Comment modéliser la réussite scolaire en tenant compte de plusieurs niveaux d'analyse ?

Julien DANHIER, Céline TENEY
Chercheurs[1]

L'analyse des résultats aux tests « PISA » a abouti aux conclusions que non seulement les élèves issus de milieux défavorisés réussissaient moins bien que les autres, mais qu'ils réussissaient d'autant moins bien qu'ils étaient scolarisés dans des établissements où l'origine socio-économique moyenne des élèves est moins favorisée. Pour établir ce genre de résultat, les chercheurs ont recours à des modèles statistiques dits « multiniveaux ».

Nous nous plaçons dans le domaine de l'éducation pour présenter un exemple simple d'analyse multiniveaux afin de mettre en lumière l´intérêt de cette méthode (voir Danhier et Martin 2014 pour une application plus développée). À cet effet, nous allons analyser les données issues de l'enquête PISA (« Program for International Student Assessment »). Celle-ci est un projet de recherche mené par l'OCDE qui vise à évaluer "dans quelle mesure les élèves qui approchent du terme de leur scolarité obligatoire possèdent certaines des connaissances et compétences essentielles pour participer pleinement à la vie de nos sociétés modernes" (OCDE 2014: 23). Plus précisément, nous nous limitons aux 1963 élèves de l'enseignement secondaire ordinaire général (excluant ainsi ceux de l'enseignement de qualification et de l'enseignement spécialisé), scolarisés dans 79 écoles de la Fédération Wallonie-Bruxelles. Ces données sont hiérarchiques puisqu'elles reflètent une réalité structurée où les élèves sont regroupés dans des écoles. Afin de tenir compte de cette hiérarchie, les élèves et les écoles ont été utilisés comme unités aux premier et second niveaux.

Pour illustrer ce type d'analyse (voir figure 1), nous présentons, successivement, des modèles multi-niveaux afin d'expliquer la dispersion des résultats en mathématiques aux épreuves PISA 2012 (dont l'échelle s'étend d'environ 235 à 810 points). Le modèle 1 nous permet d'observer l'effet de variables socio-démographiques sur la réussite scolaire. Parmi les variables disponibles, nous nous sommes limités au genre (représenté par une variable dichotomique) et à l'origine socio-économique. Cette dernière est un indice composite couvrant les niveaux d'éducation des parents, leurs situations professionnelles et diverses possessions du ménage. Il se mesure sur une échelle continue allant de -2,7 à 2,5. Dans le modèle 2, l'effet du retard scolaire accumulé par l'étudiant est ajouté. Finalement, il est possible de mesurer l'effet propre de la composition socio-économique de l'école (modèle 3). Celle-ci est l'effet spécifique du regroupement des élèves, qu'il soit dû à des interactions directes entre pairs (discussions, motivations, disputes ou tensions entre différents groupes), à des pratiques du corps professoral (ajustement du style pédagogique ou attentes différentes relatives au groupe d'élèves) et à la qualité de l'école (problèmes de management des ressources humaines ou financement) (van Ewijk & Sleegers

1. Julien Danhier : Groupe de recherche sur les Relations Ethniques, les Migrations et l'Égalité (GERME) Université libre de Bruxelles (ULB) jdanhier@ulb.ac.be
Céline Teney : Centre for Social Policy Research University of Bremen celine.teney@uni-bremen.de

2010). Le tableau 1 reprend les modèles qui seront commentés successivement ci-dessous. Notons toutefois que si ce mode de progression est conseillé par Hox (2010), il n'est pas le seul possible et d'autres progressions peuvent ouvrir à d'autres interprétations.

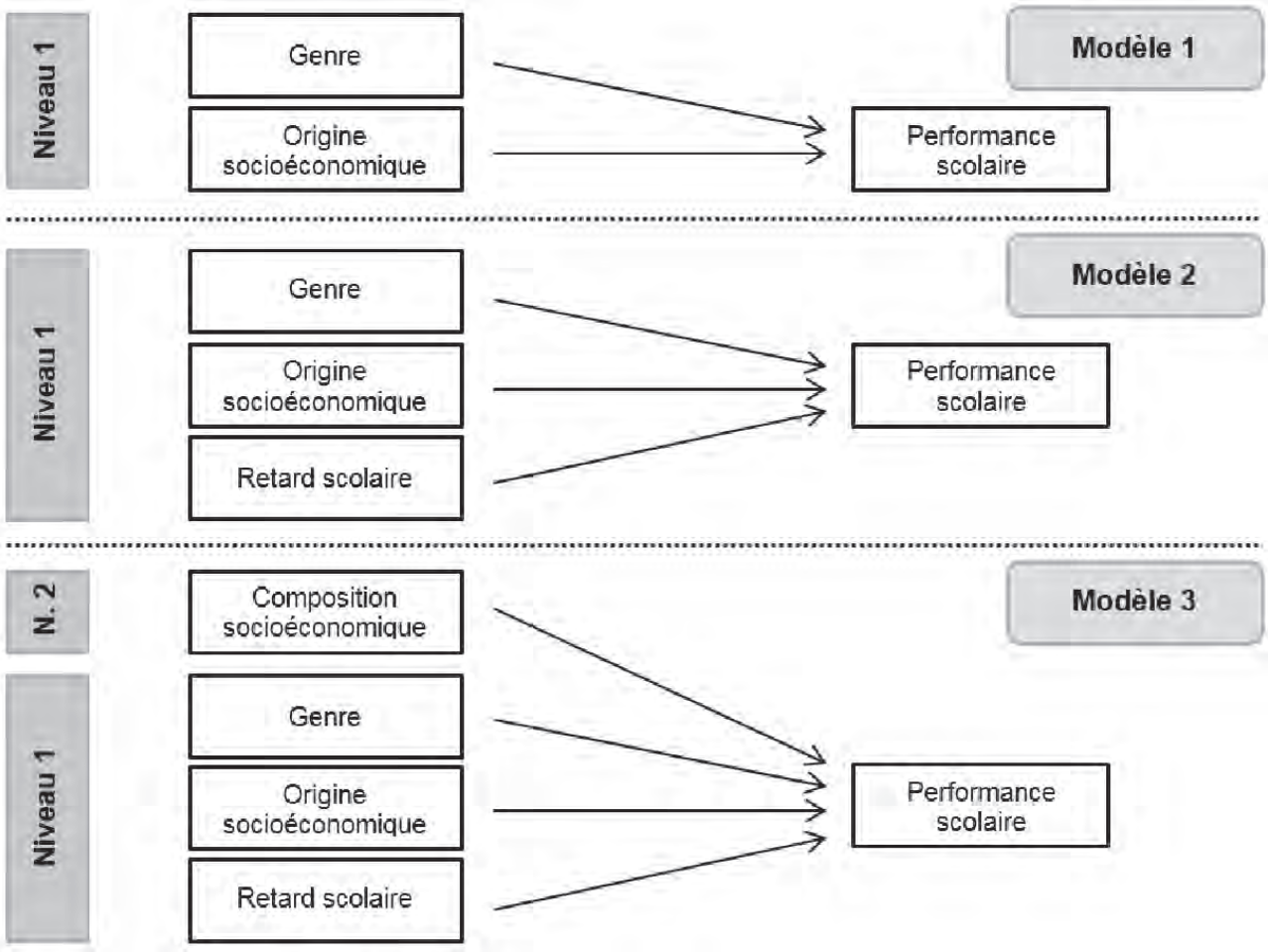

Figure 1 : Représentation des modèles successifs

Il est d'usage de commencer par produire un modèle dit « vide » dans lequel aucune variable explicative n'est spécifiée. La part dite « aléatoire » désigne les variances des « erreurs » à chaque niveau, à savoir la variance des écarts entre les résultats individuels et leurs moyennes dans chaque école (variance « élèves »: 5048) et celle des écarts entre ces dernières et la moyenne générale (variance « écoles »: 2597). Un tel modèle est utile pour deux raisons. Premièrement, il peut servir de base de comparaison pour apprécier la variance expliquée par les modèles suivants. Deuxièmement, il permet de voir comment la variance des scores obtenus par les élèves se répartit entre les niveaux. Ici, une variance « écoles » représente 34 % de la variance totale (qui correspond à la somme des variances « élèves » et « écoles »), ce qui confirme qu'il y a des agglomérats dans nos données. En d'autres termes, 34 % de la dispersion des résultats en mathématiques aux tests PISA est imputable à des différences entre écoles.

Une fois le modèle vide analysé, nous pouvons ajouter des variables explicatives, en commençant par les variables élèves, puis en ajoutant les variables écoles. Le tableau 1 reprend les valeurs prises par l'ordonnée à l'origine et les coefficients de régression (leur erreur standard apparaissant entre parenthèses) sous l'intitulé « part fixe ». Dans le modèle 1, l'ordonnée à l'origine correspond au score pour un individu présentant une valeur de 0 sur les échelles des variables explicatives. Tout comme dans une régression linéaire simple, les coefficients

associés aux variables explicatives représentent l'augmentation des scores d'un élève associée à une augmentation d'un point dans l'échelle de la variable considérée, toute chose étant égale par ailleurs. Un test de Wald permet de vérifier qu'un coefficient est significativement différent de 0. Ici donc, un garçon d'origine socio-économique moyenne obtiendra 525 points en mathématiques. S'il s'agissait d'une fille, elle aurait 19 points de moins. Enfin, si elle était d'origine plus défavorisée (d'un point sur cette échelle qui s'étend d'environ -2,7 à 2,5), elle aurait 21 points de moins. À titre de comparaison, l'OCDE a calculé qu'un écart de 41 points était, en moyenne, équivalent à une année de scolarisation.

Comme dans le cas de la régression linéaire simple, les variables peuvent être ajoutées par blocs successifs afin d'observer d'éventuels effets de médiation. Nous ajoutons donc le retard scolaire dans le modèle 2[2]. L'ordonnée à l'origine représente maintenant, le score d'un garçon d'origine socio-économique moyenne et à l'heure dans son parcours scolaire. L'effet propre du retard scolaire est énorme puisqu'une année de retard est associée à une baisse de 61 points aux tests PISA. Il est intéressant de noter que l'effet de l'origine socio-économique a fortement baissé. Ceci traduit un retard scolaire plus important chez les élèves d'origine défavorisée. On pourra ainsi dire que les scores moindres des élèves d'origine défavorisée traduisent d'une part leur origine, mais également leur présence plus importante parmi les élèves en retard.

L'ajout de variables au modèle réduit la variance des erreurs au niveau des élèves et des écoles, ce qui nous permet de calculer une variance expliquée à chaque niveau, à la manière du R2 de la régression linéaire (une mesure de l'adéquation entre le modèle et les données observées). L'introduction des variables socio-démographiques dans le modèle 1 s'associe à une réduction de la variance résiduelle attribuable aux "élèves" de l'ordre de 5,9% ((5048-4750)/5048). Lorsque l'effet du retard scolaire est pris en compte dans le modèle 2, cette réduction est de l'ordre de 39,0%. La diminution de la variance résiduelle attribuable aux "écoles" est de 25,8% (pour le modèle 1) et 76,6% (pour le modèle 2). Cela signifie que les variables mesurant des caractéristiques individuelles des élèves jouent à la fois au niveau des élèves, mais également au niveau des écoles. On pourra dès lors parler de l'effet du recrutement différentiel des écoles sur la dispersion des résultats des écoles. Il est possible de comparer plus finement les modèles entre eux et d'observer que 50 % de la variance entre écoles est uniquement expliquée par les caractéristiques scolaires des élèves tandis que 20 % de cette variance l'est par l'effet joint des caractéristiques scolaires et non scolaires.

Tableau 1 : Analyse multiniveaux

Paramètres	Modèle 0	Modèle 1	Modèle 2	Modèle 3
Part fixe				
Ordonnée à l'origine	512 (6,96) ***	525 (6,97) ***	569 (4,53) ***	575 (3,92) ***
Variables au niveau des élèves				
Genre (référence : homme)		-19,4 (4,19) ***	-26,1 (3,55) ***	-26,5 (3,54) ***
Origine socio-économique (-)		-20,5 (2,77) ***	-8,4 (2,23) ***	-6,19 (2,32) **
Retard scolaire			-61,4 (2,55) ***	-59,6 (2,47) ***
Variable au niveau des écoles				
Composition socio-économique			-49,4 (6,18) ***	
Part aléatoire				
Variance « élèves »	5048 (387)	4750 (367)	3078 (238)	3070 (237)
Variance « écoles »	2597 (549)	1928 (393)	624 (131)	249 (73)
Ajustement du modèle				
R2 « élèves »	0,0	5,9	39,0	39,2
R2 « écoles »	0,0	25,8	76,6	90,4

Niveaux de significativité : non significatif (n.s.), 0,05=*, 0,01=**, 0,001=***

2. Afin de faciliter l'interprétation pour les non-statisticiens, seules les variables dont le 0 n'a pas de sens ont été centrées autour de la moyenne générale, à savoir, l'origine et la composition socio-économique.

Dans le modèle 3, nous ajoutons la composition socio-économique (parfois appelée tonalité) qui est une variable au niveau des écoles mesurant l´origine socio-économique moyenne des élèves fréquentant la même école. Ceci permet de tester si le regroupement d'élèves dans les écoles selon leur origine a un effet sur les résultats scolaires. Comme l'origine individuelle a été précédemment modélisée, il s'agit d'un effet supplémentaire non réductible aux origines socio-économiques individuelles. Nous observons qu'avec un coefficient de -49, la composition socio-économique a un effet significatif sur les résultats scolaires. En d'autres termes, un garçon d'origine moyenne, à l'heure dans son cursus et dans une école dont la composition socio-économique est dans la moyenne aura 575 points en mathématiques. Dans une école parmi les plus défavorisées (1 point sur l'échelle de la composition qui s'étend de -1,2 à 1) ce même garçon, avec la même origine sociale, aura 49 points de moins.
L'étape suivante dans la modélisation consiste à vérifier si l'effet des variables individuelles est identique dans toutes les écoles et à complexifier ainsi la part aléatoire du modèle. Il est possible, par exemple, que l'effet de l'origine socio-économique diffère d'une école à l'autre. Nous avons testé cette spécification dans un modèle non reporté ici et ce modèle apparaît comme moins pertinent pour nos données. Si l'effet de l'origine socio-économique avait été significativement différent d'une école à l'autre, nous aurions également pu tester si l'importance de cet effet dépendait de la composition de l'école (effet d'interaction entre deux niveaux) et s'il était par exemple plus important dans les écoles plus favorisées. Une telle hypothèse n'est pas confirmée ici.

L'analyse multiniveaux

L'analyse de régression multi-niveaux (ou hiérarchique) est une famille d'analyses développées pour tenir compte des agglomérats présents dans les données, soit parce qu'ils existent dans la réalité, soit parce qu'ils sont créés par le chercheur lors de sa collecte. Un exemple typique de données dites « hiérarchiques » est issu du monde de l'éducation où les élèves sont regroupés dans des classes, elles-mêmes regroupées dans des écoles. Sans être exhaustives, deux qualités de ce type d'analyse méritent d'être soulignées et en justifient l'usage.
Premièrement, lorsque ces agglomérats sont présents dans les données, les observations ne peuvent pas être considérées comme indépendantes. Pour reprendre l'exemple de l'éducation, des élèves fréquentant une même classe ou une même école évoluent dans un même contexte scolaire, parfois, avec les mêmes enseignants et ont donc tendance à avoir un profil scolaire et socio-démographique plus similaire que des élèves provenant d´écoles ou de classes différentes. Cette relation doit être statistiquement prise en compte dans l'analyse sous peine d'obtenir des résultats faussement significatifs et l'analyse multi-niveaux est une des méthodes qui permettent de le faire.

Deuxièmement, les caractéristiques, non seulement des individus, mais aussi des agglomérats peuvent avoir une influence. Dans le cas de l'éducation, les caractéristiques des élèves et des écoles peuvent jouer différemment et intervenir à différents niveaux. Ainsi, l´origine sociale d´un élève peut exercer une influence plus ou moins importante sur sa réussite scolaire selon que l´élève fréquente l'une ou l'autre école. De plus, si les caractéristiques individuelles peuvent influer sur la réussite, leur agrégation au niveau des écoles peut également exercer une influence significative. Dit autrement, l´origine sociale moyenne d´une école peut exercer un effet supplémentaire à l´origine sociale individuelle sur la réussite scolaire d´un élève. L'analyse multi-niveaux permet de modéliser des variables à chaque niveau et d'observer leur influence selon le niveau considéré.

Pour aller plus loin

L'analyse multi-niveaux permet des modélisations complexes et son usage ne se limite pas à l'exemple simple que nous venons de présenter. Nous invitons le lecteur à consulter les ouvrages de référence pour appréhender toute l'étendue des analyses possibles (voir notamment Hox, 2010 ; Snijders & Bosker, 2012).

L'analyse multiniveaux n'est toutefois pas la panacée. Elle n'est pas adaptée à toutes les analyses. Il s'agit d'une méthode complexe ayant certains prérequis. Son usage doit ainsi être raisonné et justifié. Nous relevons rapidement certains problèmes auxquels tout utilisateur sera confronté.
L'analyse exige, tout d'abord, un type particulier de données. Premièrement, l'échantillon doit être de taille suffisante, non seulement au niveau des élèves, mais encore au niveau des écoles. En dessous d'au moins 100 écoles, la prudence sera requise, car des simulations ont relevé des biais importants dans le cas d'échantillons restreints (Maas & Hox, 2005). Deuxièmement, il faut être attentif à la définition des niveaux et à l'absence éventuelle de certains. Ainsi, dans nos données, un niveau intermédiaire aurait dû être utilisé : celui des classes. Son absence a pour conséquence une redistribution de la variance entre les deux autres niveaux. Dans le cas d'un niveau supérieur manquant, toute la variance aurait été imputée au niveau le plus haut, à savoir celui de l'école (Opdenakker & Van Damme, 2000).

Au-delà des données, certains choix méthodologiques spécifiques doivent être faits et auront des conséquences sur les résultats obtenus. Le choix de la méthode de centrage est un de ceux-ci. Comme dans le cas de la régression linéaire simple, il est possible de centrer les valeurs d'une variable autour de leur moyenne afin d'en faciliter l'interprétation. Dans l'analyse multiniveaux il est également possible de centrer ces valeurs autour de la moyenne de l'école. Ce choix guidé par la question de recherche n'est pas marginal puisqu'il produira des résultats non équivalents (Enders & Tofighi, 2007).

Malgré sa complexité, la méthode s'est répandue et de nombreux programmes permettent aujourd'hui de l'utiliser de manière relativement intuitive. Parmi d'autres, nous pouvons mentionner des programmes spécialisés, comme MlwiN et HLM (dont les manuels sont aisément accessibles pour les débutants) ou des logiciels plus généraux comme SPSS, SAS, Stata, Mplus ou R.

Références

[1] Danhier, J., & Martin, É. (2014). Comparing Compositional Effects in Two Education Systems: The Case of the Belgian Communities. British Journal of Educational Studies, 62(2), 171 189.
[2] Enders, C. K., &Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. Psychological methods, 12(2), 121 138.
[3] Hox, J. (2010). Multilevel analysis. Techniques and applications (2e éd.). New York : Routledge.
[4] Maas, C. J. M., & Hox, J. (2005). Sufficient Sample Sizes for Multilevel Modeling. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1(3), 86 92.
[5] OCDE, 2014. Résultats du PISA 2012: Savoirs et savoir-faire des élèves. Paris, OECD Publishing.
[6] Opdenakker, M.-C., & Van Damme, J. (2000). The importance of identifying levels in multilevel analysis: an illustration of the effects of ignoring the top or intermediate levels in school effectiveness research. School Effectiveness and School Improvement, 11, 103–130.
[7] Snijders, T. A. B., & Bosker, R. J. (2012). Multilevel analysis. An introduction to basic and advanced multilevel modeling (2nd éd.). London : Sage.
[8] Van Ewijk, r. & Sleegers, P. (2010). Peer ethnicity and achievement: a meta-analysis into the compositional effect. School Effectiveness and School Improvement, 21(3), 237-265.

Mini-débat : jusqu'où va le libre choix des auteurs dans la présentation d'un graphique ?

Avec la participation d'Yves GUIARD

Chercheur LTCI - CNRS et Telecom-ParisTech

et de Thomas PIKETTY

Professeur à l'Ecole d'économie de Paris

Les graphiques statistiques doivent-ils obéir à des règles strictes, ou bien au contraire les auteurs d'une étude disposent-ils d'une certaine liberté pour présenter graphiquement leurs résultats ? A cette question trop générale, il ne peut pas y avoir de réponse unique : mais il est intéressant d'examiner des cas particuliers.

Yves Guiard, chercheur émérite au laboratoire « Traitement et Communication de l'Information » de Télécom-ParisTech, a proposé à la rédaction de Statistique et Société un article où il critique un graphique paru dans le livre « Pour une révolution fiscale » de Camille Landais, Thomas Piketty et Emmanuel Saez. Pour Yves Guiard, un graphique qui représente une variable en fonction de quantiles de la distribution d'une autre variable doit utiliser en abscisses des classes d'égale importance, faute de quoi le graphique risque de donner une idée fausse du phénomène étudié. L'article d'Yves Guiard est publié dans les pages qui suivent.

Nous avons communiqué cet article aux auteurs du livre en question. Thomas Piketty a répondu en leur nom par une réaction qui est reproduite ci-dessous page 79. Il y affirme la liberté des auteurs, dans ce cas comme dans d'autres, à choisir la représentation qui leur semble la plus opportune.

Le lecteur jugera !