séminal « a Heteroscedasticity-Consistent Covariance Matrix Estimator and Direct Test of Heteroscedasticity » d'Harold White (1980).

CONCLUSION

Les analyses multiniveaux revêtent une importance capitale dans le domaine de l'éducation. C'est un moyen efficace permettant de saisir la multiplicité des effets lorsqu'ils sont emboîtés les uns dans les autres. De plus, les estimations produites par ces analyses donnent des estimations plus précises en comparaison avec les méthodes classiques de régression linéaire. Grâce à une telle modélisation, il est possible de montrer l'existence considérable d'effets-établissement dans le contexte marocain, induisant qu'il existe d'importantes différences dans les acquis scolaires entre les établissements.

Au terme de ce rapport technique, le dispositif technique qui a été mobilisé pour la mise en œuvre de l'évaluation dans le cadre du PNEA-2008 a été passé en revue. A cet effet, nous avons détaillé la méthodologie d'échantillonnage, la construction des tests, les méthodes de validation psychométrique, la logique qui sous-tend l'analyse bi-variée, ainsi que les fondements de base de la modélisation multiniveaux ont été détaillés.

Ce rapport se veut un guide méthodologique global et permettra aux lecteurs de comprendre les soubassements scientifiques derrière la production des résultats tant au niveau des fascicules qu'au niveau du rapport analytique.

ANNEXE BIBLIOGRAPHIQUE

Aitken, L.S, &West, S.G. (1991). Multiple regression testing and interpreting interactions. Newbury Park, London : Sage.

Boufrahi, S., Arseneau, M.N& Robin, R. (2003). Les facteurs clés de la réussite scolaire au primaire. Université du Québec à Montréal.

Bressoux, P. (1995). Les effets du contexte scolaire sur les acquisitions des élèves : effet-école et effets-classes en lecture, Revue française de sociologie(1995).

Bressoux, P. (1994). Les recherches sur les effets-écoles et les effets-maîtres, Revue française de pédagogie, no 108, juillet-août-septembre 1994.

Bressoux, P. (2008). Modélisation statistique appliquée aux sciences sociales , De Boeck, coll. Méthodes en sciences humaines, Bruxelles.

Bryk, A.S. & Raudenbush, S.W. (2000). Hierarchical linear model: applications and data analysis methods, 2e edition, SAGE publication, London.

Courgeau, D. (2007). Multilevel Synthesis: From The group to the Individual, Springer. Germany.

Cousin, O. (1993). L'effet-établissement. Construction d'une problématique. Revue française de sociologie, vol.34, 1993.

Goldstein, H. (1987). Multilevel Models in Educational and Social Research. London, Griffin; New York, University Press.

Goldstein, H. (1999). Multilevel statistical models, Institut of Education, London.

Hox, J. (2002). Multilevel Analysis. Techniques and applications. LEA publishers, New Jersey.

Laveault, D.& Grégoire, J. (2008). Introduction aux theories des tests. De Boeck, coll. Méthodes en sciences humaines, Bruxelles.

Leeuw.J, Meijer, E. (2008). Handbook of Multilevel Analysis, New York, Springer science & Business Media, LLC.

O'connell, A& McCoachy, D.B. (2008). Multilevel Modelling of Educational Data. IAP, Charlotte, USA.

Snijders, T & Bosker, R. (1999). Multilevel analysis. An introduction to basic and advanced multilevel modeling, London, Sage.

S. Rabe-Hesketh and A. Skrondal. (2005). Multilevel and Longitudinal Modeling Using Stata. Stata Press, College Station, TX.

StataCorp. Stata Statistical Software: Release 10. Stata Corporation, College Station, TX, 2008.