RÉPUBLIQUE FRANÇAISE
*Liberté*
*Égalité*
*Fraternité*

depp Direction de l'évaluation, de la prospective et de la performance

# Méthodologie utilisée dans les évaluations nationales du second degré

## Analyse psychométrique

Version du 21/01/22

DEPP B2-2

Série Méthodes

Document de travail n° 2022-M01

Janvier 2022

# Méthodologie utilisée dans les évaluations nationales du second degré

Analyse psychométrique

Cet ouvrage est édité par le ministère de l'Éducation nationale, de la Jeunesse et des Sports

Direction de l'évaluation, de la prospective et de la performance

61-65, rue Dutot

75732 Paris Cedex 15

Directrice de la publication

Fabienne Rosenwald

DEPP B2-2

ISBN / e-ISBN

ISSN : 2779-3532

# Table des matières



# Introduction

Depuis 2017, la Depp a mis en place des évaluations de français et de mathématiques pour tous les élèves entrant en classe de sixième, puis en 2018 pour tous les élèves entrant en classe de seconde.

Les tests standardisés sont développés avec des méthodes d'analyse psychométrique afin d'assurer la comparabilité des résultats dans l'espace (entre établissements, départements et

régions) et dans le temps (entre les générations). Les théories psychométriques fournissent des outils pour, en amont des passations [1], créer des évaluations pertinentes et équilibrées, et en aval, attribuer à chaque élève un niveau de compétence selon ses réponses à l'évaluation.

Ce document porte sur le deuxième volet : l'estimation des compétences des élèves à partir de leurs réponses aux questions de l'évaluation. La méthodologie d'analyse des évaluations nationales sera explicitée de la réception des données de passation de l'ensemble des élèves en classe de sixième et de seconde, à la consolidation des résultats individuels aux épreuves.

Avant de présenter la méthodologie, il faut rappeler le cadre théorique choisi pour les évaluations nationales : le modèle de Rasch à deux paramètres. Au lieu de calculer un score brut [2] pour chaque élève, les Modèles de Réponse à l'Item (MRI) permettent de positionner les élèves sur une échelle de compétence en fonction de la difficulté des questions auxquelles ils ont répondu juste. Chaque question $j$, appelée *item*, est ainsi modélisée par deux paramètres :

— la pente $a_j$ qui indique sa capacité à discriminer les élèves selon leur niveau de compétence ;

— la difficulté $b_j$ qui correspond au niveau de difficulté de la question.

Avec ces deux paramètres, il est pertinent de modéliser la probabilité qu'un élève $i$ avec un niveau de compétence $\theta_i$ réponde juste à l'item $j$ (i.e. $Y_{ij} = 1$) par la formule suivante :

$$\mathbb{P}(Y_{ij} = 1|\theta_i, a_j, b_j) = \frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}}$$

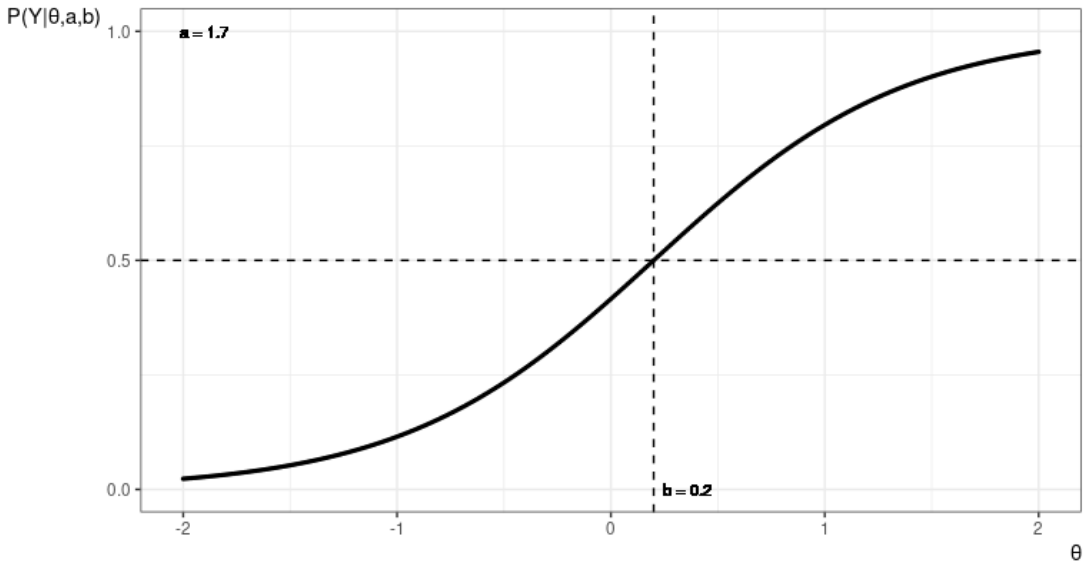


FIGURE 1 – Probabilité de réussite d'un item en fonction de la compétence des élèves

L'objectif de l'analyse psychométrique est d'estimer, à partir des réponses des élèves, les paramètres de difficulté des items $(a_j, b_j)_j$ ainsi que les compétences des élèves $(\theta_i)_i$. En pratique, les fonctions du package *TAM*, pour *Test Analysis Modules*, disponible sour *R*,

1. On nomme passations les évaluations passées par les élèves.
2. Le score brut est le nombre de bonnes réponses divisé par le nombre de questions.

sont mobilisées au cours de l'analyse. Ces fonctions permettent de déterminer les paramètres qui maximisent la vraisemblance des données observées, en supposant une distribution normale centrée et réduite des compétences des élèves. Les $(a_j)_j$ sont positifs (généralement dans l'intervalle $[0; 1, 7]$), et les $(b_j)_j$ sont sur la même échelle que les $(\theta_i)_i$.

# 1 Préparation des données

La première partie de l'analyse consiste à recueillir les données des passations des élèves et de les préparer pour la suite. Cette partie se découpe en trois phases.

## 1.1 Nettoyage initial

En premier lieu, le traitement débute par la vérification de la conformité des données de passations. Puis, plusieurs filtres sont effectués afin de ne conserver que les données pertinentes.

Sont supprimés notamment :

— les exercices d'exemple et d'entrainement ;

— les passations qui n'ont pas été effectuées par des élèves ;

— les lignes élève/item en doublon (la ligne avec le marqueur temporel le plus récent est conservée).

## 1.2 Suppression des non-réponses terminales

L'analyse psychométrique permet d'évaluer la compétence des élèves en fonction de leurs réponses aux questions et de la difficulté de celles-ci. Il n'est pas nécessaire qu'un élève ait répondu à toutes les questions du test pour estimer son niveau, tant que le nombre de réponses est suffisant.

Lors des évaluations sur support numérique, les élèves ont la possibilité de passer des questions, mais ils ne peuvent pas revenir en arrière. Par conséquent, il existe deux types de non-réponses :

— les réponses vides aux questions que l'élève a vues et a passées

— les réponses vides aux questions en fin d'évaluation que l'élève n'a pas eu le temps de voir

Pour les premières, on considère que l'élève n'a pas répondu parce qu'il ou elle ne connaissait pas la bonne réponse, et la non-réponse est traitée comme une mauvaise réponse. En revanche, pour les dernières, on considère que l'élève ne les a pas vues, et elles sont retirées des données comme si elles ne faisaient pas partie de l'évaluation de l'élève.

C'est pourquoi la deuxième partie de la préparation des données consiste à identifier puis supprimer les non-réponses terminales. Comme les questions sont présentées à chaque élève selon un ordre aléatoire, il faut classer les questions vues par l'élève selon l'ordre de passage grâce aux marqueurs temporels, et supprimer toutes les réponses vides qui ne sont suivies d'aucune réponse non vide.

Il est possible qu'une fois les non réponses terminales retirées, certaines passations ne comportent plus qu'un très petit nombre d'items, sur lesquels il est difficile d'évaluer avec précision le niveau de l'élève dans la discipline. Dans les évaluations nationales, le choix est

fait de conserver ces passations tout au long de l'analyse, et de regarder en détail à la fin les scores attribués à ces passations parcellaires.

### 1.3 Vérification des corrections automatiques

La dernière partie de préparation des données consiste à vérifier que les corrections automatiques attribuées aux réponses des élèves sont justes.

## 2 Vérification de la validité du modèle

Une fois les données nettoyées, l'analyse psychométrique peut commencer. L'objectif premier est d'attribuer à chaque élève un score qui indique son niveau de compétence dans la matière. Ce score doit être comparable à ceux de ses camarades de l'année, ainsi qu'aux scores des élèves des années précédentes.

Cette première partie de l'analyse a pour objectif de repérer les items qui ne mesurent pas correctement la compétence visée afin de les écarter par la suite.

L'analyse psychométrique se place dans un cadre théorique qui suppose que tous les items mesurent le même trait latent et de la même manière quels que soient les groupes de sujets (au sein d'une cohorte d'élèves ou entre cohortes). Plusieurs études doivent être menées afin de s'assurer de la pertinence du choix de ce modèle.

### 2.1 Analyse sur les construits unidimensionnels

Tout d'abord, il s'agit de tester l'hypothèse d'unidimensionnalité. L'hypothèse d'unidimensionnalité stipule que les réponses données par les élèves aux items d'une évaluation sont la manifestation d'une dimension latente, unique et continue. Autrement dit, la probablité de réussir un item ne dépend que de la position des élèves sur un unique trait latent $\theta$ (Rocher, 2013).

Des études sur ce point sont menées en amont des évaluations, à partir des résultats des années précédentes. Alors que l'on observe que les évaluations de mathématiques mesurent une compétence unidimensionnelle, ce n'est pas tout à fait le cas en français. En effet, les questions relatives à la compréhension de l'oral semblent évaluer une compétence sensiblement différente de celle de la compréhension de l'écrit ou du fonctionnement de la langue. Ces observations nous conduisent à traiter à part la compréhension de l'oral en français.

On choisit donc d'effectuer cette première partie de l'analyse sur chaque groupe de questions qui forme un construit unidimensionnel [3]. La composition des groupes est donnée a priori à partir d'études réalisées sur les années antérieures. Néanmoins, il s'agira par la suite de conforter la composition de ces construits unidimensionnels par une analyse des dépendances locales.

Pour chaque construit unidimensionnel, les étapes suivantes sont déroulées :

1. Suppression des items avec un r-bis (défini plus bas) trop faible ;
2. Détection des items en fonctionnement différentiel.

3. Un construit unidimensionnel est un groupe d'items qui mesurent un trait latent unique et continu.

A la fin de l'analyse, la validité du construit unidimensionnel sera vérifiée. Pour cela, il s'agit de tester l'hypothèse d'indépendance locale. Sous l'hypothèse d'indépendance locale, la réussite à un item est statistiquement indépendante de la réussite aux autres items, conditionnellement à la compétence $\theta$ de l'élève. Afin de détecter les éventuelles dépendances locales, l'ajustement du modèle est calculé, puis il faut vérifier que les dépendances entre items sont inférieures à un seuil fixé à $0,2$.

## 2.2 Suppression des items avec un r-bis trop faible

L'objectif est d'écarter les questions qui mesurent mal la compétence visée.

Le r bisérial point, nommé ici plus simplement r-bis, permet d'évaluer à quel point un item mesure la même compétence que les autres questions (Rocher, 2013). Plus précisément, c'est le coefficient de corrélation entre $Y_j$ la variable indicatrice de réussite à l'item $j$ (1 si l'élève répond juste et 0 sinon) et $S$ le score de l'élève sur l'ensemble du construit :

$$r_j^{bis} = corr(Y_j, S) = corr(Y_j, \frac{1}{m}\sum_{k=1}^{m} Y_k)$$

En pratique, on préfère utiliser le r-bis corrigé, qui exclue l'item en question du score global :

$$r_j^{\text{bis corrigé}} = corr(Y_j, S_{\text{corrigé}}) = corr(Y_j, \frac{1}{m-1}\sum_{k\neq j} Y_k)$$

Les r-bis corrigés sont calculés pour tous les items du construit unidimensionnel, puis les items dont le r-bis est inférieur à 0,15 sont retirés des données.

## 2.3 Détection des fonctionnements différentiels d'items

Chaque année, une partie des items de l'évaluation de l'année précédente est reprise dans la nouvelle évaluation. Ces items repris à l'identique, appelés items d'ancrage, vont permettre de détecter une évolution du niveau des élèves, en fonction des variations de leurs difficultés selon les années.

En effet, au fur à mesure des années, le niveau des élèves peut varier. Or, notre méthode d'estimation présentée ci-après suppose une distribution normale centrée et réduite des performances des élèves. Ainsi, par exemple, une augmentation du niveau des élèves ne transparaitra pas dans une hausse de la moyenne des $(\theta_i)_i$, fixée à 0 par l'algorithme, mais dans une baisse des difficultés des items $(b_j)_j$. En effet, si les élèves deviennent meilleurs, les questions leur sembleront plus faciles. Néanmoins, à partir des variations des difficultés des items sur plusieurs années, il est possible de déduire l'évolution du niveau des élèves. Dans notre modèle, on préfère considérer que la difficulté d'un item d'ancrage est constante au cours du temps, et les performances des élèves sont transformés en conséquence pour absorber l'effet d'une variation de niveau.

Toutefois, il est possible que la difficulté d'un item d'ancrage varie différemment par rapport aux autres questions, pour des raisons qui ne sont pas liées à l'évolution du niveau des élèves. Par exemple, si un item porte sur un sujet qui est retiré du programme scolaire [4],

4. Cet exemple permet d'illustrer simplement un fonctionnement différentiel d'item, toutefois dans le cas des évaluations nationales, il faut rappeler que tous les items font partie du programme.

sa difficulté risque d'augmenter sans correspondre à une baisse du niveau des élèves dans la matière. On ne peut pas considérer que la difficulté de cet item est constante au cours du temps, puisqu'il est devenu hors programme. Par conséquent, cet item ne peut plus servir de référence pour étudier l'évolution du niveau des élèves. On dit que l'item présente un fonctionnement différentiel.

Il s'agit alors de repérer dans le construit unidimensionnel les items d'ancrage dysfonctionnels pour les considérer comme des nouveaux items, avec une nouvelle difficulté à déterminer. Pour cela, deux critères sont choisis pour décrire un fonctionnement différentiel d'item (FDI) :

— la difficulté de l'item mis sur l'échelle de l'année dernière s'écarte de sa valeur de l'année précédente de plus de 0,5 ;

— ou la pente de l'item mis sur l'échelle de l'année dernière s'écarte de sa valeur de l'année précédente de plus de 0,25.

En pratique, les paramètres d'item sont estimés sur le construit unidimensionnel pour chaque année séparément, grâce à une méthode de maximisation de la vraisemblance marginale (Bock and Aitkin, 1981). Deux jeux de paramètres sont obtenus, un pour chaque année. Pour pouvoir comparer les deux jeux, il faut les placer sur la même échelle. Pour ce faire, la méthode dite d'*equating mean-mean* est utilisée (Loyd and Hoover, 1980). Il est possible alors de comparer les paramètres placés sur la même échelle afin de détecter les items d'ancrage en fonctionnement différentiel selon les critères énoncés plus haut. Dès que l'un des deux critères est vérifié, l'item ne peut plus faire partie de l'ancrage : il est considéré comme un nouvel item avec une nouvelle difficulté à déterminer.

# 3 Estimation des performances sur la discipline

Une fois les items sélectionnés par construit unidimensionnel, il s'agit d'estimer pour chaque élève un score qui décrive son niveau de compétence dans la discipline.

L'un des objectifs de ces dispositifs est d'inscrire ces évaluations dans une perspective de comparaison temporelle. Pour cela, les résultats de l'année en cours sont d'abord calculés sur l'échelle de l'année précédente à l'aide d'une méthode dite de calibration concurrente. Dans la partie suivante, notre méthode pour transposer ces résultats sur l'échelle de l'année de référence sera présentée.

L'estimation sur l'échelle de l'année précédente s'effectue en trois étapes :

1. Préparation des données jointes sur les deux années consécutives ;
2. Estimation des paramètres des items en calibration concurrente ;
3. Estimation des niveaux de compétence des élèves.

## 3.1 Préparation des données jointes

L'analyse est menée séparément sur le français et les mathématiques. Elle est réalisée sur l'ensemble des données de l'année en cours, auxquelles sont ajoutées celles de l'année précédente. Les données sont décomposées en deux groupes, le groupe de référence correspondant à l'évaluation de l'année précédente. Il y a une ligne par élève et par question, indiquant la réussite ou non de l'élève à l'item.

Tout d'abord, les réponses aux items qui ont un r-bis inférieur à $0,15$ sont retirées des données. Puis, les noms des items d'ancrage qui ont été détectés en fonctionnement

différentiel sont modifiées dans les données de l'année précédente, afin de les considérer comme des items différents entre les deux années. On dispose alors de deux grands jeux de données, un en français et un autre en mathématiques, qui regroupent les données de passation des deux années consécutives.

### 3.2 Estimation des paramètres des items

Il s'agit d'abord d'estimer les paramètres, difficulté et pente, des items sur les évaluations des deux années réunies. L'algorithme utilisé calcule les paramètres des items par maximisation de la vraisemblance marginale en calibration concurrente, en supposant que la distribution des scores des élèves du groupe de référence est centrée et réduite (Bock and Aitkin, 1981). Cette méthode permet de caler le niveau de l'année en cours sur celui de l'année précédente, puisqu'un seul couple de paramètre $(a, b)$ est estimé pour chacun des items d'ancrage. Dans la section suivante sera expliquée la conversion de ces paramètres sur l'échelle de référence.

### 3.3 Estimation du niveau de compétence des élèves

Une fois les paramètres des items déterminés, l'analyse se poursuit par l'estimation des performances des élèves, sur les deux années. Il s'agit de trouver les $(\theta_i)_i$ qui maximisent la vraisemblance des données observées, avec les $(a_j, b_j)_j$ fixés. Comme le nombre de paramètres à estimer, les $(\theta_i)_i$, augmente avec le nombre d'observations, la méthode du maximum de vraisemblance classique est biaisée. La méthode d'estimation alors choisie repose sur la vraisemblance pondérée proposée par Warm (1989).

## 4 Conversion des résultats sur l'échelle de référence

La méthode d'estimation a permis de calculer les $(a_j, b_j)_j$ et les $(\theta_i)_i$ qui maximisent la vraisemblance des données observées. Dans l'algorithme, la moyenne des $\theta$ de la cohorte de référence (ici les élèves de l'année dernière) est fixée à 0 et l'écart type à 1, et les $\theta$ de l'autre groupe (les élèves de l'année en cours) sont placés sur la même échelle.

L'algorithme présuppose une distribution centrée et réduite des performances pour des raisons liées à l'identification du modèle. En effet, la variable $\theta$ n'est définie qu'à une transformation linéaire près. Autrement dit, si on translate les compétences de la manière suivante :

$$\theta^* = A\theta + B, \text{ avec } A \text{ et } B \text{ deux constantes, } A \text{ non nul}$$

il existe des couples $(a_j^*, b_j^*)_j$ tels que les $(\theta_i^*)_i$ et les $(a_j^*, b_j^*)_j$ soient solution du même problème de maximisation. Ce point sera explicité à la fin de cette partie.

Ainsi, après avoir calculé les scores sur l'échelle de l'année précédente, l'objectif est de les placer sur l'échelle de référence, ce qui permettra de faire des comparaisons entre plusieurs années. L'échelle de référence a été définie au début des évaluations nationales, elle correspond à la distribution des scores en 2017 pour les évaluations de sixième et en 2019 pour celles de seconde. Pour transposer les nouveaux résultats sur cette échelle, une conversion est réalisée, appelée *equating*.

## 4.1 Calcul des coefficients de conversion

Il s'agit maintenant d'effectuer une transformation linéaire, du type $\theta \longrightarrow A\theta + B$, pour ramener les scores sur la bonne échelle. Pour déterminer les coefficients $A$ et $B$, il faudra comparer les scores de l'année dernière nouvellement calculés avec les mêmes scores obtenus l'année dernière et déjà placés sur l'échelle de référence.

Pour la suite, $(\mu^{t-1}, \sigma^{t-1})$ désignent la moyenne et l'écart type de la distribution des scores de l'année précédente qui viennent d'être calculés en calibration concurrente, et $(\mu_{ref}^{t-1}, \sigma_{ref}^{t-1})$ sont la moyenne et l'écart type des mêmes scores mais qui ont été obtenus l'année dernière et déjà placés sur l'échelle de référence.

D'après les méthodes d'estimation introduites précédemment, il est attendu que les scores de l'année passée nouvellement calculés suivent une distribution normale centrée et réduite, c'est-à-dire $(\mu^{t-1}, \sigma^{t-1}) = (0, 1)$. Or, il arrive que les moments soient légèrement différents, notamment avec un écart type plus élevé à cause des scores parfaits [5]. Par conséquent, les scores de l'année précédente vont être convertis sur l'échelle $(0, 1)$ grâce à la transformation suivante :

$$\theta \longrightarrow \frac{\theta - \mu^{t-1}}{\sigma^{t-1}}$$

Une fois centrés réduits, les scores de l'année précédente peuvent être linéairement transformés pour retrouver les résultats finaux de l'année dernière, qui avaient été placés sur l'échelle de référence. Pour cela, la transformation suivante est appliquée :

$$\theta \longrightarrow \sigma_{ref}^{t-1} \times \theta + \mu_{ref}^{t-1}$$

En regroupant les deux opérations successives, la transformation sur les scores s'écrit :

$$\theta \longrightarrow \sigma_{ref}^{t-1} \times \frac{\theta - \mu^{t-1}}{\sigma^{t-1}} + \mu_{ref}^{t-1}$$

Les coefficients d'equating s'écrivent alors :

$$\begin{cases} A = \dfrac{\sigma_{ref}^{t-1}}{\sigma^{t-1}} \\ B = \mu_{ref}^{t-1} - A\mu^{t-1} \end{cases}$$

Ces deux coefficients permettent de passer de l'échelle des scores de l'année dernière à celle des scores de l'année de référence.

Le coefficient $A$ doit être proche de 1 puisque les évaluations sont conçues pour mesurer la même compétence chaque année. Le coefficient $B$ représente l'évolution du niveau entre l'année dernière et l'année de référence.

## 4.2 Conversion des performances

Une fois ces deux coefficients obtenus, la distribution des performances des élèves peut être déplacée pour que les scores des élèves soient comparables avec ceux de l'année de référence, par la transformation : $\theta \longrightarrow A\theta + B$.

5. Les scores parfaits correspondent aux évaluations totalement réussies ou totalement ratées. Ces passations ne permettent pas de bien positionner les élèves sur l'échelle de compétence puisque leurs niveaux sont au-delà des limites prévues par l'évaluation.

Par ailleurs, en plus d'obtenir une estimation de la performance de chaque élève, il est prévu de qualifier leur niveau de maitrise dans la discipline. Pour cela, quatre niveaux de maitrise sont définis : maitrise insuffisante, maitrise fragile, maitrise satisfaisante et très bonne maitrise.

Chaque niveau est délimité par des seuils de compétence qui ont été établis par des experts grâce à la méthode des bookmarks (Cizek and Bunch, 2007). Grâce à ces seuils, chaque élève peut être placé dans un groupe de maitrise en fonction de sa performance à l'évaluation.

Par la suite, dans les publications, les résultats seront principalement présentés sous forme de taux de maitrise, i.e. du pourcentage d'élèves avec une maitrise satisfaisante ou très bonne de la discipline.

Ci-dessous, les seuils de maitrise utilisés pour les deux niveaux scolaires et les deux disciplines, français et mathématiques, sont détaillés.

| | Sixième | | Seconde | |
|---|---|---|---|---|
| | Français | Mathématiques | Français | Mathématiques |
| Maitrise insuffisante | $\theta < -1.8$ | $\theta < -1.8$ | $\theta < -1.8$ | $\theta < -1.8$ |
| Maitrise fragile | $-1.8 \leq \theta < -0.9$ | $-1.8 \leq \theta < -0.6$ | $-1.8 \leq \theta < -0.8$ | $-1.8 \leq \theta < -0.7$ |
| Maitrise satisfaisante | $-0.9 \leq \theta < 1.3$ | $-0.6 \leq \theta < 1.3$ | $-0.8 \leq \theta < 1.3$ | $-0.7 \leq \theta < 1.3$ |
| Très bonne maitrise | $1.3 \leq \theta$ | $1.3 \leq \theta$ | $1.3 \leq \theta$ | $1.3 \leq \theta$ |

Figure 2 – Seuils de maitrise

Enfin, pour des questions d'usage, les scores sont transformés pour que l'échelle de référence ne soit pas centrée et réduite, mais centrée en 250 et d'écart type 50. Cela correspond à la transformation suivante :

$$\theta \longrightarrow 50 \times \theta + 250$$

### 4.3 Conversion des paramètres d'item

L'analyse psychométrique se termine par la convertion des paramètres d'item sur l'échelle de référence. L'ojectif est d'appliquer sur les paramètres d'item $(a_j, b_j)_j$ une transformation équivalente à celle effectuée sur les performances des élèves, afin de s'assurer que les $(a_j, b_j)_j$ et les $(\theta_i)_i$ linéairement transformés sont solutions du même problème de maximisation de vraisemblance.

Pour placer les scores sur l'échelle de référence, la transformation linéaire $\theta^* = A\theta + B$ a été effectuée. Il s'agit de trouver quelle transformation appliquer aux paramètres d'item pour accompagner le changement d'échelle.

De manière explicite, il s'agit de trouver les $(a_j^*, b_j^*)_j$ tels que $(A\theta_i + B, a_j^*, b_j^*)_{ij}$ soient solution du même problème de maximisation, c'est-à-dire :

$$\mathbb{P}(Y_{ij} = 1 | A\theta_i + B, a_j^*, b_j^*) = \mathbb{P}(Y_{ij} = 1 | \theta_i, a_j, b_j) \text{ pour tout } \theta_i$$

avec $\mathbb{P}(Y_{ij} = 1 | \theta_i, a_j, b_j) = \frac{e^{a_j(\theta_i - b_j)}}{1+e^{a_j(\theta_i - b_j)}}$ la probabilité qu'un élève $i$ réussisse l'item $j$.

$$
\begin{aligned}
&\mathbb{P}(Y_{ij}=1|\theta_i^*,a_j^*,b_j^*)=\mathbb{P}(Y_{ij}=1|\theta_i,a_j,b_j) \text{ pour tout } \theta_i\\
&\iff a_j^*(\theta_i^*-b_j^*)=a_j(\theta_i-b_j) \text{ pour tout } \theta_i\\
&\iff a_j^*(A\theta_i+B-b_j^*)=a_j(\theta_i-b_j) \text{ pour tout } \theta_i\\
&\iff Aa_j^*(\theta_i-\frac{b_j^*-B}{A})=a_j(\theta_i-b_j) \text{ pour tout } \theta_i\\
&\iff Aa_j^*=a_j \text{ et } \frac{b_j^*-B}{A}=b_j\\
&\iff a_j^*=\frac{a_j}{A} \text{ et } b_j^*=Ab_j+B
\end{aligned}
$$

Ainsi, les paramètres transformés sur l'échelle de référence sont obtenus grâce aux opérations suivantes :

$$
\begin{cases}
a_j^*=\dfrac{a_j}{A}\\
b_j^*=Ab_j+B
\end{cases}
$$

# 5 Erreurs d'estimation sur les résultats agrégés

Les résultats de ces analyses sont publiés par la DEPP sous forme agrégée : on fournit à chaque établissement scolaire, chaque département et chaque académie, la moyenne des compétences de leurs élèves en français et l'équivalent en mathématiques.

Ces valeurs sont les moyennes d'estimations individuelles, qui ne sont pas des nombres exacts, mais des estimateurs sans biais avec des écarts types calculés lors du déroulement de l'algorithme. Il s'agit alors de fournir, avec les moyennes des compétences des élèves, les bornes des intervalles de confiance dans lesquels ces moyennes se situent avec une probabilité supérieure à 95 %.

Dans cette partie est présentée la méthode pour calculer les bornes des intervalles de confiance au seuil de 95 % pour les moyennes des scores des élèves.

## 5.1 Moyenne de variables aléatoires gaussiennes

Pour chaque élève $v$, on estime un score $\hat{\theta}_v$ dont la loi s'approche de la loi normale $\mathcal{N}(\theta_v,\sigma_v)$, avec $\theta_v$ la vraie compétence de l'élève et $\sigma_v$ un estimateur de l'erreur standard.

On considère maintenant un groupe de $n$ élèves. Pour fournir des résultats agrégés, on calcule la moyenne des estimateurs de score :

$$
\hat{\Theta}=\frac{1}{n}\sum_{v=1}^{n}\hat{\theta}_v
$$

$\hat{\Theta}$ est une variable aléatoire et on peut déterminer sa loi en calculant sa fonction caractéristique :

$$
\mathbb{E}(e^{it\hat{\Theta}})=\mathbb{E}(e^{it\frac{1}{n}\sum_{v=1}^{n}\hat{\theta}_v})=\mathbb{E}\left(\prod_{v=1}^{n}e^{i\frac{t}{n}\hat{\theta}_v}\right)
$$

Par indépendance des estimateurs entre élèves,

$$\mathbb{E}(e^{it\hat{\Theta}}) = \prod_{v=1}^{n} \mathbb{E}\left( e^{i\frac{t}{n}\hat{\theta}_v} \right)$$

Or, la fonction caractéristique d'une variable aléatoire $Z$ qui suit une loi normale $\mathcal{N}(\mu, \sigma)$ s'écrit $\mathbb{E}(e^{itZ}) = e^{it\mu - \frac{1}{2}t^2\sigma^2}$.

Par conséquent,

$$\mathbb{E}(e^{it\hat{\Theta}}) = \prod_{v=1}^{n} e^{i\frac{t}{n}\theta_v - \frac{1}{2}\left(\frac{t}{n}\right)^2 \sigma_v^2} = e^{it\frac{1}{n}\sum_{v=1}^{n}\theta_v - \frac{t^2}{2}\frac{1}{n^2}\sum_{v=1}^{n}\sigma_v^2}$$

On reconnait la fonction caractéristique d'une loi normale de moyenne $\mu = \frac{1}{n}\sum_{v=1}^{n}\theta_v$ et d'écart-type $\sigma = \sqrt{\frac{1}{n^2}\sum_{v=1}^{n}\sigma_v^2}$.

Donc, la moyenne des $n$ scores suit une loi normale d'écart type :

$$\sigma = \frac{1}{n}\sqrt{\sum_{v=1}^{n}\sigma_v^2}$$

## 5.2 Modélisation de l'intervalle de confiance

Comme $\hat{\Theta}$ suit une loi normale, on peut prendre comme intervalle de confiance avec un niveau de confiance au seuil de 95 % :

$$[\hat{\Theta} - 2\sigma, \hat{\Theta} + 2\sigma] \text{ avec } \sigma = \frac{1}{n}\sqrt{\sum_{v=1}^{n}\sigma_v^2}$$

L'objectif est de fournir un intervalle de confiance pour accompagner les résultats publiés. Il est possible de simplifier la formule de l'erreur d'estimation pour pouvoir la calculer rapidement à chaque échelle géographique.

Des recherches annexes ont montré que cette erreur dépend fortement du nombre d'élèves concernés, et peu des erreurs d'estimation individuelles (puisque les faibles disparités sur les erreurs d'estimation individuelles sont compensées lorsqu'on les somme). Par conséquent, pour modéliser l'erreur d'estimation, l'erreur sur chaque estimateur peut être approximée par sa valeur moyenne $\sigma_v \approx 14$. Ainsi, l'erreur d'estimation sur la moyenne de $n$ résultats peut se simplifier :

$$\sigma = \frac{1}{n}\sqrt{\sum_{v=1}^{n}\sigma_i^2} \approx \frac{\sqrt{n \times 14^2}}{n} = \frac{14}{\sqrt{n}}$$

Par conséquent, un intervalle de confiance autour du score agrégé $\hat{\Theta}$ d'un établissement avec $n$ élèves a pour bornes

$$\left[\hat{\Theta} - \frac{28}{\sqrt{n}}, \hat{\Theta} + \frac{28}{\sqrt{n}}\right]$$

Par exemple, pour un établissement avec 100 élèves scolarisés en classe de sixième, on peut estimer le niveau de compétence moyen en français avec une précision de $\pm 2,8$ points sur l'échelle des scores $(250, 50)$, avec une confiance au seuil de 95%.

# 6 Estimation des avantages des élèves par domaine

En plus de fournir un score qui quantifie les niveaux de compétence dans la discipline, l'analyse psychométrique peut permettre de repérer les points forts et les points faibles des élèves selon les domaines.

Les forces et faiblesses sont calculées pour chaque élève en comparant le nombre de points marqués dans le domaine et le nombre de points que l'élève aurait marqués si sa compétence dans le domaine était la même que sur l'ensemble de la discipline. Il s'agit de comparer le score brut obtenu dans le domaine avec le score attendu, en retranchant le résidu sur la discipline (défini par la suite).

Dans cette partie, un exemple sera proposé pour commencer à se familiariser avec les concepts de cette méthode. Puis, la méthodologie sera explicitée dans le cas général.

## 6.1 Exemple de calcul de résidus et mesure des avantages par domaine

On considère un élève $i$ qui passe une évaluation avec 10 questions. L'évaluation est composée de 2 domaines, $d_1$ et $d_2$, de 6 et 4 items respectivement.


Le score obtenu par l'élève (i.e. le score brut) est défini par :

$$\sum_{j=1}^{10} Y_{ij} \text{ avec } Y_{ij} = 1 \text{ si l'élève répond juste et 0 sinon.}$$

On peut évaluer le score attendu (i.e. le score que l'élève obtient en espérance selon sa compétence estimée $\hat{\theta}_i$) en calculant :

$$\sum_{j=1}^{10} \mathbb{P}(Y_{ij} = 1 | \hat{\theta}_i, a_j, b_j)$$

L'écriture est simplifiée en posant $P_j(\hat{\theta}_i) = \mathbb{P}(Y_{ij} = 1 | \hat{\theta}_i, a_j, b_j)$

On suppose que l'élève a obtenu un score brut de 6/10 qui se décompose selon les domaines de la manière suivante :

On suppose de plus que la compétence estimée $\hat{\theta}_i$ de l'élève laissait présager un score attendu légèrement inférieur, 5/10, et décomposé selon les domaines de la sorte :

La compétence $\hat{\theta}_i$ décrit le niveau de l'élève pour la discipline. Pour savoir si un domaine de compétence est un point fort de l'élève, il est possible de regarder si, dans ce domaine, le score obtenu par l'élève dépasse le score attendu, calculé en fonction de sa compétence sur la discipline.

Par exemple, pour le domaine $d_1$, la différence entre le score obtenu et le score attendu s'écrit :

$$\sum_{j=1}^{6}(Y_{ij} - P_j(\hat{\theta}_i)) = 1$$

Dans cette formule, on somme les $(Y_{ij} - P_j(\hat{\theta}_i))_j$ qui correspondent aux résidus sur les items, i.e. la différence entre l'observé $Y_{ij}$ et l'attendu $P_j(\hat{\theta}_i)$.

Par la suite, le résidu sur l'item $j$ pour l'élève $i$ est noté $r_{ij}$, et la somme des résidus sur le domaine $d_1$ $R_{i1}$.

De même, on calcule la somme des résidus sur le domaine $d_2$ :

$$R_{i2} = \sum_{j=7}^{10}(Y_{ij} - P_j(\hat{\theta}_i)) = 0$$

En examinant les résidus sur les domaines, il peut être déduit que $d_1$ est un point fort de l'élève et que $d_2$ n'est ni point fort ni point faible. Or, compte-tenu de la partition de l'évaluation en deux domaines, ce n'est pas logique : les points forts et points faibles devraient se contrebalancer parfaitement afin que le niveau moyen corresponde à la compétence de l'élève sur la discipline.

C'est pour cela qu'il faut retrancher à ces valeurs la somme des résidus sur la discipline. En effet, bien que l'algorithme tente de rapprocher au maximum les scores obtenus et les scores attendus sur l'ensemble des items, il est possible qu'il reste quand même un faible résidu sur la discipline.

Dans l'exemple, la somme des résidus sur la discipline vaut $R_i = \sum_{j=1}^{10}(Y_{ij} - P_j(\hat{\theta}_i)) = 1$

Ainsi, on définit $c_{ik}$ l'avantage de l'élève $i$ dans le domaine $d_k$ par rapport à son niveau sur la discipline en retranchant une partie du résidu sur la discipline, de la manière suivante :

$$c_{ik} = R_{ik} - \frac{m_k}{m}R_i$$

avec $m_k$ et $m$ le nombre d'items respectivement dans le domaine $d_k$ et sur l'ensemble de la discipline. Le facteur $\frac{m_k}{m}$ permet de retrancher pour chaque domaine une part du résidu global proportionnelle à sa taille.

Pour déterminer les avantages de l'élève $i$ de l'exemple, on calcule :

$$c_{i1} = 1 - \frac{6}{10} \times 1 = \frac{4}{10} \text{ et } c_{i2} = 0 - \frac{4}{10} \times 1 = -\frac{4}{10}$$

L'avantage s'exprime en terme de nombre d'items réussis en plus ou en moins dans le domaine, selon la compétence de l'élève moyenne sur la discipline. Par exemple ici, l'élève $i$ a marqué 0,4 points de plus dans le domaine $d_1$, et 0,4 points de moins sur $d_2$ par rapport à son niveau moyen sur la matière.

## 6.2 Méthode de calcul des avantages par domaine à différentes échelles

L'exemple plus haut a illustré l'intuition théorique qui justifie le recours aux résidus pour évaluer les points forts et faibles des élèves par domaine. Dans cette partie, sera explicitée la méthode qui permet, dans un premier temps, d'évaluer les avantages des élèves par domaine, et dans un second temps, d'élargir le champ afin de déterminer les points forts et faibles des établissements selon les domaines de compétence.

En suivant les notations précédentes, l'avantage d'un élève $i$ dans le domaine $d_k$ s'écrit :

$$c_{ik} = R_{ik} - \frac{m_k}{m} R_i$$

avec $R_{ik}$ et $R_i$ les sommes des résidus respectivement sur le domaine $d_k$ et sur la discipline, et $m_k$ et $m$ les nombres d'items respectivement dans le domaine $d_k$ et dans la discipline.

Pour rappel, le résidu $r_{ij}$ sur un item $j$ pour un élève dont la compétence a été estimée $\hat{\theta}_i$ est défini de la manière suivante :

$$r_{ij} = Y_{ij} - P_j(\hat{\theta}_i)$$

Il s'agit maintenant de déterminer les points forts et points faibles à des échelles plus grandes : établissement, département, académie et au niveau national.

Pour un groupe de $n$ élèves, la variable d'intérêt est la moyenne des scores des élèves. Le score obtenu par le groupe correspond à la moyenne des scores obtenus par les élèves :

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} Y_{ij}$$

Le score attendu pour le groupe correspond à la moyenne des scores attendus pour les élèves :

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} P_j(\hat{\theta}_i)$$

Par conséquent, on peut définir, de la même manière que précédemment, $c_k$ l'avantage d'un groupe dans le domaine $d_k$ comprenant $m_k$ items par :

$$\begin{aligned} c_k = R_k - \frac{m_k}{m} R &= \frac{1}{n} \sum_{i=1}^{n} \left( \sum_{j \in d_k} Y_{ij} - P_j(\hat{\theta}_i) \right) - \frac{m_k}{m} \frac{1}{n} \sum_{i=1}^{n} \left( \sum_{j=1}^{m} Y_{ij} - P_j(\hat{\theta}_i) \right) \\ &= \sum_{j \in d_k} \overline{Y_{.j} - P_j(\hat{\theta}_.)}^{n} - \frac{m_k}{m} \sum_{j=1}^{m} \overline{Y_{.j} - P_j(\hat{\theta}_.)}^{n} \end{aligned}$$

La formule est identique à celle appliquée à un seul élève, sauf qu'au lieu de prendre la somme des résidus sur tous les items pour un élève, on prend la somme de la moyenne des résidus sur tous les élèves pour chaque item.

Une autre manière d'interpréter cette formule, qui permettra de compiler plus facilement les avantages d'un groupe, est de faire ressortir les avantages des élèves :

$$c_k = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j\in d_k}(Y_{ij} - P_j(\hat{\theta}_i)) - \frac{m_k}{m}\sum_{j=1}^{m}(Y_{ij} - P_j(\hat{\theta}_i))\right) = \frac{1}{n}\sum_{i=1}^{n} c_{ik}$$

Pour des questions d'usage, ces valeurs, qui correspondent initialement à un nombre de points, sont mises sur l'échelle des scores de la discipline concernée. Ainsi, l'écart type des avantages par domaine sur l'ensemble des domaines d'une discipline est le même que celui des scores sur la discipline.

## Conclusion

Ce document présente les différentes méthodes qui permettent d'analyser les résultats des évaluations nationales de sixième et de seconde, en français et en mathématiques. Des explications théoriques plus poussées peuvent être trouvées dans les ouvrages cités dans la bibliographie.

## Bibliographie


R. D. Bock and M. Aitkin. Marginal maximum likelihood estimation of item parameters : application of an em algorithm. *Psychometrika*, 46(4) :443–459, 1981.

G. J. Cizek and M. B. Bunch. *Standard setting : A guide to establishing and evaluating performance standards on tests.* Sage Publications Ltd, 2007.

B. H. Loyd and H. D. Hoover. Vertical equating using the rasch model. *Journal of Educational Measurement*, 17(3) :179–193, 1980.

T. Rocher. *Mesure des compétences : les méthodes se valent-elles ? Questions de psychométrie dans le cadre de l'évaluation de la compréhension de l'écrit.* PhD thesis, Université Paris Ouest Nanterre La Défense, 2013.

T. A. Warm. Weighted likelihood estimation of ability in the item response theory. *Psychometrika*, 54(3) :427–450, 1989.