Regrouper les élèves par niveau de compétence

Une synthèse de la littérature

Nathalie Roques

Décembre 2024

Les termes anglais employés dans cette synthèse sont écrits en *italique* quand ce ne sont pas des titres ni des noms propres.

Introduction 5
Partie 1. La question et la méthode 7
1. Les modes de regroupements 7
2. Les effets mesurés 8
3. Les types d'études 9
4. La méthode 11
Partie 2. La recherche internationale 13
1. Les études primaires 13
2. Les méta-analyses 19
3. Conclure sur la recherche internationale 25
Partie 3. Les études anglaises 27
1. Le contexte anglais 27
2. La recherche sur les groupes de niveau 28
2.1. Comment sont regroupés les élèves ? 29
2.2. Groupes de niveau et résultats académiques 34
2.3. Groupes de niveau et résultats non cognitifs 36
2.4. Groupes de niveau et pratiques dans les établissements 39
2.5. Conclure sur la recherche anglaise 42
Discussion et conclusion 45
1. Les théories et les pratiques 45
2. Les résultats 48
3. Conclusion 50

Introduction

Le 5 décembre 2023, Gabriel Attal, alors ministre de l'Éducation nationale en France, annonçait que les élèves de 6ème et de 5ème seront répartis dans des « groupes de niveaux » (sous-entendu de capacité ou de compétences) pour les enseignements du français et des mathématiques dès septembre 2024, et qu'il en sera de même en 4ème et 3ème à partir de septembre 2025. Les autres disciplines ne sont pas concernées. Le 15 mars 2024, un arrêté est publié qui reprend ces annonces en remplaçant la notion de groupes de niveaux par celle de « groupes de besoins », et en conférant aux chefs d'établissements une grande marge de manœuvre. Les élèves en difficulté restent précisément ciblés (l'effectif du groupe dans lequel ils se retrouveraient pourrait être limité à 15)[1]. Cette mesure est un élément phare de la réforme Choc des savoirs qui a comme objectif d'élever le niveau des élèves français et qui s'appuie sur les conclusions de la mission Exigence des savoirs[2] mise en place deux mois plus tôt. Le dossier de presse[3] qui accompagne l'annonce ministérielle dit s'appuyer sur des études empiriques et cite deux documents pour justifier la mise en place de ces nouveaux groupes de niveau. Ce rapide survol de la recherche ne suffit pas à répondre aux questions que peuvent se poser les acteurs de la communauté éducative sur l'efficacité de cette mesure et cette synthèse a comme objectif de combler cette lacune.

Le cadre posé par la réforme rapidement décrit en introduction invite bien évidemment à rassembler les résultats de recherches rigoureuses qui répondraient précisément à notre question. Il s'agirait alors d'analyser les études qui ont évalué les effets à plus ou moins long terme d'une transformation de la politique éducative d'un État imposant à tous ses établissements secondaires inférieurs (le niveau collège en France) de passer de regroupements des élèves par classe d'âge de niveau hétérogène dans toutes les disciplines, à des regroupements des élèves par classe d'âge de niveau homogène en mathématiques et en littéracie. Mais la réforme française engagée est inédite dans le monde, et de telles études n'existent pas. Notre objectif sera donc plus modestement d'analyser les résultats de la recherche internationale sur la question des regroupements par niveau de compétences. Pour alléger l'écriture de cette synthèse, le terme « niveau » employé seul et au singulier devra être compris comme un « niveau de compétences »[4]. Dans une première partie, la question de recherche est précisée et le cadre comme les limites de cette synthèse sont posés. Les résultats de la recherche internationale (principalement américaine) sont présentés dans une seconde partie, la troisième partie étant consacrée aux études anglaises récentes qui méritent d'être bien comprises. La discussion en dernière partie intègre des éléments théoriques et pratiques mais aussi politiques afin de mettre en lumière les conclusions de ces études la plupart du temps quantitatives et d'orienter les praticiens et les décideurs vers des solutions adaptées aux objectifs qu'ils se sont fixés.

[1] NOR : MENE2407076N _ Note de service du 15-3-2024
https://www.education.gouv.fr/bo/2024/Special2/MENE2407076N

[2] Cette mission n'a publié aucun texte. Une audition de ses responsables de la mission par la commission des affaires culturelles et de l'éducation le 20 décembre 2023 peut être consultée sur internet https://www.assemblee-nationale.fr/dyn/16/organes/commissions-permanentes/affaires-culturelles/actualites/mission-exigence-des-savoirs-audition

[3] https://www.education.gouv.fr/choc-des-savoirs-une-mobilisation-generale-pour-elever-le-niveau-de-notre-ecole-380226

[4] Le terme « besoin » n'est pas utilisé par les publications analysées ici, et a été volontairement écarté.

Partie 1. La question et la méthode

Formulée de la façon la plus concise possible, notre question de recherche consiste à évaluer et à comparer les effets que peuvent avoir différents modes de regroupements des élèves, ces regroupements étant basés pour certains sur une évaluation de leur niveau dans une ou plusieurs disciplines. Les effets concernent en premier lieu le niveau scolaire des élèves, mais également d'autres caractéristiques non cognitives qui influent sur leur parcours. Ces élèves sont scolarisés dans les établissements publics du secondaire inférieur. Les types de regroupements possibles et les effets que l'on cherche à évaluer sont définis dans un premier temps. Puis les principales caractéristiques des designs des études dont nous analyseront les conclusions ici sont présentés. Cette première partie se termine par la description de la méthode suivie pour élaborer cette synthèse.

1. Les modes de regroupements

On va commencer par quatre remarques triviales :

1. Regrouper des élèves dans une classe (c'est-à-dire dans une salle avec au moins un enseignant) est un principe fondamental sur lequel repose tous les systèmes scolaires du monde. Seuls peuvent échapper à cette description les cours particuliers ou en très petit groupes (les cours de soutien par exemple).
2. Regrouper les élèves par classe d'âge[1] est une forme extrêmement fréquente de regroupement. Il s'agit en fait d'un regroupement par niveau de capacité implicite qui repose sur une conception universelle du développement de l'enfant.
3. Dans tous les pays du monde, à un moment ou à un autre, les élèves sont répartis dans des groupes qui tiennent explicitement compte de leur niveau de compétence dans une ou plusieurs disciplines, et parfois d'autres éléments (comme leur appétence, leurs objectifs, ...).
4. Certains élèves sont également regroupés car partageant une caractéristique commune, comme des élèves dits « à besoin éducatif spéciaux », des élèves porteurs de certains handicaps ou des élèves à haut potentiel intellectuel. Ces élèves sont la plupart du temps exclus des analyses que nous allons évoquer.

La question qui se pose ici ne concerne donc pas tant l'existence du regroupement par niveau que le moment et la manière dont sont opérés ces regroupements. Pour clarifier les choses, on distinguera ici 5 modes de regroupement qui n'ont d'autre vocation que de servir de points de repère somme toute théoriques. Chaque établissement, chaque classe, chaque groupe d'élèves quel qu'il soit, peut prétendre se placer, tel un curseur, quelque part entre ces points de repère. De la même façon, les mots « homogène » et « hétérogène » doivent être considérés comme les bornes extrêmes et théoriques d'un intervalle continu ; on notera que ces deux termes ne sont jamais définis[2]. Les regroupements dont il sera question ici sont tous explicites.

1. Les établissements de niveau : dans un même pays, les élèves sont regroupés dans des établissements de niveaux différents.

[1] C'est-à-dire regrouper des élèves qui ont, à quelques mois près, le même âge.

[2] Si l'homogénéité parfaite peut se concevoir facilement comme un ensemble d'entités identiques (dans notre cas ce serait une classe constituée de la copie conforme d'un même élève), il n'en va pas de même pour une hétérogénéité « totale ».

2. Les classes de niveau dans toutes les disciplines : dans un même établissement, les élèves sont regroupés dans une même classe de niveau homogène durant toute l'année ; ils ont la même équipe pédagogique et les mêmes pairs.
3. Les groupes de niveau dans certaines disciplines : les élèves sont regroupés dans un groupe de niveau dans une ou deux disciplines (souvent les mathématiques et la litéracie) ; dans ces disciplines, il peut arriver qu'ils changent de niveau, donc de groupe, donc d'enseignant et de pairs au cours de l'année ; les élèves d'une même classe (qui reste l'ensemble de référence) ont toujours les mêmes pairs.
4. Les groupes de niveau intra-classe : dans une même classe, les élèves sont regroupés pour des travaux en petits groupes de niveau ; les élèves ne travaillent pas toujours avec les mêmes pairs mais ils ont toujours le même enseignant[1].
5. Les élèves ne sont jamais regroupés par niveau ; ils ont presque toujours les mêmes pairs et presque toujours les mêmes enseignants durant toute l'année scolaire[2].

Il peut arriver que plusieurs modes de regroupement puissent concerner un même élève ; par exemple des établissements de niveau homogène peuvent répartir les élèves dans des classes de niveau homogène dans toutes ou certaines disciplines. C'est le troisième mode de regroupement qui nous intéresse plus particulièrement.

Les dénominations anglaises et américaines de ces différentes formes de regroupements d'élèves sont présentées dans le tableau X ci-dessous. En anglais, le terme *ability grouping* (regroupement par niveau de compétence ou capacité) est très souvent utilisé comme terme générique. Les chercheurs lui préfèrent souvent *achievment grouping* (regroupement par niveau de résultats scolaires).

On terminera cette présentation des modes de regroupement en remarquant que la grande majorité des études analysées ci-après considèrent trois niveaux de groupes : les élèves regroupés par niveau sont soit dans un groupe de niveau faible, soit dans un groupe de niveau moyen, soit dans un groupe de niveau élevé. On verra plus loin que cela ne correspond pas toujours à la réalité du terrain, mais plutôt à un découpage de l'échantillon analytique permettant d'intégrer des organisations différentes à l'intérieur d'une même étude.

Tableau X : les 5 catégories de regroupement

	Les élèves sont regroupés	**Anglais**
1	dans des établissements de niveau homogène	*Between school grouping (tracking* aux USA)
2	dans des classes de niveau homogène dans toutes les disciplines	*Streaming*
3	dans des classes de niveau homogène pour 1 ou 2 disciplines	*Setting, ability grouping*
4	dans des groupes de niveau homogène au sein d'une classe de niveau hétérogène	*Within-class grouping*
5	ne sont pas regroupés par niveau	*Mixed-ability grouping*

2. Les effets mesurés

Pour faciliter la définition des effets mesurés par les études, quatre dimensions sont considérées dont la description fait l'objet de ce chapitre. On verra plus loin dans les faits que ces effets concernent

[1] Cette organisation est associée au travail en groupes de petit effectif.
[2] Les exceptions peuvent concerner les langues vivantes ou certaines options en France.

- des caractéristiques d'élèves ou d'ensembles d'élèves (qui et quoi ?)
- sont obtenus par comparaison (comment ?)
- mesurés à un temps t ou sur un laps de temps permettant de comparer les évolutions de groupes d'élèves (quand ?)
- et évalués sur des territoires plus ou moins grand (où ?).

La première dimension concerne les caractéristiques évaluées. Ce sont bien sûr et en premier lieu les compétences cognitives des élèves, dites aussi compétences scolaires ou académiques qui nous intéressent plus particulièrement. Elles sont toujours associées à une ou plusieurs disciplines et mesurées par des scores. Les mathématiques vont retenir particulièrement notre attention ici, et ce pour trois raisons : cette discipline est très souvent concernée par des regroupements en classes homogènes dans les pays anglo-saxons ; les résultats des études portant sur cette discipline sont plus facilement comparables d'un pays à l'autre que les résultats portant sur d'autres disciplines (notamment la littéracie) ; les évaluations des compétences académiques mathématiques et leurs analyses sont nombreuses et fournissent une riche base de données. Mais les études se sont intéressées aussi à des caractéristiques non cognitives comme l'estime de soi des élèves. À ces caractéristiques mesurées au niveau individuel, on se doit d'en ajouter une troisième définie cette fois au niveau d'une région ou d'un pays : l'inéquité des organisations scolaires, définie comme la corrélation entre les résultats scolaires et le niveau socioéconomique des élèves.

La deuxième dimension concerne le design des études puisque les effets sont évalués en conduisant des comparaisons. Il s'agit par exemple de comparer les scores en mathématiques d'élèves scolarisés dans un établissement regroupant par niveau aux scores d'élèves scolarisés dans un établissement ne regroupant pas par niveau. De nombreuses études comparent également des élèves de groupes de niveaux différents.

À cela s'ajoute une dimension temporelle et une dimension spatiale. Les comparaisons se font soit à un temps t, soit à deux moments différents ce qui permet alors de comparer des évolutions. On signalera ici que les scores antérieurs des élèves (par exemple leur score avant d'avoir été regroupés par niveau) sont très souvent contrôlés. Enfin ces effets peuvent concerner quelques dizaines d'élèves (en s'intéressant à deux groupes d'élèves dans un même établissement par exemple) ou plusieurs milliers de sujets (quand on compare des élèves échantillonnés dans plusieurs pays).

3. Les types d'études

Toutes les études rassemblées dans cette synthèse peuvent être réparties selon leur design en cinq catégories dont les points forts et les points faibles sont présentés dans le tableau X :

1. les études expérimentales qui sont des essais contrôlés randomisés (ECR) ou quasi-expérimentales (EQE) et qui comparent des organisations différentes dans un même établissement,
2. les études observationnelles qui comparent des établissements organisant différemment les groupes d'élèves dans un même pays,
3. les études observationnelles qui comparent des caractéristiques de systèmes scolaires de plusieurs pays,

4. les synthèses systématiques qui sont susceptibles d'inclure des études des design décrits ci-dessus ; ce sont souvent des méta-analyses qui calculent des tailles d'effet pour chacune des études sélectionnées, puis une ou plusieurs tailles d'effet globales.
5. les études qualitatives qui analysent les résultats d'entretiens et/ou d'observation.

Les études dont nous analyserons les résultats sont majoritairement des études quantitatives. Elles reposent toutes sur des hypothèses théoriques mais leur objectif n'est pas de valider ou non ces dernières. Cette discussion est pourtant indispensable pour qui souhaite comprendre ce qui se passe dans une classe et les recherches qualitatives auront ici un rôle capital à jouer comme nous le verrons à la fin de cette synthèse.

Un aspect peu discuté dans ces différentes études est la direction du changement analysé. Dans certains cas, l'organisation habituelle en classes de niveau hétérogène est abandonnée au profit d'une organisation en classes de niveau homogène. Mais le cas inverse est aussi rencontré. Si l'effet Hawthorne[1] observé dans les études expérimentales tend à favoriser la nouvelle organisation, l'effet inverse peut être observé sur des études observationnelles à plus grande échelle où l'hostilité au changement peut influer sur les résultats (notamment quand on étudie l'impact d'une réforme).

Tableau X. Principales caractéristiques des types d'études

Type d'études	Points forts	Points faibles	Validité interne	Validité externe
1	Explicitation des interventions élevée.	Peu d'études souvent datées, échantillons de petite taille, interventions courtes.	Étude causale, validité interne forte.	Validité externe faible
2	Observations en milieu naturel d'organisations en place depuis parfois plusieurs années. Échantillons de grande taille.	Nombreux facteurs peuvant être à l'origine des effets mesurés en dehors de la caractéristique d'intérêt principal.	Étude non causale, des facteurs de confusion doivent être considérés et contrôlés.	Validité externe forte au niveau d'un État.
3	Concerne des systèmes éducatifs entiers, ce qui correspond à notre question de recherche initiale.	Compare des ensembles d'élèves très divers.	Étude non causale, plusieurs types de regroupement sont parfois agrégés, validité interne faible	Systèmes très différents comparés, validité externe variable.
4	Agrégation de nombreux résultats, échantillon analytique de taille maximale.	Mélange « des pommes et des poires », peu de précision sur les interventions et contextes analysés.	Les statistiques calculées ont une erreur standard faible, mais la diversité des interventions peut diminuer la validité interne.	La validité externe dépend du nombre d'études et des modèles statistiques choisis.
5	Permet de comprendre les effets mesurés en faisant le lien avec des aspects théoriques	Les résultats sont difficiles à résumer de façon objective	Non concerné	Non concerné

[1] Au début des années 1900, des expériences ont été menées qui avaient comme but d'accroitre la productivité des travailleurs de l'entreprise Hawthorne. Les auteurs en ont conclu que c'est le fait de mener une expérience et de témoigner de l'attention envers les sujets de l'expérience qui a été la cause des résultats, et non pas l'intervention elle-même (Adair, 1984).

4. La méthode

En ce qui concerne la recherche internationale dont il sera question dans la seconde partie, six publications sont à la source de cette synthèse :

1. La méta-analyse secondaire[1] de Steenbergen (2016) qui a permis d'identifier 5 méta-analyses primaires (Kulik, 1982 ; Slavin, 1990 ; Kulik, 1992 ; Mosteller, 1996 ; Steenbergen, 2016[2]), puis dans un second temps toutes les études primaires sélectionnées par ces méta-analyses et dont les références étaient publiées.
2. La méta-analyse de Rui (2009) non citée par la méta-analyse secondaire de Steenbergen (2016), puis toutes les études primaires incluses.
3. La méta-analyse d'EEF (2021) puis toutes les études primaires incluses.
4. La synthèse de Dupriez (2010)
5. L'article de Bygren (2016)
6. L'enquête PISA 2022

Les études repérées grâce aux 5 premières sources ont été systématiquement consultées. À l'exception d'un livre (Francis, 2019a), tous les textes cités dans cette synthèse ont été consultés par l'intermédiaire de la bibliothèque Diderot de l'ENS à Lyon[3] (la plupart ayant été téléchargés sur les sites internet des revues). Certains documents, comme des comptes-rendus de conférences, n'ont pas été retrouvés.

Les études observationnelles citées dans les méta-analyses, portant sur les mathématiques et publiées après 1980 ont été systématiquement analysées. D'autres études observationnelles ont été rajoutées quand elles rentraient dans le cadre de notre recherche ; elles se devaient alors d'étudier des regroupements par niveau explicites d'élèves scolarisés dans le secondaire inférieur. C'est pour cette raison que certaines études ont été exclues de notre analyse, comme par exemple celles de Gamoran (1989, 1995) très souvent citées dans les articles, mais qui concernent les élèves américains du secondaire supérieur ou encore celle de Duru-Bellat (1997) qui s'est intéressée à l'effet du niveau moyen des classes sur des élèves qui n'étaient pas explicitement regroupés par niveau. La méta-analyse de Terrin (2023) qui inclue de très nombreuses études du secondaire supérieur voire de l'enseignement supérieur a également été écartée. Seule exception à cette règle, les enquêtes PISA qui concernent des élèves de 15 ans scolarisés en majorité dans le secondaire supérieur en France, mais dont l'utilisation systématique dans les discours politiques impose une présentation.

Comme on le verra plus loin, les ECR et les EQE sont datés, leur examen exhaustif n'a pas paru pertinent et l'analyse de leurs résultats se fondera sur les éléments fournis par leurs méta-analyses. Toutes les tailles d'effet (notée ES pour *Effect Size*) de toutes les études sélectionnées par les méta-analyses précédemment citées sont présentées en annexe X. Elles sont positives quand elles sont en faveur des regroupements en classes de niveau homogène. Pour alléger l'écriture, quand elles sont statistiquement significativement différentes de zéro au seuil de 95%, elles seront décrites comme « statistiquement significatives ».

En ce qui concerne la recherche anglaise analysée en troisième partie, l'article de Hodgen (2023) et deux rapports d'EEF (Roy, 2018a et 2018b) ont servi de point de départ. En l'absence d'une recherche

[1] Méta-analyse de méta-analyses
[2] L'article décrit également les résultats d'une méta-analyse primaire.
[3] https://www.bibliotheque-diderot.fr/

exhaustive de la littérature, il n'est pas certain que toutes les études réalisées sur ce territoire après 2000 aient été repérées. Pour toutes les études identifiées, les articles repérés ont fait l'objet d'une analyse systématique.

Certaines statistiques ont été recalculées ou calculées ici pour la première fois. Ce sont des tailles d'effet globales, des coefficients de corrélations ou des pourcentages. Elles sont repérées dans ce texte en étant déclarés comme « recalculées » pour les premières et « calculées ici » pour les secondes.

Partie 2. La recherche internationale

Les chercheurs s'intéressent à la question du regroupement des élèves par niveau depuis les années 1920, et la plupart des publications contemporaines introduisent leur sujet en évoquant cette longue tradition de recherche. La grande majorité des études dont il sera question ici sont américaines et les niveaux d'étude des élèves n'ont pas été traduits dans cette synthèse (voir le tableau X pour les correspondances avec le système français). Pour satisfaire à notre question de recherche, les articles sélectionnés ici doivent concerner les élèves de la 6ème à la 3ème (système français) et donc les élèves de grade 6 à 9 (système américain). Certains articles qui ont intégré des résultats d'élèves de grade 10 ont tout de même été inclus quand ils permettaient d'évaluer les effets d'une organisation mise en place les années précédentes.

Tableau X : correspondance entre les différents niveaux d'étude en France et aux USA

<table>
<tr><th rowspan="2">Âges</th><th colspan="2">France</th><th colspan="2">USA</th></tr>
<tr><td>Niveau</td><td>Établissement</td><td>Niveau</td><td>Établissement</td></tr>
<tr><td>6 – 7 ans</td><td>CP</td><td rowspan="5">Établissement primaire</td><td>grade 1</td><td rowspan="5">Elementary school ou Primary school</td></tr>
<tr><td>7 – 8 ans</td><td>CE1</td><td>grade 2</td></tr>
<tr><td>8 – 9 ans</td><td>CE2</td><td>grade 3</td></tr>
<tr><td>9 – 10 ans</td><td>CM1</td><td>grade 4</td></tr>
<tr><td>10 – 11 ans</td><td>CM2</td><td>grade 5</td></tr>
<tr><td>11 -12 ans</td><td>6ème</td><td rowspan="4">Collège</td><td>grade 6</td><td rowspan="3">Middle school ou Junior high school</td></tr>
<tr><td>12 – 13 ans</td><td>5ème</td><td>grade 7</td></tr>
<tr><td>13 – 14 ans</td><td>4ème</td><td>grade 8</td></tr>
<tr><td>14 – 15 ans</td><td>3ème</td><td>grade 9</td><td rowspan="2">High school ou Senior high school</td></tr>
<tr><td>15 - 16 ans</td><td>2nde</td><td>Lycée</td><td>grade 10</td></tr>
</table>

On distinguera les études primaires[1] d'une part de leur méta-analyses d'autre part. Il a été décidé ici de présenter en premier lieu les résultats d'études primaires afin de comprendre comment la recherche s'est construite sur le terrain, et de présenter leurs synthèses quantitatives dans un second temps.

1. Les études primaires

Ces études peuvent être réparties dans trois catégories qui prennent appuis sur les trois premiers designs d'études décrits précédemment et qui correspondent plus ou moins à trois périodes historiques : les études expérimentales de petite taille (souvent des thèses) de 1930 à 1990, les études observationnelles sur plusieurs établissements de 1980 à 2010 et les enquêtes internationales à partir des années 2000. Toutes les tailles d'effet des études incluses dans au moins une méta-analyse sont rassemblées en annexe X. Toutes les études primaires sont répertoriées en annexe X.

Dans sa méta-analyse sur les groupes de niveau, Kulik (1982) a sélectionné 51 études expérimentales (ou quasi-expérimentales) : 9 ont été publiées avant 1951, 9 entre 1951 et 1960, 27 entre 1961 et 1970 et 6 entre 1971 et 1980. On peut en conclure que l'« âge d'or » de ce type études se situe dans les années 1960 ; elles ont ensuite laissé la place à des études observationnelles de plus grande ampleur. Un bon exemple de ce type d'étude est l'ECR de Bicak (1964) où 77 élèves de grade 8 d'un même établissement ont été aléatoirement répartis dans une classe hétérogène (groupe contrôle) et un groupe de deux classes homogènes (une de niveau faible et une de niveau élevé) en se basant sur leur QI. Cette expérience a concerné les cours de sciences où les élèves étaient auparavant répartis dans 3

[1] Une étude est dite primaire quand elle fournit des données susceptibles d'être intégrées à une méta-analyse (qui est une étude secondaire).

classes hétérogènes. Les élèves de la classe homogène de niveau élevé sont comparés aux élèves de niveau élevé de la classe hétérogène (et c'est la même chose pour les élèves de faible niveau). Bicak conclue à l'absence d'effet du regroupement sur les élèves, aussi bien pour les bons élèves que pour les plus faibles. Cette étude a été incluse dans quatre méta-analyses et a donné lieu à des calculs de tailles d'effet différents comme on le verra plus loin. Une seconde période débute alors dans les années 1980. Des enquêtes nationales sont conduites au Royaume-Uni et aux Etats-Unis qui ont permis aux chercheurs de réaliser des études observationnelles en milieu naturel et de comparer les élèves scolarisés dans des établissements qui regroupent leurs élèves par niveau avec des élèves scolarisés dans des établissements qui ne regroupent pas leurs élèves par niveau. Les analyses statistiques sont basées sur des régressions linéaires multiples qui contrôlent un certain nombre de covariables (il s'agit le plus souvent du niveau scolaire antérieur, du genre, du niveau socioéconomique et de l'origine ethnique des élèves). Plus récentes que les études expérimentales ou quasi-expérimentales évoquées précédemment, elles ne peuvent cependant prétendre démontrer directement une relation de cause à effet. Toutes les études observationnelles postérieures à 1980 et incluses dans les méta-analyses de Slavin (1990, 1993), de Rui (2009) et d'EEF (2021) ayant publié des résultats pour des élèves du secondaire inférieur sont décrites dans ce chapitre. Elles concernent toutes l'enseignement des mathématiques, parfois également l'enseignement de l'anglais. D'autres études citées dans l'article de Bygren (2016) ont été également incluses dans cette description. Aucune des études citées par Dupriez (2010) en dehors de celles sélectionnées par les méta-analyses précédemment citées n'ont été retenues, car elles ne répondaient pas aux critères définis ici. Les études sont présentées par ordre chronologique et leurs principales caractéristiques sont résumées dans le tableau X (les tailles d'effet présentées ont été calculées par les méta-analyses ayant retenu ces études).

Kerckhoff (1986) utilise les données de la National Child Development Study (NCDS) qui a collecté pendant presque 20 ans des informations sur tous les enfants nés entre le 3 et 9 mars 1958 au Royaume-Uni. Les résultats d'élèves fréquentant des établissements où ils étaient regroupés par niveau (groupe traitement) ont été comparés aux résultats d'élèves fréquentant des établissements qui n'appliquaient pas cette organisation (le groupe contrôle). Le score en mathématiques (et en anglais qui fait l'objet d'une analyse séparée) des élèves âgés de 16 ans est la variable dépendante d'une régression linéaire multiple tenant compte du niveau initial des élèves. Les résultats montrent que les scores des élèves des classes de niveau faible des établissements regroupant par niveau sont plus faibles que les scores d'élèves de niveaux initiaux comparables scolarisés dans des établissements ne regroupant pas les élèves par niveau. Et que l'inverse est vrai : les scores des élèves des classes de niveau élevé des établissements regroupant par niveau sont plus élevés que les scores d'élèves de niveaux initiaux comparables scolarisés dans des établissements ne regroupant pas les élèves par niveau. Ces conclusions seront critiquées par Slavin (1990, voir plus loin) qui met en doute la capacité d'un modèle statistique, quel qu'il soit, de prendre en compte des niveaux scolaires antérieurs qui sont parfois très différents.

Hoffer (1992) utilise les données de la Longitudinal Study of American Youth (LSAY) et compare les scores d'élèves qui fréquentent des établissements regroupant leurs élèves par niveau, à ceux d'élèves d'établissements ne regroupant pas les élèves par niveau, en sciences et en mathématiques. L'échantillon est séparé en quintiles selon les scores obtenus au grade 7. Dans chaque quintile, l'évolution des scores entre les grades 7 et 9 des élèves regroupés par niveau est comparée à l'évolution des scores des élèves non regroupés par niveau. Selon Hoffer, cette méthode est plus apte à évaluer

les effets du regroupement sur les scores que la méthode utilisée par Kerckhoff (1986). Il n'en reste pas moins que sa conclusion reste la même : le regroupement par niveau est favorable aux élèves de niveau élevé et défavorable aux élèves de niveau faible.

En 1996, Argys utilise les données publiées en 1988 de la National Education Longitudinal Study (NELS). Le score en mathématiques des élèves de grade 10 est analysé en tenant compte de leur niveau initial (mesuré par leur score obtenu au grade 8). Les enseignants de mathématiques ont répondu à un questionnaire pour déterminer si leur classe (de grade 8 et de grade 10) est de niveau global élevé, ou moyen, ou faible, ou encore hétérogène. La comparaison des moyennes des scores prédits par les modèles linéaires entre ces différents types de classe montre à nouveau que les élèves de niveau faible réussissent mieux dans les classes hétérogènes et que c'est l'inverse pour les élèves de niveau élevé.

Hawkins (1999) réalise une étude avant/après sur un seul groupe d'élèves, donc sans groupe contrôle, et analyse les scores des élèves scolarisés dans une classe hétérogène au grade 8 après avoir été scolarisés au grade 7 dans des groupes de niveau homogène dans un établissement américain dans lequel le chercheur enseigne. Ces résultats doivent donc être considérés avec précaution. Aucun effet sur les scores n'est montré.

Betts (2000) utilise les mêmes données que Hoffer (1992) et propose également plusieurs régressions linéaires pour comparer les élèves scolarisés dans des établissements regroupant par niveau en mathématiques à des élèves de même niveau initial mais fréquentant des établissements ne regroupant pas par niveau en mathématiques. Cette fois et dans les deux types d'établissements, les enseignants des élèves de grade 9 ont évalué le niveau de leur classe sur une échelle comportant cinq niveaux, allant de « beaucoup plus élevé que le niveau moyen de mon établissement » (niveau 1) à « beaucoup plus faible que le niveau moyen de mon établissement » (niveau 5). Pour chacun de ces niveaux (et donc pour des élèves qui ont des niveaux scolaires similaires selon l'auteur), une régression linéaire compare les scores des élèves qui sont dans des établissements regroupant par niveau avec les élèves qui sont dans des établissements ne regroupant pas par niveau. La conclusion reste dans la lignée des conclusions précédentes, avec cependant un effet différentiel des niveaux de regroupement moins marqué. On notera tout de même que pour les classes de niveau faible comme pour les classes de niveau élevé (selon le jugement de l'enseignant), les élèves scolarisés dans des établissements regroupant par niveau ont de meilleurs résultats que les autres, mais cette différence n'est pas statistiquement significative. Inversement, les élèves des classes de niveau moyen des classes homogènes ont de moins bons résultats que les élèves des classes de niveau moyen des établissements ne regroupant pas par niveau, et cette fois la différence est statistiquement significative. Selon Betts, ses résultats se rapprochent plus de ceux de Kerckoff (1986) et de Hoffer (1992) que de Slavin (1990) (voir plus loin).

Hallinan (2000) conduit une étude longitudinale sur des élèves de grade 9 dans 6 établissements du Midwest aux USA. L'objectif de cette étude n'est pas de comparer les résultats des élèves groupés en classe de niveau à ceux d'élèves en classes hétérogènes : la question de recherche porte sur l'effet pour un élève d'être placé dans un groupe de niveau supérieur à son groupe d'origine. Ces établissements regroupaient leurs élèves en mathématiques sur cinq niveaux différents. Certains élèves de l'échantillon analytique étaient scolarisés au grade 8 dans des classes hétérogènes et les coefficients de régression de la variable « avoir été ou non dans une classe hétérogène » permettent à Rui (2009)

qui sélectionne cette étude dans sa méta-analyse, de calculer une taille d'effet. Les résultats montrent que les élèves qui viennent d'établissement qui ne regroupent pas par niveau au grade 8 ont de meilleurs résultats que les autres après avoir tenu compte des scores prétests, du genre, de l'origine ethnique et du niveau socioéconomique. Mais ces différences ne sont significatives que pour les scores des élèves des classes de niveau élevé. Hallinan répond enfin à sa question initiale et conclue que faire passer un élève dans un groupe de niveau plus élevé a un effet positif sur ses résultats en mathématiques (comme en anglais), quel que soit son niveau initial.

Figlio (2002) utilise les mêmes données qu'Argys (1996) pour étudier l'évolution des scores des élèves du grade 8 au grade 10. Dans un premier temps il reproduit les mêmes analyses que Hoffer (1992) et Betts (2000), et retrouve des résultats conformes à ses prédécesseurs. Dans un second temps, tous les élèves de l'échantillon sont classés par ordre croissant de niveau scolaire en mathématiques (scores au grade 8) ce qui lui permet de diviser l'échantillon total en terciles. Pour chacun de ces terciles Figlio analyse l'effet d'être (ou non) scolarisé dans un établissement qui regroupe par niveau (au grade 8 et au grade 10) sur l'évolution des scores entre les grades 8 et 10. Les résultats montrent que le regroupement n'a pas d'effet sur les résultats académiques des élèves, et ce quel que soit le niveau scolaire de l'élève. Figlio souligne qu'il y a là une contradiction avec les conclusions précédentes. Il poursuit son analyse en introduisant d'autres variables. En effet, les recherches précédentes n'ont pas pris en compte la possibilité que le placement d'un élève dans une établissement qui regroupe (ou non) soit liée aux autres variables indépendantes. Trois variables sont alors introduites dans un nouveau modèle : le nombre de cours académiques requis pour l'obtention du diplôme d'Etat, le nombre d'établissements dans le comté et la fraction des votants pour Reagan en 1984, toujours dans le comté. Ces éléments permettent à Figlio de montrer que l'effet du regroupement est cette fois positif pour les élèves de faible niveau, nul pour les élèves de niveau moyen et négatif (mais non statistiquement significatif) pour les élèves de niveau élevé. C'est ce résultat inhabituel qui est repris par Bygren (2016) qui cite cette étude dans son article. Ces trois dernières variables sont spécifiques au système américain, et cette conclusion ne peut sans doute pas concerner des élèves français.
On termine cette série d'études avec Burris (2006) dont l'étude longitudinale (série chronologique interrompue) analyse les données de 6 cohortes dans le comté de Naussau aux USA. Les trois premières cohortes sont entrées au secondaire (grade 6) en 1995, 1996, 1997 et ont suivi des cours de mathématiques dans des classes de niveau homogène. Les trois suivantes sont entrées au secondaire en 1998, 1999 et 2000, et ont suivi des cours en classes hétérogènes avec un programme en mathématiques « accéléré ». Burris utilise un modèle de régression logistique qui prend en compte plusieurs covariables (notamment le score initial, le niveau socio-économique, l'origine ethnique, le genre). Ses résultats montrent que pour tous les sous-groupes d'élèves que l'on peut considérer, avoir été scolarisé dans des classes hétérogènes et avoir bénéficié d'un programme accéléré en maths a été favorable aux élèves, quel que soit leur niveau initial. Le fait que deux changements simultanés soient analysés (le passage d'une organisation avec des classes homogènes à une organisation avec des classes hétérogènes d'une part et un changement de programme d'autre part) et la durée importante de l'étude ne permettent pas d'affirmer le lien causal de façon univoque.
Toutes ces études publiées aux Etats-Unis ont cherché à éclairer le débat sur le *detracking*[1], sujet de nombreuses discussions à la fin des années 1990 dans ce pays et trois études citées ci-dessus (Argys,

[1] Passer d'une organisation avec des classes regroupant par niveau à une organisation avec des classes hétérogènes

1996 ; Hawkins, 1999 ; Burris, 2006) ont posé leurs questions de recherche en ce sens. Leurs résultats ont été parfois intégrés dans des méta-analyses dont les conclusions sont présentées à la fin de cette partie (tableau X).

Tableau X : caractéristiques des 8 études observationnelles

Auteur (date)	**Données**	**Type de regroupement**	**Taille échantillon**	**Variable dépendante**	**Cité par**	**Taille d'effet**
Kerckhoff (1986)	Enquête longitudinale NCDS, 1969 et 1974	*Streaming* et *setting* non distingués	9 340	Grade 10 (1974)	Slavin (1990), Bygren (2016) Dupriez (2010)	Ensemble : 0,03 Élevé : n.p Moyen : n.p Faible : n.p
Hoffer (1992)	Enquête longitudinale LSAY de 1987 à 1989	Pas d'information	1 900	Score grade 8 et grade 9	Slavin (1993), EEF (2021), Bygren (2016), Dupriez (2010)	Ensemble : -0,01 (Slavin) ; -0,15 (EEF) Élevé : 0,13 Moyen : -0,03 Faible : -0,27
Argys (1996)	Enquête longitudinale NELS à partir de 1988	Classes de niveau élevé, moyen, faible ou classe hétérogène (questionnaire enseignant)	3 405	Score grade 10	Rui (2009), Bygren (2016), Dupriez (2010)	Ensemble : 0,08 Élevé : 0,44 Moyen : 0,16 Faible : -0,25
Hawkins (1999)	Étude avant / après de 1996 - 1998	Pas d'information	une classe	Score grade 7 et 8	Rui (2009)	Ensemble : -0,06 Élevé : n.p Moyen : n.p Faible : n.p
Betts (2000)	Enquête longitudinale LSAY de 1987 à 1989	Regroupement en maths ; niveau de la classe (entre 1 et 5) dans l'établissement ; répartition des scores en quartiles	5 442	Élèves de grade 7 à 10. La variable dépendante n'est pas précisée.	Bygren (2016), Dupriez (2010)	Ensemble : n.p Élevé : n.p Moyen : n.p Faible : n.p
Hallinan (2000)	Enquête longitudinale, 1989 et 1990	Cinq niveaux de regroupements en maths	2 574	Score grade 9	Rui (2009)	Ensemble : -0,29 Élevé : -0,32 Moyen : n.p Faible : -0,27
Figlio (2002)	Enquête longitudinale NELS à partir de 1988	Regroupement en maths ; niveau de la classe (élevé, moyen, faible, hétérogène) dans l'établissement	5 948	Évolution des scores entre grade 8 et 10	Bygren (2016)	Ensemble : n.p Élevé : n.p Moyen : n.p Faible : n.p
Burris (2006)	6 séries chronologiques interrompues entre 1995 et 2006	3 cohortes avec groupes de niveau, puis trois cohortes avec classes hétérogènes + *accelerated Maths*	985	Scores grades 8, 9, 10, 11 et 12	Rui (2009), EEF (2021)	Ensemble : -0,28 (Rui) ; -,32 (EEF) Élevé : -0,22 (Rui) Moyen : n.p Faible : n.p

n.p. : non publié ; ensemble : tous les élèves ; Élevé : taille d'effet du groupe de niveau élevé (même chose pour moyen et faible).

À partir des années 2000, les enquêtes PISA (Programme for International Student Assessment) ont produit un nombre considérable de données qui ont été utilisées par de nombreux chercheurs. Contrairement aux études citées au chapitre précédent, le champ de l'analyse inclue de très nombreux pays (par exemple européens). Ces enquêtes sont des études transversales qui ne récoltent presqu'aucune données rétrospectives et analyse les performances des élèves de 15 ans révolus[1] : elle sort donc de notre cadre de recherche. Mais le rôle important qu'elles jouent actuellement dans les discours politiques impose que leurs résultats soient présentés. Les premières conclusions de l'enquête PISA 2022 ont été publiées le 5 décembre 2023 (le même jour Gabriel Attal publiait sa lettre présentant les grands traits de sa réforme Choc des savoirs). Un chapitre entier est consacré à la sélection et à la stratification des élèves et certaines données concernent les regroupements par niveau (OCDE, 2022, p.158 – 159)[2]. Elles ont été récoltées à partir des questionnaires complétés par les chefs des établissements participants à l'enquête. Interrogés sur les regroupements par niveau, ils devaient sélectionner l'une de ces trois réponses : « pas de regroupement », « regroupement en classes de niveau » ou « regroupement dans la classe ». Et si une forme de regroupement était reconnue, ils précisaient s'il concernait « toutes les disciplines » ou « quelques disciplines » (aucune information n'est donnée au sujet de ces dernières). Des coefficients de corrélations ont été calculés (tableau II.4.1 p.128) qui estiment l'association entre le regroupement en classes de niveau pour toutes les disciplines et les scores en mathématiques d'une part mais aussi l'indice d'équité[3] d'autre part. Ils n'ont pas été calculés pour les « regroupements en classes de niveau pour quelques disciplines » qui nous intéresse. La seule conclusion du rapport que l'on puisse citer ici indique que « dans les pays équitables et performants, …/… peu d'élèves sont regroupés dans des classes de niveau ». Il est probablement sous-entendu ici que ce sont des regroupements pour toutes les disciplines. Les auteurs déconseillent d'inférer à partir de ces résultats (p.160) : « Établir des inférences causales n'est pas conseillé, étant donné la nature transversale de l'enquête PISA et la complexité des relations entre les politiques de stratifications et les résultats des élèves ». Une autre conclusion attribuable à l'OCDE vient d'Andreas Schleicher (qui supervise les enquêtes PISA depuis plusieurs années) cité en ces termes dans un rapport EEF (Roy, 2018 b, p.6) : « Selon les données de l'enquête PISA, lorsque les établissements utilisent les regroupements par niveau pour des disciplines spécifiques, de manière sélective et en autorisant une mobilité des élèves entre les groupes, cela n'a aucun effet néfaste sur les résultats, mais lorsqu'ils appliquent le regroupement à toutes les disciplines - c'est-à-dire mettent en œuvre le *streaming* - cela creuse les disparités socio-économiques. » Dans la note sur la France[4], on lit que « les données de PISA 2022 montrent que les pays où une proportion plus élevée d'élèves sont regroupés par compétences entre les classes d'un établissement (ou dans chaque classe) uniquement pour certaines matières tendent à obtenir des performances plus élevées en mathématiques (corrélation de + 0,32 quand la comparaison porte sur ceux regroupés entre les classes d'un établissement et + 0,51 quand la comparaison porte sur ceux regroupés dans chaque classe). » La formulation laisse entrevoir ici la possibilité d'un lien de causalité. Ils ont été recalculés et les résultats publiés ont été retrouvés. Les mêmes coefficients de corrélation ont également été calculés ici quand on se limite aux seuls pays de

[1] S'ils n'ont pas redoublé, les élèves français sont en classe de seconde (générale et technologique ou professionnelle) ou en CAP depuis 7 mois.

[2] Elles sont rassemblées dans 7 tableaux (du tableau II.B1.26 au tableau II.B1.4.32). Toutes les données PISA 2022 sont en accès libre https://www.oecd.org/en/data/datasets/pisa-2022-database.html

[3] L'équité se définit d'abord comme un objectif pour l'OCDE : tous les élèves, quelle que soit leur origine, devraient avoir une chance équitable de réaliser leur plein potentiel (OCDE, 2023, p.44). C'est également un indice égal à 100 - R^2, R^2 étant le coefficient de détermination de la régression des scores en mathématiques sur l'indice socioéconomique utilisé par l'OCDE comme indice d'inéquité.

[4] https://www.oecd.org/pisa/publications/Countrynote_FRA_French.pdf

l'OCDE cette fois, ce qui est d'usage dans les rapports PISA (tableau X). On notera que ces derniers sont moins favorables au regroupement que ceux publiés par la note française.

Tableau X : Coefficients de corrélation entre pourcentages de regroupement et scores en maths[1]

	Regroupés par niveau dans différentes classes		Regroupés par niveau dans leur classe	
	Pour toutes les disciplines	Pour quelques disciplines	Pour toutes les disciplines	Pour quelques disciplines
OECD	-0,12*	0,20**	-0,56*	0,34**
Tous les pays	-0,42*	0,32***	-0,54*	0,51***

* : publiés dans le volume II (p.128, tableau II.4.1) ; ** : calculés ici ; ***: publiés dans la Note PISA 2022 pour la France (p.16)

2. Les méta-analyses

À partir des années 1980, des méta-analyses ont été conduites dans le but d'apporter des réponses aux questions qui nous préoccupent. La procédure générale suivie par ces synthèses systématiques et quantitatives repose sur 4 étapes :

1. la recherche exhaustive et systématique de publications répondant à une question de recherche
2. la sélection d'études primaires à partir de critères définis *a priori*
3. le calcul de tailles d'effet pour chacune de ces études
4. le calcul de tailles d'effet globales (pour l'ensemble des études ou pour des sous-groupes d'études partageant une même caractéristique).

Ces synthèses évaluent de façon objective les effets d'interventions ou de certaines caractéristiques de l'enseignement en évitant notamment le picorage (*cherry-picking*). Les méthodes qui concernent aussi bien la recherche de la littérature, que l'évaluation et la sélection des études et l'analyse statistique des données se sont développées et affinées au cours des 40 dernières années (Roques, 2022) et les résultats des huit méta-analyses présentés ci-dessous par ordre chronologique doivent être replacés dans ce contexte historique. Seules les conclusions sur les scores académiques sont présentées. Certaines méta-analyses ont étudiés des caractéristiques non cognitives, mais les auteurs reconnaissants eux-mêmes ne pouvoir en tirer de conclusions cohérentes, il a été décidé ici de ne pas les présenter (voir la partie 3 pour l'analyse de ces caractéristiques). Leurs principales caractéristiques sont résumées dans le tableau X et les tailles d'effet globales calculées par chacune des 8 méta-analyses sont présentées dans le tableau X. Elles permettent de comparer l'ensemble des élèves regroupés en classes de niveau par rapport à l'ensemble des élèves non regroupés par niveau. Mais aussi de comparer des élèves d'un certain groupes de niveau (soit élevé, moyen ou faible) à des élèves du groupe contrôle de même niveau antérieur. La liste de toutes les études primaires sélectionnées est présentée en annexe X et la liste de toutes les tailles d'effet publiées en annexe X.

Kulik est le premier à publier en 1982 une méta-analyse sur l'effet du regroupement des élèves dans des classes homogènes à l'intérieur d'un même établissement dans le secondaire. Sont exclues de cette synthèses les études sur des regroupements intra-classe et les comparaisons entre établissements proposant des organisations différentes. Les études doivent inclure un groupe traitement et un groupe contrôle, ces deux groupes doivent avoir un niveau initial comparable et la mesure des performances

[1] Recalculés ou calculés à partir des tableaux II.B1.4.29 et I.B1.2.1.

ne doit pas en favoriser injustement un (habituellement le groupe traitement). La significativité statistique qu'il convient d'associer aux tailles d'effet est parfois mentionnée dans le texte, mais aucun intervalle de confiance ni valeur-p ne sont publiés. La liste des 52 études sélectionnées n'est pas publiée (ni bien sûr les tailles d'effet associées à chacune de ces études). Kulik constate que le regroupement par niveau n'a aucun effet. Cet auteur publiera d'autres méta-analyses, par exemple pour les élèves scolarisés en primaire (Kulik, 1984) ou en intégrant également les regroupements intra-classe (Kulik, 1987). Bien que datées, ces méta-analyses restent fréquemment citées par les chercheurs encore de nos jours.

Tableau X : caractéristiques des 7 méta-analyses primaires

Auteurs (année)	**Regroupements**	**Niveau d'étude**	**Design des études primaires**	**1. Calculs des tailles d'effet des études 2. Calculs des tailles d'effet globales 3. Estimation de la significativité statistique**
Kulik (1982)	Inter-classe dans le même établissement	Secondaire	Études de comparaison	1. Δ de Glass 2. Moyennes des ES 3. Parfois dans le texte
Slavin (1990)	Tous	Secondaire	ECR et EQE ; études longitudinales	1. Δ de Glass 2. Médiane des ES 3. Parfois dans le texte
Kulik (1992)	Inter-classe dans le même établissement	Primaire et secondaire	Études de comparaison	1. Pas d'information 2. Pas d'information 3. Parfois dans le texte
Mosteller (1996)	Tous	Primaire et secondaire	ECR	1. Pas d'information 2. Moyenne pondérée par la taille d'échantillon 3. Non
Rui (2009)	Tous	Primaire et secondaire	Tous	1. Pas d'information 2. Modèles de l'effet fixe et modèle des effets aléatoires 3. Quelques intervalles de confiance et valeurs-p
Steenbergen (2016)	Inter-classe	Primaire et secondaire	ECR	1. g de Hedges 2. Modèle des effets mixtes 3. Intervalles de confiance et valeurs-p
EEF (2021)	Inter-classe	Primaire et secondaire	Tous	1. d de Cohen sans tenir compte des scores prétests 2. Modèle des effets aléatoires 3. Intervalles de confiance

Tableau X : tailles d'effet globales des 8 méta-analyses sur les résultats scolaires

Méta-analyses	**Tous les élèves**	**Élèves des groupes de niveau**		
		élevé	**moyen**	**faible**
Kulik (1982)*	0	0,00	-0,06	0,002
Slavin (1990)	-0,02	0,01	-0,08	-0,02
Kulik (1992)	0,03	0,10**	-0,02	-0,04
Mosteller (1996)	0	0,08	-0,04	-0,06
Rui (2009)***	-0,087** ; -0,202**	-0,075** ; -0,170	n.c.	-0,113** ; -0,283**
Steenbergen (2016) [#]	-0,03	0,06	-0,04	0,03
Steenbergen (2016) [##]	0,15 **	n.p.	n.p.	n.p.
EEF (2021)	0,020[###]	n.p.	n.p.	n.p.

Les tailles d'effet sont positives quand elles sont en faveur des regroupements en classes de niveau homogène. n.p. : non publiée. * : résultats pour les élèves représentatifs de la population générale ; ** : statistiquement significatif ; *** = modèle de l'effet fixe puis des effets aléatoires ; [#] : méta-analyse secondaire ; [##] : méta-analyse primaire ; [###] : la taille d'effet affichée sur le site est égale à 0,028 (non significative) et a été corrigée ici (communication personnelle)

En 1990, c'est au tour de Slavin de publier sa première méta-analyse sur les regroupements des élèves scolarisés au secondaire dans des groupes homogènes. Elle fait suite à sa méta-analyse publiée 3 années plus tôt sur le même sujet concernant les élèves scolarisés en primaire (Slavin, 1987). Il s'agit plus précisément d'une *best evidence synthesis,* qui ajoute aux résultats habituels d'une méta-analyse une description de chacune des études incluses dans la méta-analyse. Elle mérite d'être décrite dans le détail car c'est certainement l'article le plus souvent cité dans toutes les études dont nous présentons les résultats[1]. Cette fois, les regroupements intra-classe sont également considérés et non différenciés des regroupement inter-classe dans l'analyse. L'effet sur les résultats scolaire (dans plusieurs disciplines) est étudié en tenant compte des scores initiaux. La significativité statistique qu'il convient d'associer aux tailles d'effet est parfois mentionnée dans le texte, mais aucun intervalle de confiance ni valeur-p ne sont publiés. Parmi les 29 études sélectionnées, six études sont antérieures à 1936, et plusieurs sont des thèses. Slavin indique que toutes les études des méta-analyses de Kulik (1982, 1987) sont incluses sauf deux. Pour 12 études, les élèves étaient regroupés par niveau dans toutes les disciplines, pour les 17 autres seules quelques disciplines étaient concernées. Aucune taille d'effet n'est publiée pour ces deux sous-groupes d'études, Slavin mentionnant seulement une absence de différence significative. D'autres caractéristiques, comme le nombre de niveaux, la durée et la localisation de l'étude ou la discipline analysée n'ont pas montré avoir une influence sur l'effet (là encore, aucune taille d'effet n'est publiée). Pour chaque étude, les tailles d'effets ont été calculées pour l'ensemble des élèves d'une part, mais aussi pour les élèves des groupes de niveau élevé, pour les élèves des groupes de niveau moyen et pour les élèves des groupes de niveau faible d'autre part quand cela était possible. Dans ce dernier cas de figure, la question du groupe contrôle n'est pas clarifiée (on a vu que dans certaines études primaires, les élèves regroupés dans le groupe de niveau élevé étaient comparés aux élèves de niveau antérieur similaire mais scolarisés dans des classes hétérogènes par exemple). Les tailles d'effet globales sont des médianes. Elles ont été calculées ici pour les études postérieures à 1936 (c'est-à-dire en excluant les études les plus datées) pour l'ensemble des élèves (tableau X) et pour les différents niveaux de groupes (tableau X). On remarque alors que les différences entre les tailles d'effet pour des groupes de niveau différents sont plus marquées.

Tableau X. Tailles d'effet globales pour l'ensemble des élèves

	Effectifs	Taille d'effet totale
Toutes les études	20	-0,02
Tous les ECR et EQE	13	-0,06
Toutes les études postérieures à 1936*	14	-0,04
Tous les ECR et EQE postérieures à 1936*	10	-0,05

* : calculées ici.

Tableau X. Tailles d'effet globales pour les sous-groupes d'élèves

	Effectifs	Tailles d'effets des		
		groupes de niveau élevé	groupes de niveau moyen	groupes de niveau faible
Toutes les études	15	0,01	-0,08	-0,02
Tous les ECR et EQE	11	0,05	-0,10	-0,06
Toutes les études postérieures à 1936*	11	0,07	-0,10	-0,06
Tous les ECR et EQE postérieurs à 1936*	9	0,05	-0,13	-0,10

* : calculées ici.

[1] Le nombre de citations mentionnées dans Google Scholar en aout 2024 était de 1 742 ; il était égal à 965 en aout 2016 (Steenbergen, 2016). Pour comparaison, la méta-analyse de Kulik (1982) est citée 999 fois et l'étude de Kerckhoff (1986) 479 fois.

Slavin conclue sur l'absence d'effet des regroupements par niveau et rejette l'hypothèse d'un effet différentiel : contrairement à ce qui est parfois dit, les élèves de niveaux élevés ne sont pas gagnants dans les regroupements par niveau et les élèves de niveau faible ne sont pas perdants (p.486). Il se montre également dubitatif sur les conclusions présentées par certaines des études incluses, leur reprochant de mal évaluer les effets du regroupement sur les scores pour des élèves appartenant à des groupes d'un niveau déterminé (faible, moyen ou élevé). Ainsi aucune taille d'effet pour les trois groupes de niveau analysés dans l'étude de Kerckhof (1986) n'est calculée, les méthodes analytiques ne permettant pas selon Slavin de tenir réellement compte des niveaux initiaux (scores prétests) qui sont trop différents. Il ajoute que les élèves de niveau élevé augmentent plus leurs scores que les élèves de niveau faible, qu'ils soient regroupés dans des classes homogènes ou hétérogènes (p.490).

Slavin publie une autre méta-analyse en 1993 sur la même question mais en limitant le champ de l'étude au secondaire inférieur : trois des études de sa précédente méta-analyse sont ainsi exclues[1] et l'étude de Hoffer (1992) qui venait d'être publiée est rajoutée. Mais ses conclusions restent les mêmes.

En 1992, Kulik publie une dernière méta-analyse. Cette fois, les élèves sont scolarisés en primaire et au secondaire, et plusieurs organisations de l'enseignement sont analysées : les classes de niveau dans un même établissement, le saut de classe, les classes de niveau regroupant plusieurs niveaux d'études ou les groupes de niveau dans la classe. Les études repérées en 1982, 1984 et 1987 (qui ne sont toujours pas citées) sont à nouveau évaluées avec des critères qui ont été affinés en suivant quelques-unes des critiques émises par Slavin et certaines des études analysées par ce dernier sont ajoutées à cette nouvelle méta-analyse. La conclusion reste inchangée en ce qui concerne les groupes de niveau dans les établissements.

En 1996 Mosteller publie une méta-analyse qui inclue 10 essais contrôlés randomisés. Parmi eux, 8 étaient déjà sélectionnés par Slavin (1990). Les deux autres sont une étude de Vakos (1969) exclue par Slavin pour avoir étudié le regroupement des élèves en groupes homogènes seulement 2 jours par semaine (les 3 autres jours, les élèves étant en classes hétérogènes). Et l'étude de Wardrop (1967) qui ne pouvait pas avoir été retenue par Slavin car concerne des élèves de grade 3. Mosteler reprend à son compte les résultats globaux de Slavin et de Kulik, et conclue à une absence d'effet des groupes de niveau homogène sur les scores de l'ensemble des élèves. Il se montre très prudent sur l'effet faiblement positif pour les élèves des groupes de niveau élevé comme sur l'effet faiblement négatif pour les élèves des groupes de niveau faible.

Dans sa méta-analyse de 2009, Rui a pour objectif d'analyser les effets de l'abandon du regroupement des élèves par niveau (*detracking*). Aucune définition sur les types de regroupements n'est donnée, et les élèves peuvent être scolarisés de la maternelle à la fin du lycée. Les publications sont postérieures à 1970, et une préférence est donnée aux études expérimentales ou quasi-expérimentales (mais certaines études non expérimentales sont sélectionnées). À l'instar de Slavin (1990), Rui conduit une *best evidence synthesis*. À partir des 15 études sélectionnées, 22 tailles d'effet sont calculées : elles sont positives quand elles favorisent les groupes hétérogènes, contrairement à toutes les autres méta-analyses et leur signe a été changé ici. Aucune information n'est donnée sur leur mode de calcul. Les tailles d'effet globales sont estimées pour plusieurs sous-groupes d'études, en utilisant soit le modèle des effets aléatoires soit le modèle de l'effet fixe. Certains intervalles de confiance et valeurs-p sont publiés. Une analyse de l'hétérogénéité montre que l'erreur d'échantillonnage n'explique pas toute la

[1] Ces études analysaient des scores d'élèves de *grade* 10 et plus.

variation des tailles d'effet. Selon Rui, les résultats des analyses de sous-groupes doivent être considérés avec prudence, car les tailles d'effet sont très hétérogènes et le nombre d'études dans chaque groupe souvent inférieur à 10. La conclusion reste toujours dans la lignée des études précédentes : les élèves des groupes de niveau faible semblent profiter de l'abandon des groupes de niveau.

En 2016, Steenbergen publie une méta-analyse secondaire, c'est-à-dire une méta-analyse de méta-analyses, accompagnée d'une petite méta-analyse d'études primaires. La qualité des méta-analyses sélectionnées est évaluée. Tous les niveaux scolaires sont concernés. Les résultats concernant le regroupement des élèves en classes de niveau à l'intérieur de leur établissement font l'objet d'une analyse de sous-groupe ; ils proviennent :

- des 5 méta-analyses des Kulik (1982, 1984, 1985[1], 1987, 1992), des 3 méta-analyses de Slavin (1987, 1990, 1993) et de la méta-analyse de Mosteller (1996). La qualité des méta-analyses des Kulik est estimée comme étant faible, les autres méta-analyses sont de qualité modérée.
- D'une méta-analyse qui s'intéresse aux élèves à haut potentiel (Goldring, 1990)
- De deux méta-analyses qui ne concernent que les élèves du primaire (Henderson, 1989 ; Lou, 1996)
- Et d'une méta-analyse qui a fait l'objet d'une présentation à une conférence, qui n'a pas été publiée dans une revue et n'a pas pu être retrouvée ici (Noland, 1986).

La méta-analyse de Rui (2009) n'est pas incluse[2]. Comme nous l'avons déjà constaté, les méta-analyses primaires sélectionnent en partie les mêmes études, ce qui donne lieu à des recouvrements entre leurs échantillons analytiques (*overlapping*). De ce fait, les tailles d'effet globales (recalculées par Steenbergen) pour chacune d'entre elles ne sont pas indépendantes. Si une description de ces recouvrements est bien proposée en annexe par l'auteur, la taille d'effet globale calculée reste tout de même égale à la moyenne pondérée par l'inverse des erreurs standards des tailles d'effet. Or pour être valide, cette méthode de calcul suppose que les tailles d'effet sont indépendantes. On notera que les méta-analyses secondaires autrefois mises en œuvre par EEF (voir ci-dessous) ont été abandonnées en 2021 en raison des biais méthodologiques importants qui ne sont la plupart du temps pas contrôlés. En ce qui concerne les quatre méta-analyses décrites ci-dessus (Kulik, 1982 ; Slavin, 1990 ; Kulik, 1992 ; Mosteller, 1996), les résultats obtenus par Steenbergen sont similaires à ceux publiés dans les articles originaux et aucune taille d'effet n'est statistiquement significative.

Steenbergen (2016) présente également les résultats d'une petite méta-analyse affirmant inclure uniquement des ECR issus des trois méta-analyses de Slavin (1987, 1990, 1993). Les études concernant le regroupement des élèves en classes de niveau à l'intérieur des établissements font l'objet d'une analyse de sous-groupes. Les cinq ECR inclus analysent des scores d'élèves scolarisés au secondaire. Parmi eux, l'étude de Mikkelson (1962) avait pourtant été exclue par Slavin pour n'avoir analysé que le regroupement des élèves de niveau élevé dans des classes homogènes. Inversement, l'étude de Marascuilo (1972) incluse par Slavin dans ses méta-analyses, est exclue par Steenbergen qui indique que la taille d'effet n'a pas pu être calculée. Les tailles d'effet calculées pour chacune des études sont différentes de celles publiées par ailleurs (voir tableau X). Elles sont notamment toutes positives pour Steenbergen. Si des valeurs différentes peuvent s'expliquer par des méthodes de calcul différentes, les signes eux ne devraient pas changer (l'effet du regroupement est soit positif quand les scores des élèves

[1] Steenbergen fait référence ici à une présentation des Kuliks à la Convention annuelle de l'American Psychological Association (APA) dont le texte n'a pas été retrouvé.
[2] Aucune information ne permet de dire si elle a été exclue ou juste non repérée.

regroupés par niveau sont en moyenne supérieurs aux scores des élèves non regroupés, soit négatif dans le cas contraire). On ne peut écarter ici l'hypothèse d'une erreur de calcul, hypothèse que Steenbergen semble ne pas considérer puisqu'il conclue que « les méta-analyses précédentes ont sous-estimés l'effet (du regroupement par niveau) (Steenbergen, 2016 , p.890). »

Tableau X. Tailles d'effet des études incluses par Steenbergen (2016) et d'autres méta-analyses

Auteur de l'étude	Elève des groupes de niveau			Tous les élèves	Auteur de la méta-analyses
	Elevé	Moyen	Faible		
Drews (1963)	-0,16	0,01	-0,04	-0,05	Slavin (1990)
	-0,18	0,02	-0,08	-0,04	Mosteller (1996)
	0,17	0,08	0,14	n.c.	Steenbergen (2016)
Ford (1974)	n.c.		n.c.	n.s.	Slavin (1990)
	n.c.		n.c.	0,29	Mosteller (1996)
	0,5		0,45	n.c.	Steenbergen (2016)
Bicak (1964)	-0,39		-0,1	-0,25	Slavin (1990)
	-0,55		-0,16	-0,33	Mosteller (1996)
	0,31		0,12	n.c.	Steenbergen (2016)
	n.c.		n.c.	-0,028	EEF (2021)
Fick (1963)	0,01	0	-0,04	-0,01	Slavin (1990)
	0,25	0,009	-0,27	0,02	Mosteller (1996)
	n.c.	n.c.	n.c.	0,10	Steenbergen (2016)

n.c. = non calculable ; n.s. = non significatif (ES non calculable)

En 2021, Education Endowment Foundation (EEF) publie une méta-analyse sur le regroupement des élèves dans des classes de niveau (*Setting and Streaming*) [1]. Une première version avait été publiée en 2016 suivie d'une seconde version en 2018, qui n'est plus consultable actuellement mais qui reste souvent citées par les chercheurs. Ces deux premières versions sont des méta-analyses secondaires. Pour les raisons évoquées plus haut, la réalisation de méta-analyse secondaire a depuis été abandonnée par EEF, et la troisième version publiée en 2021 est cette fois une méta-analyse d'études primaires (réalisée à partir des études sélectionnées par leurs méta-analyses précédentes). Pour l'enseignement secondaire, le Toolkit a sélectionné treize études, dont dix sont également incluses dans les méta-analyses précédemment citées. Parmi les trois autres, deux études ont été exclues par Slavin (1990) : l'étude de Zweibelson (1965) car l'intervention consistait à regrouper les élèves par niveau deux jours par semaine (les trois autres jours les élèves étaient en classes hétérogènes) et celle d'Adamson (1972) car les différences initiales entre les élèves des groupes intervention et contrôle étaient trop importantes. La troisième est un ECR réalisé par EEF (Roy, 2018a) trop récent pour avoir été cité par les autres méta-analyses. Cet ECR a évalué les effets d'une intervention visant à améliorer les pratiques autour des regroupements d'élèves en classes de niveau hétérogène et n'avait pas comme objectif de comparer les deux types de regroupements (voir partie suivante). Il est donc étonnant qu'il ait été sélectionné dans cette méta-analyse. Les conclusions ne distinguent pas le regroupement par niveau uniquement dans certaines disciplines (*setting*) du regroupement par niveau dans toutes les disciplines (*streaming*) et indique que l'impact est nul en moyenne, avec des résultats moins bons pour les élèves peu performants avec un niveau de preuve limité.

Certaines limites aux conclusions de ces 8 méta-analyses doivent être posées qui concernent la sélection des études mais aussi les méthodes statistiques utilisées. Tout d'abord, ces méta-analyses regroupent toutes des études analysant plusieurs disciplines différentes. Les mathématiques en font

[1] https://educationendowmentfoundation.org.uk/education-evidence/teaching-learning-toolkit

presque toujours partie, accompagnées parfois de la littéracie, des sciences sociales et des sciences. À l'exception de Kulik (1982, 1992), de Steenbergen (2016) et d'EEF (2021), tous les types de regroupement sont inclus et indistinguables. De plus, à l'exception des méta-analyses de Kulik (1982) et de Slavin (1990), des niveaux d'étude différents (qui vont du primaire à l'enseignement supérieur) sont également inclus et indistinguables. Et finalement, parmi les 7 méta-analyses primaires, aucune n'a calculé une taille d'effet globale pour (1) des scores en mathématiques (2) d'élèves regroupés par classes de niveaux différents en mathématiques et en anglais uniquement et (3) scolarisés au secondaire inférieur.

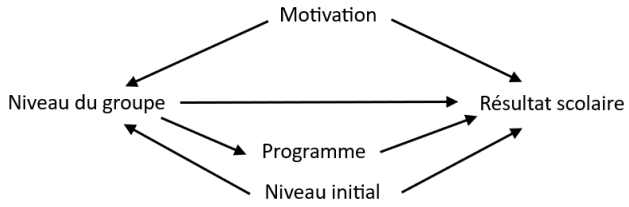
Si certaines études sélectionnées ont utilisé des scores issus de tests standardisés, pour d'autres ce sont des tests « maison » mis au point par les chercheurs qui ont fourni les données (ce sont souvent des études portant sur de faibles échantillons). Or on sait que les scores de tests standardisés donnent des tailles d'effet plus faible que les autres (Zhang, 2013 ; Wolf, 2023) : il conviendrait de prendre ces différences en compte, en donnant par exemple moins de crédit aux petites études, ce que ne font pas toujours les méta-analyses. Il est prudent donc de considérer les tailles d'effet comme potentiellement surestimées. Enfin, aucune d'entre elle n'a tenu compte du regroupement des élèves dans les classes (*clusters*) pour le calcul des tailles d'effet de chaque étude, ni du regroupement des tailles d'effet pour le calcul des tailles d'effet globales (les tailles d'effet provenant d'une même étude ne sont pas indépendantes) ce qui a comme effet de surestimer la signification statistique des tailles d'effets calculées (Hodgen, 2022). Des méthodes existent de nos jours qui permettent de corriger ces estimations (WWC, 2022). Il est vrai que de toutes façons, bien peu de tailles d'effet calculées ici sont statistiquement significatives et cela n'aurait donc probablement pas changer les conclusions.

3. Conclure sur la recherche internationale

Presque toutes les études analysées dans ce chapitre sont des études comparatives qui ont cherché à mesurer l'effet pour les élèves d'être regroupés ou non par niveau sur leurs scores académiques. Ces études sont en majorité observationnelles (les études expérimentales sur cette question ayant été abandonnées) et doivent surmonter des défis méthodologiques importants afin de démontrer le lien de cause à effet qu'elles interrogent. Il s'agit notamment de contrôler des biais de confusion, c'est-à-dire des facteurs qui risquent d'augmenter les effets que l'on pourrait à tort attribuer au regroupement par niveau. Mais aussi de tenir compte de variables médiatrices qui sont parties intégrantes du lien causal que l'on cherche à mettre en évidence. Le score initial est un biais de confusion évident (il est à l'origine du niveau du groupe auquel l'élève est affecté, et fortement corrélé à son score ultérieur). Et depuis la mise en garde de Slavin (1990), les chercheurs ont adapté le design de leurs études de manière à ce que les élèves des groupes de niveau élevé (ou moyen ou faible) soient comparés à des élèves de même niveau initial dans les classes hétérogènes. Mais d'autres biais de confusion existent, comme par exemple la motivation des élèves, qui n'ont presque jamais été contrôlés[1]. En ce qui concerne les variables médiatrices, on citera comme exemple les programmes suivis par les élèves évoqués ci-dessus. Pour quelques études (Kerckhoff, 1986 ; Argys , 1996 ; Figlio, 2002), il est à peu près certain que les élèves de niveau élevé ont suivi des programmes plus ambitieux que les autres ce qui a comme conséquence de surestimer les effets (voir figure X).

[1] Figlio (2002) est une exception à cette règle, mais peu convaincante.

Figure X. Variables de confusion ou médiatrices



Une autre limite des études par comparaison concerne la définition et la précision même de ce qui est comparé. Très peu de détails sont donnés sur les élèves regroupés par niveau. Par exemple on ne sait pas toujours combien de disciplines sont concernées par les regroupements, même si les mathématiques en font souvent partie. Les organisations sont complexes et il est difficile pour les études sur de grands échantillon de répartir les élèves dans des catégories cohérentes. De plus, cette répartition étant souvent le fuit d'une interprétation des enseignants, des erreurs peuvent s'être produites qui limitent la validité des conclusions. Le manque d'informations concerne également les élèves de classes dites « hétérogènes ». On a vu par exemple qu'ils peuvent parfois se trouver dans des classes de niveau informelles (Betts, 2000).

Ces limites obligent à une attitude conservatrice. Et au final, les conclusions de ce pan de la recherche montrent que le regroupement des élèves en classes de niveau homogène (comparé au regroupement des élèves en classes de niveau hétérogène) :

- n'a pas d'influence sur leurs résultats quand on considère l'ensemble des élèves,
- profite probablement un peu aux élèves des groupes de niveau élevé,
- défavorise probablement un peu les élèves des groupes de niveau faible.

On peut retenir que très peu de résultats plaident en faveur des regroupements par niveau pour les élèves les plus faibles[1]. Ce constat repose sur l'analyse d'études observationnelles de grande ampleur mais aussi sur des méta-analyses qui ont calculé des tailles d'effet la plupart du temps non significatives. Cela concerne aussi bien les tailles d'effet calculées pour chacune des études, les tailles d'effet globales qui comparent l'ensemble des élèves regroupés par niveau à l'ensemble des élèves non regroupés par niveau, et les tailles d'effet estimant l'effet pour chaque niveau de groupe (élevé, moyen ou faible). L'enquête PISA n'apporte pas de faits nouveaux.

La plupart de ces études ont été conduites sur le territoire américain et datent d'au moins 20 ans. La Royaume-Uni et les USA présentent des visages différents sur plusieurs plans (culturel, socioéconomique, scolaire) qui sont en lien direct avec notre question, ce qui limite la transférabilité des conclusions de ces recherches au territoire français. Il est intéressant alors de tourner notre regard vers l'Angleterre, plus proche de la Royaume-Uni, où plusieurs études d'envergures ont cherché à répondre plus récemment à notre question de recherche.

[1] Là encore, Figlio (2002) échappe à la règle.

Partie 3. Les études anglaises

Plusieurs recherches ont été conduites en Angleterre depuis la fin des années 1990 dont les résultats méritent d'être étudiés pour les raisons suivantes :

- les populations scolaires de la Royaume-Uni et de l'Angleterre sont comparables (par exemple en ce qui concerne le niveau socioéconomique et la proportion d'élèves issus de l'immigration) et les objectifs des politiques scolaires affichés sont similaires et peuvent se résumer par l'élévation du niveau des élèves et de l'équité du système scolaire ;
- les établissements anglais jouissent d'une grande autonomie et sont notamment libres de regrouper les élèves en classes de niveau plus ou moins homogène, ce qui devrait permettre de comparer plusieurs organisations différentes ;
- les autorités publiques anglaises ont récemment financé des études de grande ampleur sur le regroupement des élèves par niveau au secondaire dans leur pays.

100. Le contexte anglais

L'organisation du système scolaire incombe au Department for Education[1] (DfE) dirigé par le secrétaire à l'éducation. Les établissements scolaires sont publics ou privés[2]. Tous les établissements recevant des fonds publics sont soumis en Angleterre à l'évaluation par les inspecteurs de l'Ofsted (Office for Standards in Education) et chaque établissement reçoit un niveau Ofsted (Ofsted *grade*) après son inspection. Le système anglais désigne les niveaux d'étude en les numérotant à partir de la première année de scolarisation. Les élèves débutent ainsi en *year* 1 à 5 – 6 ans, puis passent en *year* 2 l'année suivante et ainsi de suite. Les intitulés anglais ont été conservés dans le texte. Le parcours scolaire est divisé en 4 cycles (voir tableau X). Les élèves anglais quittent l'école primaire à la fin de *year* 6 qui clôt le deuxième cycle de l'enseignement scolaire, le Key Stage 2 (KS2). Ils entrent alors au collège (*secondary school*) pour 5 années en passant de *year* 7 à *year* 11 (ce qui correspond pour le système français de la 6ème à la seconde). Les trois premières années (de *year* 7 à *year* 9) forment le Key Stage 3 (KS3). Les années qui suivent (*year* 10 et *year* 11) constituent le KS4. L'instruction scolaire est obligatoire jusqu'à 16 ans.

Tableau X : organisation de la scolarité en Royaume-Uni et en Angleterre

<table>
<tr><th rowspan="2">Âges</th><th colspan="3">Royaume-Uni</th><th colspan="3">Angleterre</th></tr>
<tr><td>Niveau</td><td>Cycles</td><td>Établissements</td><td>Niveau</td><td>Cycles</td><td>Établissements</td></tr>
<tr><td>6 – 7 ans</td><td>CP</td><td rowspan="3">Cycle 2</td><td rowspan="5">École primaire</td><td>year 2</td><td>Fin du KS 1</td><td rowspan="5">Primary school</td></tr>
<tr><td>7 – 8 ans</td><td>CE1</td><td>year 3</td><td rowspan="4">KS 2</td></tr>
<tr><td>8 – 9 ans</td><td>CE2</td><td>year 4</td></tr>
<tr><td>9 – 10 ans</td><td>CM1</td><td rowspan="2">Cycle 3</td><td>year 5</td></tr>
<tr><td>10 – 11 ans</td><td>CM2</td><td>year 6 (KS2 SAT*)</td></tr>
<tr><td>11 -12 ans</td><td>6ème</td><td rowspan="3">Cycle 4</td><td rowspan="3">Collège</td><td>year 7</td><td rowspan="2">KS 3</td><td rowspan="4">Secondary school</td></tr>
<tr><td>12 – 13 ans</td><td>5ème</td><td>year 8</td></tr>
<tr><td>13 – 14 ans</td><td>4ème</td><td>year 9 (KS2 SAT*)</td></tr>
<tr><td>14 – 15 ans</td><td>3ème (DNB**)</td><td>year 10 (GCSE***)</td><td rowspan="2">KS 4</td></tr>
<tr><td>15 – 16 ans</td><td>2nde</td><td>x</td><td>Début du lycée</td><td>year 11 (GCSE***)</td></tr>
</table>

* : Standard assessment test ; ** : Diplôme National du Brevet ; *** : General certificate of secondary school

[1] https://www.gov.uk/government/organisations/department-for-education

[2] 7 % des établissements du secondaire sont privés (Francis, 2019 a)

Les élèves issus de milieux socioéconomiques défavorisés sont éligibles au Free School Meal[1] (FSM). Ce critère est très souvent utilisé dans les études comme indicateur du niveau socioéconomique car il est très bien renseigné. En fin de *year* 6 (donc de KS2) et en fin de *year* 9 (donc de KS3), les élèves passent une évaluation en mathématiques et en anglais. Les épreuves sont corrigées par des professionnels extérieurs aux établissements et les scores sont standardisés. Il sera fait plus loin dans le texte référence au « score KS2 ou KS3 » des élèves pour plus de simplicité. Les résultats publiés en juillet servent notamment à classer les établissements et à constituer des groupes de niveau comme nous le verrons plus loin. Durant les deux années de KS4 les élèves passent une série d'examens dans le but d'acquérir le GCSE (General certificate of secondary education) qui détermine leur parcours ultérieur. Les épreuves diffèrent en fonction du niveau de l'élève, ce qui explique que certains programmes étudiés en *year* 10 et *year* 11 ne soient pas les mêmes d'une classe ou d'un groupe à l'autre.

Durant la première moitié du siècle dernier, l'idée de répartir les enfants en classes de niveau (*streaming*) était si naturelle en Angleterre qu'elle paraissait banale. Presque tous les établissements secondaires ont utilisé cette organisation scolaire et beaucoup d'établissements primaires de grande taille ont fait de même. À partir des années 1960, l'intérêt porté à une éducation centrée sur l'enfant a conduit un grand nombre d'établissements primaires à abandonner le *streaming*. Dans les établissements secondaires, bien que le *streaming* ait pu réduire l'éventail des résultats au sein d'une classe, ce dernier restait encore très large et le regroupement en classes de niveau dans certaines disciplines (*setting*) se superposait au *streaming*, en particulier en mathématiques. Dans les années 1960 et 1970, un certain nombre d'études ont montré que le *streaming* n'améliorait pas les résultats des élèves et cette organisation scolaire fut progressivement abandonnée dans les établissements secondaires (Wiliam, 2004). Des classes de niveau homogène en *year* 7 et 8 ont alors été remplacées par des regroupements en classes hétérogènes (*mixed ability grouping*). Mais à partir des années 1980, les critiques contre le 'progressisme' ont conduit à un retour en faveur des classes de niveau homogène (Taylor, 2022). Plus récemment, le Department for Education (DfE, 2005 a) a encouragé les établissements à regrouper les élèves par niveau. Quelques années plus tard, l'Ofsted (2013) met l'accent sur les élèves de plus haut niveau en s'inquiétant de leur sort dans les classes hétérogènes notamment au cycle KS3. Les établissements pour lesquels un retour vers des classes de niveau homogène semble avoir un impact positif sur les élèves les plus performants, sont cités en exemple. En 2014, Nicky Morgan[2] alors secrétaire à l'éducation, annonce vouloir imposer aux établissements de regrouper les élèves dans des classes de niveau homogène. Aucune suite ne sera donnée à cette proposition. On terminera ici cette revue des positions officielles en rappelant que EEF se montre assez critique sur les regroupements par niveau dans son Toolkit (EEF, 2021). Très peu d'informations sont données par les autorités publiques sur les modes de regroupements par niveau. On sait seulement qu'en 2003-2004, 53% des établissements regroupaient les élèves par niveau en mathématiques en *year* 7 et ils étaient 100% à le faire en *year* 10 selon les inspecteurs de l'Ofsted (Taylor , 2022).

2. La recherche sur les groupes de niveau

Les chercheurs anglais de la fin des années 1990 connaissaient bien évidemment les conclusions de la recherche internationale dont nous venons d'exposer les principaux résultats. Ils savaient donc que le regroupement des élèves dans des classes de niveau n'avait montré aucune preuve de son efficacité au

[1] Ils bénéficient de repas gratuit dans l'établissement
[2] https://www.theguardian.com/politics/2014/sep/03/schools-separate-pupils-ability-setting

niveau des résultats des élèves. Six études longitudinales ont été conduites entre 1995 et 2018 sur la question qui nous intéresse. Elles ont donné lieu à un grand nombre d'analyses et de publications. Leurs résultats concernent principalement 4 thèmes qui constituent l'ossature des chapitres suivants. Les différents modes de regroupement des élèves anglais tels qu'ils sont mis en œuvre dans les établissements secondaires seront présentés dans un premier temps. Les trois thèmes suivants correspondent à 3 questions de recherche puisqu'il s'agira d'analyser les liens entre le regroupement des élèves par niveau et

- leurs résultats scolaires ;
- leurs résultats non cognitifs (leur estime de soi, leurs souhaits , ...) ;
- les pratiques à l'intérieur des établissements et des classes.

Les résultats concernant l'enseignement des mathématiques nous intéressent plus particulièrement ; ils seront parfois comparés avec des données concernant l'enseignement de l'anglais.

Les principales caractéristiques des six études anglaises (repérées par les lettres A à F dans la suite du texte pour plus de simplicité) sont présentées ci-dessous dans un ordre chronologique. La première (**étude A**) est une étude longitudinale de 4 années (1996 à 2000) qui analyse les effets du regroupement des élèves en classes de niveau homogène sur les résultats académiques et non cognitifs d'élèves scolarisés dans six établissements du grand Londres de *year* 7 à *year* 11. L'**étude B** est également une étude longitudinale (1998 – 2000) et concerne les élèves de *year* 7 à *year* 9 scolarisés dans 45 établissements. Elle a donné lieu à de nombreuses analyses et publications, qui détaillent les modes de regroupements des élèves, analysent les effets de ces regroupements sur les résultats académiques et non cognitifs ainsi que sur les méthodes d'enseignements des enseignants. Une étude transversale conduite sous l'égide du DfE (**étude C**) a analysé les données d'une enquête à laquelle 44 établissements secondaires ont répondu entre 2006 et 2007. Son objectif principal était de décrire les modes de regroupements des élèves de *year* 7 et les méthodes pédagogiques déployées en s'intéressant plus particulièrement aux élèves de niveau faible. Une étude de cas portant sur cinq établissements complète ces analyses. Deux essai contrôlé randomisé par cluster ont été parallèlement mis en œuvre par EEF entre 2015 et 2017 et concerne les élèves de *year* 7 et *year* 8. Le premier est une étude pilote (**étude D**) dont l'objectif était d'analyser l'intervention Best Practice in Mixed Attainment. Dix-huit établissements (12 dans le groupe intervention, 6 dans le groupe contrôle) y ont participé. Le second (**étude E**) est une étude d'efficacité (*efficacy trial*) réalisé dans le but d'évaluer l'intervention Best Practice in Setting dont l'objectif était d'améliorer les pratiques de 61 établissements regroupant leurs élèves par niveau (un groupe contrôle était constitué de 60 établissements regroupant leurs élèves par niveau mais ne recevant pas l'intervention). Les résultats de cette étude ont également permis de décrire les modes de regroupements, et de comparer l'effet des regroupements dans des classes de niveaux homogènes différents sur les résultats académiques et non cognitifs des élèves. Enfin, l'**étude F** (financée par EEF) a analysé les données récoltées auprès de 197 établissements ayant répondu à un questionnaire envoyé à tous les établissements d'Angleterre en 2018 dans le but de décrire les modes de regroupements mis en place dans ce pays. On terminera en signalant qu'une étude EEF est actuellement en cours dont certains éléments publiés dans un plan d'étude (Hodgen, 2019) seront évoqués dans ce chapitre.

2.1. Comment sont regroupés les élèves ?

Il va s'agir de répondre à trois questions :

1. Quels modes de regroupements des élèves (regroupement des élèves dans des classes hétérogènes ou homogènes) sont retenus par les établissements anglais pour l'enseignement des mathématiques ?
2. Quelles sont les règles appliquées pour répartir les élèves dans ces différents groupes en mathématiques ?
3. En dehors de leur niveau scolaire, quelles caractéristiques des élèves sont associées au niveau des groupes dans lesquels ils sont affectés ?

Prévalence des différents modes de regroupements

Les résultats de l'enquête PISA 2022 permettent d'affirmer que les élèves de 15 ans britanniques (qui sont presque tous en *year* 11 car très peu ont redoublé) sont regroupés dans des classes de niveau. Les informations rassemblées dans le tableau X sont issues des questionnaires auxquels ont répondu les chefs d'établissements participant à l'enquête. Aucune précision n'est donnée, ni sur les pratiques concernées par les regroupements par niveau « dans différentes classes » ou « dans leur classe », ni les disciplines concernées par la catégorie « pour quelques disciplines ». Mais d'après ce que nous verrons plus loin, il est très probable que ces regroupements prennent la forme du *setting* et non du *streaming* et concernent les mathématiques.

Tableau X : regroupement (%) des élèves en Royaume-Uni et au Royaume-Uni (tableau II.B1.4.26)

Les élèves sont	regroupés par niveau dans différentes classes			regroupés par niveau dans leur classe		
	pour toutes les disciplines	pour quelques disciplines	pour aucune discipline	pour toutes les disciplines	pour quelques disciplines	pour aucune discipline
Royaume-Uni	5,0	92,4	2,6	2,3	50,9	46,8
Moyenne OCDE	**6,7**	**30,8**	**62,5**	**5,7**	**42,4**	**51,9**

En 2018, les 3 105 établissements secondaires anglais financés par l'État, non sélectifs et scolarisant des élèves de *year* 7 à *year* 11 ont été destinataires d'un questionnaire les interrogeant sur le regroupement des élèves par niveau (**Etude F**, Taylor, 2022). L'objectif était de décrire précisément les organisations retenues. Pour les mathématiques, les informations proviennent des réponses de 197 établissements, ce qui représente un taux de réponse de 6,3 %. Les résultats montrent que le regroupement des élèves en classes de niveau homogène est largement dominant en mathématiques et que la fréquence de ces regroupements augmente au fil du temps : seuls 8% des élèves en *year* 7 sont en classes hétérogènes, ils sont moins de 2 % à partir de *year* 9. L'auteur souligne que contrairement à une idée répandue, le *streaming* qui concerne 10% des élèves en *year* 7 n'a pas disparu. En anglais les regroupements par niveau sont moins fréquents, puisque par exemple 31% des élèves sont scolarisés dans des classes hétérogènes en *year* 7, et cette proportion est encore de 11 % en *year* 11. Les établissements de l'**étude B** ont été choisis de manière à ce que plusieurs types de regroupements soient représentés, ils ne sont donc pas représentatifs des établissements anglais, tout du moins en ce qui concerne cette caractéristique (Ireson, 2002a et 2002b). Sur les 45 établissements de l'échantillon, seuls 5 ne regroupent pas leurs élèves en mathématiques en *year* 9. Ici aussi les regroupements par niveau sont moins fréquents en anglais. La grande difficulté rencontrée par les auteurs de l'**étude D (**détaillée plus loin) à recruter des établissements regroupant les élèves en classes hétérogènes en mathématiques en *year* 7 montre encore que bien peu d'élèves ne sont pas regroupés par niveau dans cette discipline dès le début du secondaire.

Pour terminer cette analyse descriptive, on remarquera que les établissements regroupent les élèves souvent sur plus de 3 niveaux de compétence. Ainsi parmi les 46 établissements du groupe contrôle de l'**étude E**, 41 % regroupent leurs élèves sur 4 niveaux de compétence, 24 % sur 3 niveaux et 17 % sur 5

niveaux (Connolly, 2019). Plus l'établissement est grand, plus le nombre de niveaux est susceptible d'être élevé (certains établissements ont jusqu'à 10 niveaux). Pour intégrer dans une même analyse des établissements qui organisent des regroupements avec des nombres de niveaux différents, les chercheurs constituent toujours 3 catégories (la catégorie de niveau faible, la catégorie de niveau moyen et la catégorie de niveau élevé), et ce parfois de façon différente comme nous le verrons par la suite. Les groupes de niveau faible ont de plus petits effectifs que les groupes de niveau élevé (en général deux fois plus faibles que les effectifs des autres groupes). C'est par exemple le cas de 44 établissements sur 45 de l'**étude B**[1].

Comment sont constitués les groupes ?

On s'intéresse ici aux caractéristiques explicitement liées au regroupement par niveau et qui sont associées au niveau scolaire des élèves. Les autres caractéristiques des élèves également corrélées à leur regroupement, mais non utilisées intentionnellement pour répartir les élèves dans les groupes sont analysées plus loin. La question porte sur la constitution initiale des groupes (en début d'année scolaire) mais également sur les éventuelles modifications qui peuvent y être apportées au cours de l'année. Les études sont quantitatives et qualitatives ; elles ont exploité les données issues de questionnaires mais aussi d'entretiens. Les informations issues des déclarations des personnels sont présentées en premier lieu, suivi par une présentation des données quantitatives non déclaratives.

<u>Déclaration des personnels</u>

Pour être inclus dans l'**étude E**, les établissements devaient regrouper par niveau en mathématiques sans pratiquer de *streaming*. 37% des enseignants de mathématiques ont répondu à un questionnaire leur demandant de préciser les méthodes utilisées pour répartir les élèves de *year* 7 dans les classes de niveau (Taylor, 2019). 74 % des enseignants utilisent les scores KS2 (d'autres tests sont également cités, comme des tests propres à l'établissement pour 53% d'entre eux) et 31 % se basent aussi sur leur jugement. Ces conclusions sont confirmées par des entretiens avec les enseignants qui ont déclaré ne pas faire toujours confiance aux scores KS2 ni aux évaluations des professeurs des écoles primaires. Dans certains établissements les élèves sont regroupés selon leur niveau général, méthode qui s'apparente à du *streaming*, et des regroupements par niveau dans certaines disciplines peuvent alors se surajouter. Les pratiques de regroupements paraissent parfois peu claires et les enseignants se contredisent, notamment dans les établissements présentant une faible stabilité des équipes. Les répartitions des élèves sont également contraintes par des facteurs pratiques comme les emplois du temps ou des considérations financières qui impactent les effectifs des groupes. Dans les 44 établissements de l'**étude C**, 89 % des enseignants disent prendre leur décision sur la base des scores de l'élève (l'étude ne précise pas lesquels), 11 % sur une évaluation de leur capacité générale et 2,3 % pour des motifs liés au comportement (Mujis, 2010). Les responsables en mathématiques qui ont répondu à une enquête dans le cadre de l'**étude B** disent utiliser plus souvent des tests propres à l'établissement que les scores KS2 ou KS3 pour répartir les élèves par niveau (Ireson, 2002b). Les regroupements sont également fondés sur l'opinion des enseignants qui pour certains font référence à l'attitude ou au comportement des élèves quand d'autres au contraire affirment porter une attention particulière à ne pas justifier la répartition des élèves pour des motifs de ce type. Cette grande différence des pratiques d'un établissement à l'autre est également retrouvée par les auteurs de l'**étude A** : sur les six établissements de l'échantillon, les règles régissant la répartition des élèves dans les

[1] Betts (2000) a montré que les effectifs moyens des classes homogènes de niveau faible sont inférieurs à 19 ; ce nombre est supérieur à 23 pour les autres classes.

différents groupes n'étaient pas transparentes dans quatre d'entre eux, et les élèves (voire dans certains cas leurs enseignants) ne savaient pas ce qui différenciait leur groupe d'un autre (William, 2004). Cette mauvaise connaissance des enseignants concernant leur propre établissement alerte les auteurs sur les précautions à prendre quant à l'interprétation des résultats d'enquêtes basés sur des rapports d'enseignants ou d'élèves.

De grandes variations existent quant aux règles qui permettent (ou non) aux élèves de changer de groupe de niveau. Certains enseignants interrogés dans le cadre de l'**étude E** mentionnent l'attitude au travail comme un élément à prendre en considération (Taylor, 2019). Ils se montrent favorables à transférer un élève dans un groupe de niveau supérieur, mais peu désireux de faire l'opération inverse. Également, l'instauration d'une bonne relation entre l'enseignant et l'élève est un élément qui peut intervenir dans les décisions. Les auteurs remarquent que les valeurs et convictions des enseignant influent fortement sur la mobilité des élèves. Les résultats de l'**étude B** (Ireson, 2002 b) vont dans le même sens et montrent que les organisations peuvent grandement différer d'un établissement à l'autre, ce sujet pouvant (ou non) faire l'objet d'une attention particulière ; aucun chiffrage n'a été possible car les établissements ne gardent pas trace de ces mouvements. Les transferts sont plus nombreux en *year* 7 et *year* 8 qu'au-delà, et se font généralement après des évaluations (et donc leur fréquence dépend de la fréquence de ces dernières). Certains enseignants ont mentionné avoir été sollicités par des parents pour que leur enfant change pour un groupe de niveau plus élevé. Le nombre d'élèves dans ces groupes de niveau élevé constitue souvent un plafond qui restreint les mouvements. La question du comportement revient encore (parfois pour souligner qu'il faut éviter de transférer un élève dans un groupe de niveau faible pour ce type de motif).

Ce que disent les chiffres

Les données de ce chapitre ne concernent que les mathématiques. Tous les élèves de *year* 7 participant à l'**étude C** sont regroupés par niveau (Muijs, 2010, Dunne, 2011). Certains établissements utilisant plus de trois niveaux de regroupements, les auteurs ont donc demandé aux enseignants de décrire leurs répartitions en affectant chacun des élèves dans une des trois catégories : niveau faible, moyen ou élevé. Les scores KS2 de l'échantillon total ont été classés en terciles, et pour chacun d'entre eux les proportions des élèves de chacune des trois catégories de niveau ont été calculées (tableau X) : un peu moins de la moitié des élèves ayant obtenu un score KS2 faible sont scolarisés dans un groupe de niveau faible, et seul un peu plus de la moitié des élèves ayant obtenu un score KS2 élevé se trouvent dans un groupe de niveau élevé.

Tableau X : proportion (%) des élèves de *year* 7 dans les catégories en fonction des scores KS2 en maths

Catégorie	de niveau faible	de niveau moyen	de niveau élevé
Scores KS2 faibles	46,6	38,9	14,5
Scores KS2 moyen	21,8	52,6	25,6
Scores KS2 élevés	11,8	33,2	54,9

Les élèves de *year* 7 de l'**étude E** sont tous regroupés par niveau et ont été répartis par les chercheurs dans 3 catégories : les élèves appartenant au groupe de niveau le plus faible niveau ; les élèves appartenant au groupe de niveau le plus élevé ; les élèves appartenant aux autres groupes (Connolly, 2019). L'auteur analyse ce qu'il définit comme une mauvaise allocation des élèves en comparant l'allocation des élèves dans l'une des trois catégories de niveau à ce qu'elle aurait été si elle avait été basée uniquement sur les scores KS2. Dans un premier temps, deux types d'élèves sont définis : les

élèves *borderline* (leur score KS2 aurait pu légitimement les placer dans un autre groupe de niveau que celui dans lequel ils ont été placé) et les autres. Parmi les élèves de cette dernière catégorie, 69 % sont placés dans le bon groupe, 16 % sont mal placés dans un groupe de niveau trop élevé et 16 % sont mal placés dans un groupe de niveau trop faible.

Les données de l'**étude B** ont été analysées en définissant là aussi trois catégories de groupes : la catégorie de niveau élevé avec 25% des élèves des groupes de niveau élevé, la catégorie de niveau faible avec 25% des élèves des groupes de niveau faible, et les autres élèves dans la catégorie de niveau moyen (Ireson, 2005a). La répartition des élèves de *year* 9 dans ces catégories en fonction de leur niveau de scores KS3 (ces niveaux allant de 2 à 8) montre que du niveau 4 au niveau 7, des élèves ayant des scores similaires pouvaient se trouver dans l'une de ces trois catégories.

Corrélations avec d'autres caractéristiques

Dans l'**étude E** (Connolly, 2019), le genre, le niveau socioéconomique et l'origine ethnique des élèves ont été analysés pour chacune des trois catégories définies plus haut (tableau X). Une attrition de 36 % (non discutée par les auteurs) est observée pour les données caractérisant l'origine ethnique, et il a été décidé de ne pas présenter ces résultats ici. Aucun test statistique n'a été conduit par l'auteur sur la significativité qu'il convient d'associer aux différences observées. L'auteur analyse la différence entre les deux taux d'élèves mal placés (vers un groupe de niveau trop élevé ou vers un groupe de niveau trop faible), et remarque notamment que les garçons sont plus souvent mal placés dans un groupe de niveau trop élevé que les filles (tableau X).

Tableau X : répartition (%) dans les groupes de niveau en mathématiques

	Groupe de niveau le plus élevé (%)	Groupes de niveau moyen (%)	Groupe de niveau le plus faible (%)
Tous	29,6	57,0	13,4
Garçons	32,1	55,8	12,1
Filles	27,1	58,3	14,6
Éligible FSM	23,5	59,9	16,5
Non éligible FSM	33,1	54,9	12,0

Lecture : 32,1 % des garçons sont dans le groupe de niveau le plus élevé

Tableau X : proportion d'élèves (%) correctement ou mal placés

	Placé dans un groupe de niveau trop élevé	Placé dans le bon groupe	Placé dans un groupe de niveau trop faible
Tous	15,7	68,9	15,5
Garçons	16,7	70,3	13,0
Filles	14,7	67,4	17,9
Éligible FSM	18,8	62,5	18,7
Non éligible FSM	14,4	71,7	13,9

Lecture : 14,7% des filles sont placées dans un groupe de niveau trop élevé.

Certains facteurs indépendants pouvant être corrélés, des modèles statistiques basés sur des régressions linéaires multiples multiniveaux ont été utilisés pour en différencier les effets. Ils ont montré que les élèves qui sont éligibles au FSM sont plus souvent mal placés dans un groupe de niveau trop faible que les autres, que les garçons sont plus souvent mal placés dans un groupe de niveau trop élevé que les filles, et moins souvent mal placés dans un groupe de niveau trop faible que les filles. Comme nous l'avons signalé, de nombreuses données étaient manquantes ; de plus l'indicateur être (ou non) éligible au FSM est reconnu par l'auteur comme un indicateur pauvre du niveau socioéconomique des élèves ; d'autres modèles statistiques ont alors été mis en œuvre (analyse de

sensibilité en ce qui concerne les données manquantes et utilisation d'un autre indicateur socioéconomique pour la seconde remarque). Les conclusions restent les mêmes concernant le genre des élèves, mais le lien entre le niveau socioéconomique et le mauvais placement des élèves disparait. Enfin, l'affectation des élèves *borderline* définis plus haut n'est pas influencé par leur genre, leur niveau socioéconomique ni leur origine ethnique.

Les auteurs de **l'étude C** (Muijs, 2010) ont analysé l'appartenance des élèves à l'une de trois catégories (niveau faible, niveau moyen ou niveau élevé), en fonction de leur niveau socioéconomique (élèves éligibles au FSM ou non et niveau d'étude et d'emploi des parents), de leur genre, de leur origine ethnique et de la présence ou non de besoins éducatifs particuliers (*Special Educational Needs*, *SEN*). Ces résultats concernent les mathématiques et l'anglais ; ils doivent être considérés avec précaution, car aucun test statistique n'est conduit sur les différences observées et le mode de calcul des proportions lui-même manque de clarté. Là aussi on constate une surreprésentation des élèves non éligibles au FSM dans les groupes de niveau élevé en mathématiques et anglais. On rajoutera à cette liste les élèves sans besoin éducatif particulier. Une régression logistique multinomiale a été conduite en classant les scores KS2 en quintiles. Cette variable indépendante est le prédicteur le plus fort : les élèves dans les deux catégories des scores les plus faibles ont une probabilité supérieure d'être dans les groupes de niveau faible que les élèves qui sont dans les trois catégories de scores les plus élevés. Le genre et l'origine ethnique ne sont pas des prédicteurs statistiquement significatifs. Mais, à score initial égal, le niveau socioéconomique faible, l'éligibilité au FSM et avoir des besoins particuliers sont des prédicteurs statistiquement significatifs de l'appartenance des élèves à un groupe de niveau faible. Dans l'**étude A**, un modèle linéaire a permis d'analyser le niveau du groupe en fonction des scores au test KS3, du niveau socioéconomique et du genre de l'élève. Les résultats montrent que, à score KS3 équivalent, les élèves de la classe ouvrière sont placés dans des groupes de niveau plus faible que les élèves de la classe moyenne (William, 2004).
Finalement, la quasi-totalité des élèves anglais scolarisés dans le secondaire inférieur suivent leurs cours de mathématiques dans des classes de niveau (ce n'est pas le cas en anglais). On a vu que malgré ce qui est dit, les scores initiaux (notamment les scores KS2 plébiscités par les chercheurs) ne sont pas les seuls critères de répartition des élèves et que certaines caractéristiques individuelles sociales jouent un rôle non négligeable : les élèves de milieux socio-économiques défavorisés et les élèves ayant des besoins particuliers sont surreprésentés dans les groupes de niveaux faibles, y compris quand leur niveau initial est pris en compte. Les organisations sont très diverses et cette diversité concerne les méthodes retenues pour répartir les élèves dans les groupes, le nombre de niveaux, le nombre de disciplines regroupant par niveau, les transferts éventuels, etc. Enfin près d'un élève sur trois serait placé dans un groupe ne correspondant pas à son niveau (tel qu'évalué par un test national standardisé) au début du collège. Si les résultats concernant l'origine ethnique n'ont pas été présentés, on gardera néanmoins à l'esprit que les élèves issus de minorités ethniques étant également souvent de milieux socioéconomiques défavorisés, ils sont inévitablement surreprésentés dans les groupes de niveau faible.

2.2. Groupes de niveau et résultats académiques

Mesurer l'effet du type de regroupement sur les résultats scolaires des élèves suppose qu'il soit possible d'isoler les facteurs « être scolarisé (ou non) dans une classe de niveau en mathématiques » ou « être scolarisé dans une classe de niveau élevé (ou moyen ou faible) » de tous les autres facteurs qui ont un effet sur les compétences et résultats académiques des élèves. Et les facteurs de confusions,

qui influent à la fois sur la variable dépendante (le score final de l'élève) et la variable indépendante (le fait d'être dans une classe homogène ou non ; ou bien le fait d'être dans une classe de niveau élevé ou moyen ou faible) doivent être contrôlés. Le niveau initial des élèves en fait bien évidemment partie. Comme nous venons de le voir, presque tous les élèves anglais sont regroupés en classes de niveau en mathématiques à leur rentrée au secondaire, et comparer des élèves de niveau équivalent dans les deux situations qui nous intéressent (être ou on dans une classe de niveau) s'est avéré difficile. Les chercheurs ont donc surtout analysé les scores d'élèves regroupés dans des classes de niveaux différents en utilisant des modèles de régressions linéaires qui tous incluent comme variable indépendante un score initial (ou score prétest).

Les données de **l'étude E** ont été analysées pour évaluer l'impact que le niveau du groupe dans lequel l'élève est affecté peut avoir sur ses résultats (Hodgen, 2023). Des régressions linéaires ont modélisé les variations des scores obtenus en fin de *year* 8 en fonction du score KS2 (score prétest), de la catégorie de niveau, de l'origine ethnique, du niveau socioéconomique et du genre de l'élève ; du nombre de niveaux de groupes (qui peut aller jusqu'à 10) et de la participation (ou non) à l'intervention *Best Practice in Setting* de l'établissement. Quatre modèles statistiques multiniveaux ont été utilisés pour tenir compte des données manquantes (qui concernent ici aussi l'origine ethnique et le niveau socioéconomique des élèves). Pour chacun des modèles statistiques des tailles d'effet ont été estimées qui comparent les élèves de la catégorie de niveau élevé aux élèves de la catégorie de niveau moyen d'une part, et les élèves de la catégorie de niveau faible aux élèves de la catégorie de niveau moyen d'autre part. Deux groupes d'analyses sont conduites pour distinguer les mathématiques de l'anglais. Les résultats montrent que le score prétest a un impact important, tout particulièrement en mathématiques. Être affecté dans un groupe de la catégorie de niveau élevé a un faible effet positif dans tous les modèles en mathématiques et un effet positif plus marqué en anglais ; toutes les tailles d'effet sont statistiquement significatives (tableau X). Cela signifie donc que, à niveau initial équivalent, les élèves de la catégorie de niveau élevé réussissent mieux que les élèves de la catégorie de niveau moyen. Être affecté dans un groupe de la catégorie de niveau faible a un effet négatif dans tous les modèles, mais parmi les 8 tailles d'effet calculées seuls les résultats estimés à partir de deux modèles pour les scores en anglais sont statistiquement significatifs.

Tableau X. Étendue des tailles d'effet calculées à partir des 4 modèles

	Catégorie de niveau élevé	Catégorie de niveau faible
Mathématiques	de 0,094 à 0,124	de –0,003 à –0,046
Anglais	de 0,273 à 0,288	de –0,089 à –0,252

Comme le souligne l'auteur, l'attrition importante durant les deux années de l'étude et l'absence de groupe contrôle limitent la validité interne des résultats de cette étude. On ne peut donc être certain que les scores des élèves des groupes de niveau le plus élevé auraient été différents s'ils avaient été scolarisés dans des classes hétérogènes. Enfin la qualité des enseignements et les éléments du programme enseigné n'ont pas été pris en compte. C'est d'ailleurs pour pallier à ces défauts qu'une étude comparative devait être réalisée par EEF en 2020 – 2021. La pandémie du Covid en a empêché la réalisation selon le planning prévu[1]. Ses objectifs publiés dans son plan d'étude (Hodgen, 2019) sont de déterminer quelle organisation (regroupement homogène ou hétérogène) produit l'impact le plus

[1] Les résultats devraient être publiés en 2025 (communication personnelle)

favorable sur les résultats en mathématiques et sur l'estime de soi de l'ensemble des élèves, mais aussi plus particulièrement des élèves défavorisés, en conduisant une étude longitudinale observationnelle.

Les auteurs de l'**étude B** ont analysé les scores d'élèves en mathématiques en fonction de leur expérience passée de *year* 7 à *year* 11 dans des classes plus ou moins homogènes ou hétérogènes. L'influence du type de regroupement a été analysée sur les scores KS3 des élèves, donc 3 ans après leur arrivée dans le secondaire (Ireson, 2002 a) mais aussi sur les résultats au GCSE des élèves, donc 5 ans après leur arrivée dans le secondaire (Ireson, 2005a). En ce qui concerne la première analyse, un modèle statistique de régression linéaire à deux niveaux (niveau élève et niveau établissement) a été utilisé pour analyser les scores KS3 en mathématiques en fonction du score KS2, du genre, de l'éligibilité au FSM, de l'assiduité et de l'intensité du regroupement homogène[1]. En mathématiques (mais pas en anglais ni en sciences) cette intensité influe positivement sur les scores KS3 des élèves ; l'analyse de l'interaction entre les scores KS2 et l'intensité du regroupement homogène montre que les élèves de niveau faible au KS2 progresse plus dans les classes hétérogènes et que les élèves de niveau élevé au KS2 progressent plus dans les classes homogènes. L'impact du niveau du groupe sur le score des élèves indique que pour trois élèves ayant le même score KS2, la performance de l'élève placé dans un groupe de niveau élevé sera significativement plus élevée et la performance de celui placé dans un groupe de niveau faible sera significativement plus faible que la performance de l'élève placé dans un groupe de niveau moyen. La seconde analyse (Ireson, 2005a) sort de notre cadre de recherche, l'auteur soulignant que les élèves en *year* 10 et *year* 11 pouvant avoir suivi des programmes différents en fonction de leur préparation au GCSE. Les résultats sont tout de même intéressants, puisqu'ils montrent notamment que l'intensité du regroupement homogène n'a aucun effet significatif sur les scores GCSE (en mathématiques mais aussi pour les autres disciplines).

Finalement, si on analyse globalement l'ensemble des élèves, avoir été regroupé par niveau ou non n'a pas d'effet sur les scores. C'est le niveau, de l'élève ou du groupe auquel il est affecté, qui compte. À niveau initial équivalent, être scolarisé dans un groupe de niveau élevé a un effet positif sur les résultats académiques en mathématiques au bout de deux ou trois années passées dans un établissement secondaire. Une étude (Ireson, 2002a) montre que les scores KS3 sont meilleurs, pour les élèves de niveau faible dans les classes hétérogènes d'une part et pour les élèves de niveau élevé dans les classes homogènes d'autre part.

2.3. Groupes de niveau et résultats non cognitifs

Le bien être à l'école impacte le parcours des élèves, et des caractéristiques associées à ce bien-être ont également été étudiées, toujours en relation avec les classes de niveau. Sous l'intitulé « résultats non cognitifs » sont regroupés des éléments disparates comme l'estime de soi des élèves[2] (dans plusieurs domaines), leur niveau de satisfaction (associée à une discipline ou à une organisation de l'enseignement) ou leurs souhaits. Dans ces analyses quantitatives, les variables dépendantes sont issues des réponses d'élèves à des questionnaires, et elles ont été étudiées en comparant des sous-groupes différents ou en estimant des coefficients de régression dans des modèles linéaires.

[1] Cet indicateur varie pour chaque établissement entre 0 à 5 (0 pour un établissement ne regroupant pour aucun *year*, 5 pour le cas contraire).

[2] Ce terme a été utilisé ici d'une façon générique, les auteurs des articles faisant parfois référence aux concepts de *self-esteem*, *self-concept*, *self-confidence* ou *self-perception* en reconnaissant que leur distinction ne repose pas sur des règles très claires (voir Francis, 2017 b).

Les données de **l'étude E** ont été analysées pour évaluer l'impact que le niveau du groupe dans lequel l'élève est affecté peut avoir sur son estime de soi en début de *year* 7 (Francis, 2017b) puis en fin de *year* 8 (Francis, 2020). L'estime de soi est mesurée dans trois domaines différents : les mathématiques, l'anglais et les apprentissages d'une façon plus générale (avec ici deux analyses séparées, l'une évaluant l'effet du regroupement en mathématiques et l'autre l'effet du regroupement en anglais). En mathématiques, l'analyse statistique se base sur les données de 77 % de l'échantillon de départ (cette attrition n'est pas commentée dans l'article). Pour l'analyse menée en début de *year* 7 (Francis, 2017b), des régressions linéaires multiple à trois niveaux (niveau élève, niveau groupe, niveau établissement) incluent, en plus de la catégorie de niveau, les covariables permettant de tenir compte du genre, de l'origine ethnique, du statut socio-économique et du nombre de niveaux de groupes de l'établissement. Le score initial (qui aurait pu être un score KS2) n'est pas inclus dans le modèle. Des tailles d'effet sont publiées par les auteurs, qui comparent l'estime de soi des élèves de la catégorie de niveau élevé aux élèves de la catégorie de niveau faible. La taille d'effet est égale à 0,71 pour l'estime de soi concernant les mathématiques et de 0,57 pour l'estime de soi concernant les apprentissages et pour des élèves regroupés par niveau en mathématiques ; ces tailles d'effet sont statistiquement significatives. Le niveau scolaire de l'élève, très probablement lié à son estime de soi, est un facteur de confusion qui n'a pas été contrôlé et aucun effet causal n'est de ce fait démontré. Ces mesures de l'estime de soi obtenus en *year* 7 gardent tout de même un certain intérêt car elles sont introduites dans les modèles linéaires comme variables explicatives pour l'analyse de l'évolution de l'estime de soi en fin de *year* 8 cette fois. Dans un premier modèle les variables indépendantes sont l'estime de soi mesuré deux ans auparavant, le niveau du groupe dans lequel l'élève est affecté, le nombre de niveaux de groupes de l'établissement, le genre, le niveau socioéconomique et l'origine ethnique de l'élève. Dans un second modèle le score KS2 (score prétest) est ajouté à ces variables. Des tailles d'effet ont été calculées en comparant les résultats des élèves de la catégorie de niveau élevé aux résultats des élèves de la catégorie de niveau moyen d'une part ; et les résultats des élèves de la catégorie de niveau faible aux résultats des élèves de la catégorie de niveau moyen d'autre part. Si on s'en tient aux résultats obtenus à partir du second modèle statistique, pour trois élèves de même niveau académique en fin de primaire et partageant un même sentiment de confiance pour les apprentissages en général en début de *year* 7, au bout de deux ans l'élève placé dans un groupe de niveau élevé en mathématiques a plus confiance en lui que l'élève placé dans un groupe de niveau moyen (ES = 0,057), qui lui-même a plus confiance en lui que l'élève du groupe de niveau faible (ES = 0,056). Les tailles d'effet sont statistiquement significatives mais très faibles. Par contre l'effet sur l'estime de soi associée à l'apprentissage des mathématiques qui était significatif dans le premier modèle disparait quand on tient compte du niveau initial de l'élève.

Les données de l**'étude B** ont fait l'objet de plusieurs analyses concernant les élèves de *year* 9. Un premier groupe d'étude a analysé l'effet du regroupement des élèves sur leur estime de soi (Ireson, 2001) et sur leur niveau de satisfaction vis-à-vis de l'école et des enseignements (Ireson, 2005 b). Un second groupe a rassemblé des données sur leurs préférences pour les différents types de regroupement (classes hétérogènes, classes homogènes ou autre) (Hallam, 2006) et sur leur niveau de satisfaction concernant leur affectation dans tel ou tel groupe de niveau (Hallam, 2007). Trois types d'établissements ont été sélectionnées : les établissements ne regroupant presque pas les élèves par niveau, les établissements regroupant par niveau partiellement et les établissements regroupant de façon stricte. Les résultats de cinq modèles de régressions linéaires, s'intéressant chacun à une facette particulière de l'estime de soi, ont été exploités. Les résultats montrent que les élèves ont une meilleure

estime de soi associée aux apprentissages mais aussi non rattachée à l'école dans les établissements qui regroupent par niveau partiellement que dans les établissements qui regroupent de façon plus stricte ; et que les élèves ont une meilleure estime de soi associée aux apprentissages dans les établissements qui regroupent par niveau partiellement que dans les établissements qui ne regroupent presque pas les élèves par niveau. En mathématiques, l'intensité du regroupement[1] n'a aucun impact sur l'estime de soi associée à cette discipline (Ireson, 2001).

Les niveaux de satisfaction des élèves de *year* 9 vis-à-vis de l'école d'une façon générale, ainsi que sur leurs enseignements en mathématiques, anglais et sciences, ont été analysés d'après leurs réponses à un questionnaire (Ireson, 2005 b). Les élèves des établissements regroupant strictement par niveau apprécient significativement moins l'école que les élèves des autres types d'établissements ; cette remarque est également valable quand on restreint l'échantillon aux élèves de niveau faible (dernier quartile des scores KS3). Les élèves de la catégorie de niveau élevé apprécient significativement plus leur enseignement que les élèves de deux autres catégories, et ce dans les trois disciplines analysées. Enfin, l'analyse de ces données par une régression linéaire multiple à deux niveaux (le niveau élève et le niveau établissement) montre que si l'on tient compte de l'estime de soi associée à l'école mesurée en *year* 7, de la perception globale des enseignements, du score KS3, du genre et de l'éligibilité (ou non) au FSM des élèves, leur opinion sur l'école ne dépend pas du type de l'établissement.

Toujours dans l**'étude B**, les préférences des élèves pour les différents types de regroupement ont été exploitées pour 73 % des élèves (l'attrition de 27 % n'est pas commentée dans l'article). Les auteurs remarquent que les élèves ont tendance à préférer le mode de regroupement qu'ils ont expérimentés : 47 % des élèves des établissements ne regroupant presque pas les élèves par niveau préfèrent les classes homogènes, ils sont 71 % dans les autres établissements (Hallam, 2006). Mais y compris dans les premiers établissements, les élèves sont souvent favorables au regroupement par niveau homogène. Enfin, la proportion des élèves préférant les classes homogènes est la plus forte pour les élèves des groupes de niveau élevé (tableau X). Toutes ces différences sont statistiquement significatives.

Tableau X : proportion (%) des préférences des élèves regroupés en mathématiques

	Classes hétérogènes	Classes homogènes (*setting*, *streaming*)	Ne sait pas
Tous	24	68	7
Groupe de niveau élevé	11	85	5
Groupe de niveau moyen	20	73	7
Groupe de niveau faible	38	53	9

Une analyse discriminante a montré que les élèves qui préfèrent les classes homogènes ont de meilleurs scores KS3, ont connu une intensité plus forte de regroupements[2] en classes homogènes, sont de niveau socioéconomique plus élevé, ont une estime de soi plus élevée et sont affectés à la catégorie de niveau élevé. Les élèves qui préfèrent les classes hétérogènes ont un niveau de satisfaction vis-à-vis de l'école en général plus important.

En ce qui concerne le niveau de leur groupe, 38% des élèves de *year* 9 n'en étaient pas satisfaits (Hallam, 2007). Parmi eux, 77% souhaitaient changer pour un groupe de niveau plus élevé et une

[1] Cet indicateur varie pour chaque établissement entre 0 à 5 (0 pour un établissement ne regroupant pour aucun *year*, 5 pour le cas contraire).
[2] Cet indicateur varie pour chaque établissement entre 0 à 5 (0 pour un établissement ne regroupant pour aucun *year*, 5 pour le cas contraire).

différence statistiquement significative entre les filles et les garçons est notée : 83% des garçons souhaitant changer vers un groupe de niveau supérieur (13 % vers un groupe de niveau inférieur) pour seulement 68 % des filles (elles sont 24 % à souhaiter changer pour un groupe de niveau plus faible). 62 % des élèves de la catégorie de niveau faible, contre 45 % de la catégorie de niveau moyen et 16% des élèves de la catégorie de niveau élevé ne sont pas satisfaits du niveau de leur groupe. Dans chacune des trois catégories, les proportions des élèves souhaitant changer pour le groupe de niveau le plus élevé, ou pour un groupe de niveau plus élevé ou encore pour un groupe de niveau plus faible ont été calculés ici à partir des données de l'article (tableau X). La proportion la plus importante concerne les élèves des groupes de niveau faible puisque 58 % de ces élèves souhaitent aller dans un groupe de niveau plus élevé (y compris le *set* 1[1]) ; ils sont 37 % quand ils sont dans un groupe de niveau moyen.

Tableau X. Souhaits (%) des élèves

	Vers le groupe de niveau le plus élevé (set 1)	Vers un autre groupe de niveau plus élevé	Vers un groupe de niveau plus faible	Autre
Groupes de niveau élevé	2	0	13	1
Groupes de niveau moyen	16	22	6	1
Groupes de niveau faible	9	49	1	3

Être regroupé par niveau ou non a finalement peu d'effet sur l'estime de soi des élèves si l'on ne tient pas compte du niveau du groupe des élèves. Le regroupement des élèves en classe homogène est plus apprécié des élèves ayant des scores élevés, ayant souvent fréquenté des classes homogènes et qui sont dans un groupe de niveau élevé. Quand on tient compte des scores des élèves, l'effet du niveau de groupe sur l'estime de soi qui est favorable aux élèves des groupes de niveau élevé, tend à s'estomper. Plus de la moitié des élèves des groupes de niveau faible veulent changer de niveau.

2.4. Groupes de niveau et pratiques dans les établissements

On sait maintenant qu'en Angleterre les élèves sont en majorité regroupés par niveau en mathématiques au secondaire. Et que tous les acteurs de la communauté éducative ont conscience que cette organisation a comme conséquence de regrouper ensemble des élèves de niveau faible, de milieu socioéconomique défavorisé, issus de minorités ethniques et repérés comme ayant parfois des besoins particuliers. Et que leurs résultats académiques et non cognitifs ne sont probablement pas avantagés par ce type d'organisation. Les regards se tournent donc naturellement vers ces élèves perçus comme les victimes potentielles d'une organisation qui ne serait pas optimale, et c'est sous cet angle que nous aborderons la question des « pratiques dans les établissements ». Les études sont quantitatives et qualitatives ; elles ont exploité les données issues de questionnaires mais aussi d'entretiens et d'observations de cours.

Les interventions Best Practice d'EEF

Deux interventions ont été mises en œuvre et étudiées par EEF avec comme objectif principal l'amélioration des résultats des élèves de *year* 7 et *year* 8 en mathématiques et en anglais ; pour Best practice in setting (**étude E**) les élèves sont regroupés par niveau (Roy, 2018b) et pour Best Practice in Mixed Attainment (**étude D**), ils sont scolarisés en classe hétérogènes (Roy, 2018a). Une description générale peut être consultée dans Reassessing 'Ability' Grouping (Francis, 2019 a). Les deux interventions avaient comme élément commun une formation des enseignants visant à élever le niveau des élèves en difficulté. Pour la première, il s'agissait également de limiter le nombre de niveaux, de

[1] C'est traditionnellement le groupe de niveau le plus élevé.

constituer les groupes uniquement sur la base des scores KS2 et d'éviter que les meilleurs enseignants soient affectés aux élèves des groupes de niveau élevé. Pour la seconde, la constitution de classes hétérogènes en *year* 7 devaient se faire sur la base exclusive des score KS2.

Les résultats de ces études n'ont pas été concluants tant en ce qui concerne l'impact de l'intervention sur les scores des élèves que sur leur estime de soi, qu'en terme de mise en œuvre, de nombreux établissements ayant mis fin à leur participation avant la fin de l'intervention. Aucune des tailles d'effets calculées n'étaient significatives[1]. Ce faible impact tient probablement au fait que les groupes traitement et contrôle ont été, au regard des éléments des interventions, moins différents que prévu (Hodgen, 2019). En ce qui concerne l'**étude E**, l'intervention semble avoir eu un effet positif sur l'utilisation des scores KS2 pour la constitution des groupes[2] et la diminution du nombre de niveau dans les établissements du groupe intervention. La répartition des enseignants a tout de même souvent tenu compte des préférences ou compétences des enseignants (certains enseignants étant considérés comme plus aptes à enseigner à tel ou tel groupe de niveau) et de contraintes pratiques (d'emploi du temps par exemple). Cette remarque a été confirmée par Francis (2019 b) qui montre que l'intervention n'a pas significativement modifié la répartition des enseignants en fonction de leur niveau de qualification. C'est l'occasion de constater ici que les enseignants les plus qualifiés sont tout de même moins nombreux à enseigner dans les groupes de niveau le plus faible, même si ces différences ne sont pas significatives (tableau X).

Tableau X. Répartition des enseignants en fonction de leur diplôme (mathématiques ou anglais)

	Groupe contrôle		**Groupe intervention**
	Groupe de niveau le plus faible	**Les autres groupes**	**Groupe de niveau le plus faible**
Licence ou plus	46 %	65 %	62 %
A level (niveau bac + 1)	41 %	29 %	26 %
GCSE (niveau lycée)	14 %	6 %	13 %
Total	100 %	153 (100 %)	100 %

Lecture : 46% des enseignants des groupes de niveau le plus faible ont une licence ou plus.

Les enseignants ayant participé à l'intervention de l'**étude D** ont apprécié les effets positifs sur les élèves : une même attention a pu être donnée à chacun d'entre eux, quel que soit son score KS2 ; l'absence de stigmatisation des élèves qui auraient été autrefois affectés dans le groupe de niveau faible a été reconnue ; les élèves de niveau élevé ont eu l'occasion d'approfondir leurs connaissances plutôt que de passer (peut-être trop rapidement) à un autre sujet. L'effet le plus remarquable concerne les élèves de niveau faible qui ont exprimé une plus grande confiance en eux et le sentiment de pouvoir progresser plus facilement dans une classe hétérogène. Une étude approfondie a été menée auprès de 3 établissements ayant participé à l'intervention et de niveau Ofsted exceptionnel (Taylor, 2017). Elle a permis de définir les éléments jugés par les enseignants comme favorables à l'enseignement en groupes hétérogènes mais aussi des éléments jugés comme étant plus négatifs (tableau X). On doit souligner ici que la crainte de ne pas mettre en place un enseignement satisfaisant avec des groupes hétérogènes est particulièrement vive pour les mathématiques, et que très peu d'enseignants de cette discipline ont enseignés à des élèves non regroupés par niveau.

[1] Ceci étant, et comme nous l'avons signalé, la méta-analyse publiées par EEF sur la question du *Setting et Streaming* (EEF, 2021) a tout de même inclue les résultats de la seconde étude.

[2] Le genre et l'origine ethnique des élèves ont tout de même été pris en considération lors de la constitution des groupes, dans un souci de mixité

Tableau X. Éléments positifs ou négatifs vis-à-vis de l'enseignement en groupes hétérogènes

Éléments positifs	Éléments négatifs
• Bénéfices pour les élèves • Augmentation de l'inclusion et de l'équité • Amélioration de la qualité de l'enseignement	• Résultat défavorable possible après une inspection de l'Ofsted • Manque de soutien de la direction • Réactions défavorables des parents • Réticence au changement de certains collègues • Charge de travail élevée • Facteurs pédagogiques : différenciation, rythme, difficultés spécifiques aux mathématiques, …

Enseigner aux élèves de niveau faible

Une étude de cas incluse dans l'**étude C** a concerné 5 classes de *year* 5 (donc des élèves de primaire), 7 classes de *year* 8 et 7 classes de *year* 10 (Dunne, 2007 et 2011) dans des établissements où les élèves des groupes de niveau faibles progressaient de façon satisfaisante. Ces données ont conduit les auteurs à établir une liste d'éléments positifs regroupés en deux points : le premier concerne l'établissement et l'organisation des enseignements placée sous la responsabilité de la direction et des autorités de tutelle (tableau X), le second concerne l'enseignement en grande partie sous la responsabilité des professeurs (tableau X). Ces recommandations peuvent concerner les élèves du primaire (qui scolarisent parfois les élèves de niveau faible dans des classes hétérogènes) et plusieurs disciplines.

Tableau X. Organisation des enseignements

Les ressources • effectifs des classes de niveau faible réduits • Les assistants et tuteurs[1] disponibles pour soutenir l'enseignant dans son enseignement et l'élève dans son apprentissage • collaboration entre enseignants et assistants • niveau de qualification élevé des enseignants	**Les programmes et évaluations** • programmes similaires dans les différents groupes de niveau pour faciliter la mobilité • utilisation de tests nationaux entre autres, feedback individuel • au niveau KS4, différenciation des programmes, avec parfois une partie professionnalisante • élaboration des programmes, des évaluations, des EDT ciblée sur les apprentissages et les résultats des élèves de niveau faible
Faciliter des relations positives • un ethos de l'établissement explicite, inclusif et stimulant • prix et récompenses pour stimuler la réussite des élèves • consultation des élèves et de leurs parents pour les responsabiliser	**Les relations extérieures** • implication des parents (visite à domicile, réunions dans l'établissement, devoirs à la maison) • organisation éventuelle de cours pour les parents • communication renforcée avec certains spécialistes (thérapeutes, traducteurs, ...)

Tableau X. Enseignement

Les ressources • les assistants jouent un rôle important, au-delà de leur attachement à un ou deux élèves •les élèves des groupes de niveau faible ont le même accès au matériel pédagogique que les autres	
Les programmes et évaluations •vitesse réduite, soutien renforcé, niveau de difficulté diminué, soutien des pairs, feedback et encouragement •opportunité pour les élèves de choisir le niveau de difficulté des exercices qui est varié •renforcer les acquisitions : répétition et/ou nouvelles activités[2] •utilisation des TIC, du tableau interactif, parfois comme récompense[3]	**Faciliter des relations positives** • négociations plus fréquentes ; activités ludiques et amusantes comme récompense • règles disciplinaires explicites pour faciliter la concentration des élèves • participation des élèves encouragée, utilisation positive des erreurs

[1] Les assistants *(teacher assistant)* jouent un rôle pédagogique important en soutenant l'enseignant dans son travail ; les tuteurs (*learning mentor*) aident les élèves dans leurs apprentissages.

[2] Certains enseignants favorisent la répétition des tâches, d'autres au contraire la juge peu pertinente.

[3] Concerne plutôt le primaire.

Ce que disent les élèves

Dans le cadre de l'**étude E**, l'analyse qualitative d'entretiens menés avec 118 élèves de *year* 7 montre que certains élèves considéraient que les enseignants des groupes de niveau plus élevé ont des exigences plus élevées, aussi bien en ce qui concerne le comportement des élèves que leurs apprentissages. À l'inverse les enseignants des groupes de niveau plus faible sont parfois perçus comme peu exigeants, plus indulgents voire infantilisants (Francis, 2019 b). Les auteurs affirment que les élèves de *year* 7 sont conscients de l'effet d'étiquetage (voir plus loin) induit par leur répartition dans des groupes de niveau. Ces résultats sont confirmés par les réponses d'élèves de *year* 8 et de *year* 9 de l'**étude A** (Boaler, 2000), les élèves du groupe de niveau le plus élevé (*set* 1) se plaignant d'une trop grande rapidité des cours, d'une difficulté importante des exercices et du manque d'informations données par l'enseignant. Les élèves du groupe de niveau le plus faible regrettent d'avoir des exercices trop faciles et d'être condamné à rester dans leur niveau : 32 % des élèves du groupe de niveau le plus bas trouvent les exercices trop faciles, ils ne sont que 7 % dans les autres groupes.

Les élèves interrogés dans le cadre de l'**étude B** ont expliqué leurs préférences pour les regroupements hétérogènes ou homogènes (Hallam, 2006). Seul 57% des élèves ont répondu au questionnaire et les réponses concernent essentiellement les classes homogènes. Ainsi 47% des élèves présentent ce type de regroupement comme permettant de travailler à son propre niveau, en compagnie d'élèves de niveau similaire. Toutes les autres réponses ont recueilli des taux inférieurs à 4%. Enfin, la raison majeure invoquée par les élèves pour changer de groupe de niveau concerne leurs apprentissages, puisqu'il s'agissait de travailler sur des tâches plus difficiles pour 27% d'entre eux, ou sur des tâches plus faciles pour 15% ; 11% évoquent un motif lié au statut (les *set* 1 ou 2 qui sont les meilleurs groupes sont aussi les plus respectés) et 7 % d'entre eux indiquent être suffisamment intelligents pour aller dans un groupe de niveau plus élevé[1] (Hallam, 2007)

2.5. Conclure sur la recherche anglaise

La recherche anglaise est sans doute arrivée au bout d'une démarche d'investigation et les résultats de l'étude EEF actuellement en cours (Hodgen, 2019) pourraient apporter un point final à plus d'un quart de siècle d'analyses. Si la promesse est tenue, ce sera l'unique étude comparative anglaise sur cette question, seul type d'étude permettant d'éliminer (ou tout du moins de contrôler) la plupart des biais de confusion. Les études dont nous avons détaillé les résultats ici montrent que pour trois élèves de même niveau scolaire, celui qui est affecté dans un groupe de niveau élevé aura probablement de meilleurs résultats que celui qui est affecté dans un groupe de niveau moyen lui-même favorisé par rapport à celui se retrouvant dans un groupe de niveau faible ; cette situation est rendue possible car, comme nous l'avons vu, des élèves de scores similaires sont parfois répartis dans des groupes de niveau différent. La corrélation entre le niveau socioéconomique et le niveau des groupes a également été confirmée, y compris après contrôle des scores initiaux : les élèves des groupes de niveau élevé sont plutôt d'un niveau socioéconomique favorisé tandis que les élèves des groupes de niveau faible sont plutôt d'un niveau socioéconomique défavorisé. Les études anglaises, en intégrant plusieurs types d'analyses et en multipliant les champs d'investigation, ont apporté des précisions sur les modalités des regroupements par niveau ainsi que sur leurs effets sur des caractéristiques non cognitives des élèves. La question complexe de l'organisation des enseignements en lien avec la répartition des élèves en groupe de niveau a également été abordée. Au final, quand un chef d'établissement ou des enseignants considèrent la possibilité de regrouper leurs élèves par niveau homogène, ils se doivent

[1] Une seule réponse par élève.

de considérer six points d'attention qui ont comme corolaire des situations problématiques qu'il convient d'éviter (tableau X), en accordant une attention particulière aux groupes de niveau faible (Francis, 2017a, 2019a). Ces différents éléments vont être repris en relation avec des points théoriques dans la dernière partie de cette synthèse.

Tableau X. Groupes de niveau : points d'attention et problèmes

Points d'attention	Situations problématiques
La répartition des élèves dans les groupes de niveau	Les élèves ne sont pas toujours répartis en fonction de leur niveau académique et leur niveau scolaire ne correspond pas toujours à leur groupe de niveau
La flexibilité des groupes de niveau	Les élèves restent dans leur groupe initial et l'évolution de leurs compétences ne sont pas prises en compte
La qualité de l'enseignement	Les enseignants les moins expérimentés sont affectés aux groupes de niveau faible
Les attentes des enseignants envers les élèves	Les enseignants ont des attentes trop faibles pour les élèves des groupes de niveau faible et trop élevées pour les élèves des groupes de niveau élevé et l'enseignement n'est pas différencié
Les programmes et les évaluations	Les élèves des groupes de niveau faible ne sont pas évalués de la même façon et ne suivent pas les mêmes programmes que les autres, ce qui limite leur opportunité d'apprendre.
La mixité sociale des groupes de niveau	Les élèves de milieux socioéconomiques faibles sont surreprésentés dans les groupes de niveau faible, y compris après contrôle de leur score initial.

Plusieurs limites à cette recherche doivent tout de même être soulignées ici. Si la qualité des études EEF est incontestable, il n'en est peut-être pas de même pour les autres analyses[1]. Certains choix dans les résultats calculés comme dans les méthodes suivies semblent plus ou moins consciemment pilotés par les convictions des chercheurs qui ne sont pas favorables aux regroupements des élèves par niveau tels que pratiqués en Angleterre actuellement. On a constaté que le regroupement des élèves par niveau se fait en suivant des organisations très différentes, qui peuvent varier d'un *year* à l'autre mais aussi d'une année à l'autre, et leur caractérisation s'avère plus complexe que prévue. Certains établissements regroupent dès *year* 7, d'autres un voire deux ans après. Le nombre de niveaux de groupes est aussi très variable pouvant aller de deux à 10 (aucune donnée exhaustive n'a été trouvée). Cette diversité des pratiques tout comme leur grande variabilité dans le temps complique la tâche des chercheurs. Pour inclure plusieurs établissements différents et donc plusieurs organisations différentes dans une même étude, ces derniers ont toujours défini 3 catégories de groupes : les catégories de niveau élevé, moyen et faible. Mais d'une étude à l'autre, ces catégories ne sont pas déterminées de la même façon : parfois la catégorie de niveau élevé n'inclue que les élèves du groupe de niveau le plus élevé, dans d'autres cas elle inclue 25% des élèves rassemblés dans les groupes de niveau(x) élevé(s). On peut raisonnablement penser que les résultats des études utilisant des catégories du premier type sont plus tranchés que les résultats des études utilisant des catégories du second type.

L'attrition des données a été constatée pour presque toutes ces études et on ne peut exclure que les caractéristiques de l'échantillon analytique soient différentes de celles de l'échantillon de départ, ce qui diminue la qualité des inférences. Cette attrition concerne les études randomisées par

[1] Y compris des analyses des données récoltées lors des études EEF.

établissements (avec des établissements qui quittent parfois le groupe intervention comme pour les **études D** et **E**), mais aussi les données récoltées par l'intermédiaire de questionnaires. C'est pour cette raison qu'aucun des résultats concernant l'origine ethnique des élèves n'a été présenté ici[1].

On terminera cette liste en soulignant une dernière limite pour qui s'intéresse plus particulièrement à l'enseignement des mathématiques. En effet, l'Angleterre n'a pas été le « laboratoire de recherche » espéré, étant donné que dans cette discipline presque tous les élèves du secondaire inférieur sont regroupés par niveau. Concrètement, la comparaison entre des élèves non regroupés par niveau d'une part et des élèves regroupés par niveau d'autre part n'a pu concerner qu'un très faible échantillon. Et il est possible que le côté atypique voire non conformiste du regroupement hétérogène ait influencé les résultats observés.

[1] L'OCDE signale dans les enquêtes PISA les pays qui ne satisfont pas à leurs exigences pour ce point ; le What Works Clearinghouse a également fixé des règles de conduites en la matière (WWC, 2022) et certaines études sont parfois exclues de leurs méta-analyses pour ce motif.

Discussion et conclusion

Nous aborderons dans cette dernière partie trois points. En premier lieu, il s'agira de développer les arguments issus de théories psychologiques ou sociales mais aussi de la pratique des enseignants et qui soutiennent pour certains le regroupement des élèves en classes de niveau homogène, pour d'autres le regroupement des élèves en classes de niveau hétérogène. D'autres arguments alertent sur les inconvénients de telle ou telle organisation (et dans ce cas, une absence d'inconvénient peut constituer un élément favorable). Dans un second temps, nous récapitulerons les résultats des recherches à la lumière de ces apports théoriques, puis les limites de cette synthèse seront à nouveau soulignées avant de conclure.

1. Les théories et les pratiques

Presque tous les articles dont nous avons exploité les résultats proposent en préambule des arguments qui soutiennent ou au contraire qui sont en défaveur des regroupements des élèves en classes homogènes et/ou hétérogènes. Il a été volontairement décidé dans cette synthèse d'évoquer cet aspect après avoir présenté les résultats de la recherche, afin que ces derniers puissent apporter leur éclairage. Ces arguments sont des concepts théoriques, fruits des réflexions et recherches conduites par des psychologues ou des sociologues, mais ils peuvent également être conçus de façon pragmatique par des enseignants expérimentés. Tous concernent *in fine* les compétences scolaires dont l'élévation est l'objectif principal des systèmes scolaires.

Mais avant tout, on se doit de répéter une fois encore que tout regroupement d'élèves est susceptible d'être considéré à la fois comme un regroupement d'élèves de niveau hétérogène et comme un regroupement d'élèves de niveau homogène. Pour la première partie de cette remarque, il suffit de reconnaitre que les élèves sont tous différents les uns des autres, et ce quelle que soit l'organisation mise en place. La seconde partie vient de ce que chaque classe d'élèves est constituée de façon à en limiter l'hétérogénéité, ne serait-ce qu'en regroupant des élèves d'âge similaire : il est évident que l'hétérogénéité est une question (voire un problème) que le système scolaire se doit de traiter. Certains des éléments présentés ci-dessous peuvent donc concerner à la fois les groupes de niveau homogène et les groupes de niveau hétérogène (tels qu'ils ont été définis dans cette synthèse). Il peut également être intéressant de traduire un argument par exemple favorable aux regroupements homogènes en un argument défavorable aux regroupements hétérogènes. C'est ce qui a été fait dans le tableau X qui synthétise ces éléments.

L'argument principal en faveur des regroupements par niveau s'appuie sur les concepts de l'opportunité d'apprendre et de la zone proximale de développement (tous les deux tournés vers l'apprenant), ainsi que sur la différenciation pédagogique (en direction de l'enseignant cette fois). On peut le résumer ainsi : regrouper les élèves par niveau réduit l'étendue des capacités des apprenants ce qui permet au professeur d'adapter plus facilement son enseignement aux besoins de chacun, en sélectionnant les connaissances et compétences visées ainsi que le niveau de difficulté des activités. D'un point de vue pragmatique, cela permet à l'enseignant de diminuer le spectre des contenus à transmettre et de les adapter au mieux à chacun de ses élèves. Le travail se resserrant sur des objectifs plus ciblés gagne en pertinence, en efficacité et finalement en qualité. Les mathématiques occupent ici une place particulière, leur substance étant perçue comme une succession logique et ordonnée de concepts qui oblige les élèves à s'appuyer sur des notions déjà acquises pour en comprendre de nouvelles plus complexes. Cet élément théorique est pour les enseignants d'une telle évidence qu'on pourrait tout aussi bien considérer qu'il est issu de leurs réflexions professionnelles.

Un autre aspect théorique très souvent convoqué dans les textes traitant des regroupements par niveau est l'effet d'étiquetage (*labelling*). Il se traduit ici en ces termes : le niveau du groupe auquel les élèves sont affectés est associé à chacun de ces élèves, ce qui favorise les auto prophéties réalisatrices. « Je suis un élève du groupe faible (moyen, fort), donc je suis de niveau faible (moyen, fort), et je serai toujours de niveau faible (moyen, fort) ». L'élève et son entourage adoptent de façon souvent inconsciente des comportements qui favorisent ces prophéties qui finissent par se réaliser. Si ces comportements sont favorables aux élèves des groupes de niveau élevés (on parle de l'effet pygmalion avec une estime de soi augmentée), ils impactent négativement les élèves des groupes de niveau faible (ici c'est l'effet effet golem qui est cité, avec une estime de soi diminuée). Cet effet négatif sur les élèves des groupes de niveau faible est unanimement reconnu et constitue de fait une question à part entière pour les chercheurs ; ne pas en faire cas c'est courir le risque de stigmatiser une partie des élèves. Les précautions langagières que tout un chacun est amené à prendre en témoigne. Derrière tout ceci se cache une inévitable hiérarchisation des niveaux qui n'échappe à personne : aucune famille (aucun enseignant) n'a comme objectif de placer son enfant (son élève) dans le groupe de niveau le plus faible[1].
Il faut tout de même reconnaitre ici que le regroupement en classes homogènes n'est pas à lui seul responsable de l'étiquetage des élèves et du mal être qu'il peut engendrer. La tendance naturelle à attacher à chaque individu des caractères plus ou moins permanents explique que certains élèves sont considérés par leurs pairs, leurs enseignants, leurs parents ou par eux-mêmes comme peu performants (dans telle ou telle discipline), ou inversement comme très compétents, et cela concerne aussi les élèves en classes hétérogènes. Mais l'effet d'étiquetage associé aux regroupements explicites amplifie cette catégorisation en limitant notamment la possibilité pour tout un chacun d'échapper à des classements qui peuvent marquer de façon indélébile.
Les avocats des groupes hétérogènes s'appuient sur la théorie du *growth mindset* que l'on peut traduire par état d'esprit de croissance, théorie développée avec succès dans les années 2010 par Dweck (2010). Cette chercheuse a montré que les élèves comme les adultes peuvent être répartis dans trois catégories : ceux qui pensent que l'intelligence (ou les capacités) est un trait de personnalité fixe qui n'évolue pas (*fixed mindset*), ceux qui pensent que l'intelligence est malléable et peut se développer grâce notamment au travail (*growth mindset*), et enfin ceux chez qui ces deux conceptions coexistent[2].
Une synthèse des recherches effectuées sur cette question (Boaler, 2017) a montré que les interventions ayant comme but d'encourager les élèves à concevoir leur intelligence et leurs capacités comme malléables ont eu des effets positifs sur leurs scores. Cet effet concerne les élèves de niveau faible qui sont encouragés à travailler pour progresser. Mais également les élèves de niveau élevé qui se montrent souvent inquiets de ne pas réussir voire de décevoir (Boaler, 1997), et ne craignent plus de commettre des erreurs considérées dans cette théorie comme favorisant la compréhension quand elles sont correctement traitées.
On termine cette liste d'arguments psychologiques par l'effet « du gros poisson dans le petit étang[3] » (Marsh, 1984) qui favorise cette fois les regroupements homogènes en affirmant que l'estime de soi des élèves de niveau faible est renforcée quand ils sont placés dans des groupes de niveau faible. N'étant entourés que de pairs de niveau scolaire équivalent, ils ne souffrent pas alors de la comparaison avec des élèves de niveau plus élevé telle qu'elle peut être vécue dans les classes hétérogènes. Dans

[1] Dans un questionnaire proposé aux élèves, si la possibilité pour eux d'aller dans le meilleur groupe était évoquée, celle d'être affecté dans le groupe de niveau le plus faible était absente des propositions (Hallam 2006).
[2] Les deux interventions étudiées par EEF dont nous avons présenté quelques résultats sont basées sur cette théorie.
[3] *Big fish little pond effect*

un environnement plus bienveillant et compréhensif, les élèves se sentent accueillis pour ce qu'ils sont, osent davantage s'exprimer et progressent plus facilement.

D'autres avantages (plus pragmatiques cette fois) peuvent concerner soit les élèves de niveau élevé, soit les élèves de niveau faible d'une part, et soit les classes homogènes soit les classes hétérogènes d'autre part. En ce qui concerne les avantages pour les élèves de niveau élevé d'être regroupé par niveau, on devine plus qu'on ne lit que l'absence d'élèves de niveau faible est favorable à leurs apprentissages. Dans les classes hétérogènes, la présence d'élèves de niveau faible est susceptible de freiner les acquisitions des élèves de niveau élevé, car les premiers monopolisent une part importante de l'attention de l'enseignant et ralentissent de ce fait le rythme des leçons et des acquisitions des seconds. On sous-entend ici que les élèves de niveau faible ont souvent des comportements inadaptés au bon fonctionnement des apprentissages en perturbant les cours et donc les autres élèves. Nous avons volontairement laissé de côté ici les études concernant les élèves de niveau très élevé (c'est-à-dire les élèves à haut potentiel intellectuel ou *gifted children*). Ils ont été la source de l'attention de quelques auteurs des méta-analyses dont nous avons discuté les résultats, cette attention exprimant par elle-même la crainte ressentie par certains acteurs de la communauté éducative que l'hétérogénéité des classes ne leur nuise. Toujours sur la question de la cohabitation d'élèves de niveaux très différents, un autre argument vient cette fois en faveur des classes hétérogènes pour les élèves de niveau élevé. En effet, cette cohabitation peut être aménagée de façon à ce que ces élèves viennent en aide à leurs pairs moins performants ce qui peut leur permettre d'approfondir certaines connaissances. Le seul fait d'expliquer une notion permet parfois d'en découvrir des aspects qui n'ont pas été perçus au premier abord[1] ; certaines questions ou réflexions inattendues d'élèves moins compétents sur un point particulier peuvent aussi avoir le même effet. Enfin, ces classes peu sélectives apparaissent également comme plus accueillantes pour qui n'apprécie pas la compétition.

Pour les élèves de niveau faible cette fois, l'avantage des classes homogènes clairement mis en avant par les autorités concerne les effectifs allégés de leurs groupes quand on les compare aux groupes de niveau plus élevé. Là aussi, cela permet aux professeurs d'adapter au mieux leur enseignement aux élèves dont ils ont la charge, et d'individualiser les apprentissages. Dans les classes hétérogènes cette fois, ces élèves bénéficient d'interactions stimulantes avec des pairs de niveau plus élevé (les enseignants parlent souvent d' « entrainement vers le haut »). Ces interactions peuvent expliquer en partie l'effet positif qu'il y a pour tout élève de se trouver dans une classe de niveau élevé tel que cela a été observé par Duru-Bellat (1997) et Hallimam (2000).

Dans tout ce qui vient d'être dit, on notera l'absence remarquable des élèves de niveau moyen. Ils disparaissent ici pour laisser la place à ceux qui peuvent poser problème : l'élève de niveau élevé qui risque de ne pas réussir autant qu'il pourrait et l'élève de niveau faible qui risque de décrocher (voire de déstabiliser le système). Finalement, ces élèves des groupes de niveau moyen ont de la chance. Contrairement aux élèves des groupes de niveau faible qui peuvent être vus comme des « idiots » (par les autres élèves, par les adultes qui les entoure ou par eux-mêmes), et aux élèves des groupes de niveau élevé qui peuvent souffrir d'être considérés comme des « intellos », les élèves des groupes de niveau moyen passent inaperçu. Ces classes paraissent alors plus accueillantes et satisfont au besoin de « normalité » de certains adolescents : c'est ce qui explique que certains élèves de niveau élevé souhaitent aller dans un groupe de niveau plus faible (probablement perçu comme « moyen »).

Les auteurs des articles sur les études anglaises soulignent souvent l'absence de recherches et de connaissances sur l'enseignement en classes hétérogènes. Il en résulte bien sûr que peu d'éléments

[1] Il s'agit finalement de penser à voix haute, stratégie conseillée par certains psychologues.

pédagogiques plaident en faveur de ces groupes hétérogènes, mais aussi et surtout en leur défaveur. Au contraire, l'enseignement en classe homogène ayant fait l'objet de nombreuses analyses, parfois « à charge », ses défauts sont bien identifiés et ses qualités peu soulignées. Cet état de fait surestime probablement les inconvénients des regroupements en classes de niveau, et par contre coup les avantages des regroupements hétérogènes.

Tableau X. Points forts et faibles des regroupement homogènes ou hétérogènes.

	Les points favorables	**Les points défavorables**
Groupes homogènes	a. Faible étendue des compétences entrainant une efficacité accrue de l'enseignant (théorie et pratique). b. Effet d'étiquetage : les élèves de niveau élevé ont une estime de soi plus forte (théorie). c. Les élèves des groupes de niveau faible osent s'exprimer d'avantage (théorie). d. Les acquisitions des élèves des groupes de niveau élevé ne sont pas limitées par les élèves de niveau faible (pratique). e. Les effectifs réduits des groupes de niveau faibles permettent à l'enseignant d'être plus efficace (pratique).	f. Effet d'étiquetage : les élèves des groupes de niveau faible ont une estime de soi plus basse (théorie). g. Moins d'entraide entre les élèves (pratique). h. Les élèves des groupes de niveau faible ne sont pas entrainés par des élèves de niveau élevé (pratique). i. Les élèves de niveau élevé sont sous pression (pratique). j. Favorise un état d'esprit fixe décourageant les élèves de niveau faible et stressant les élèves de niveau élevé (théorie).
Groupes hétérogènes	f. Pas d'effet d'étiquetage : les élèves de niveau faible ont une estime de soi plus élevée (théorie). g. Entraide entre élèves de niveau différents (pratique). h. Les élèves de niveau faible sont entrainés par les élèves de niveau élevé qui servent de modèles (pratique). i. Les élèves de niveau élevé ne sont pas sous pression (pratique). j. Favorise un état d'esprit de croissance qui encourage les élèves de niveau faible et rassure les élèves de niveau élevé (théorie).	a. Grande étendue des compétences entrainant une charge importante de travail pour l'enseignant (théorie et pratique). b. Pas d'effet d'étiquetage : les élèves de niveau élevé ne sont pas suffisamment reconnus et leur estime de soi n'est pas stimulée (théorie). c. Les élèves de niveau faible souffrent du regard de leurs pairs de niveau plus élevé (théorie). d. Les acquisitions des élèves de niveau élevé sont limitées par les élèves de niveau faible (pratique). e. Les effectifs des classes sont élevés et les élèves de niveau faible sont moins aidés par l'enseignant, parfois même abandonnés (pratique).

Chaque élément est présenté deux fois : par exemple en faveur des groupes homogènes et en défaveur des groupes hétérogènes (ils sont repérés par des lettres et reformulés) ; il est précisé également s'ils sont issus de théories psychosociales ou de la pratique des enseignants.

2. Les résultats

Les résultats concernent essentiellement deux points : la description des groupes de niveau (leurs caractéristiques et les règles présidant à leur constitution), puis les effets de l'enseignement en groupe de niveau sur les élèves.

Les avantages théoriques de la constitution de groupes de niveau vus précédemment imposent que ces derniers soient constitués sur la seule base des compétences des élèves (en fonction de la discipline concernée) : pour répondre aux besoins des élèves, ces derniers doivent être identifiés et leur niveau de connaissance doit être mesuré précisément et objectivement. En Angleterre on a vu que les scores KS2 (issus de tests standardisés) ne sont pas toujours les seules mesures utilisées pour constituer les groupes, ce qui contrevient aux recommandations énoncées par une partie des chercheurs qui

considèrent ces scores comme étant les plus impartiaux possible. Nous avons constaté également que certains élèves n'étaient pas placés dans le groupe correspondant à leur niveau scolaire antérieur et que cela concerne essentiellement les élèves de niveau moyen. D'autres éléments comme la recherche d'une certaine mixité (de genre ou d'origine ethnique) ou des contraintes pratiques (comme la gestion des emplois du temps) peuvent également peser sur les affectations des élèves dans un groupe et parfois dans un niveau. Enfin, la mobilité des élèves dans les groupes, c'est-à-dire la possibilité pour chacun d'entre eux de changer de niveau en fonction de l'évolution de leurs résultats, est parfois mise à mal pour des questions concrètes d'organisation des enseignements. Bien peu de données précises ont été trouvées à ce sujet.

Plusieurs études ont montré que les élèves issus de catégories sociales défavorisées ou présentant des troubles de l'apprentissage sont surreprésentés dans les groupes de niveau faible, y compris quand le niveau scolaire antérieur a été pris en compte. Cela revient à dire que pour deux élèves ayant obtenu des scores similaires, l'élève de niveau socioéconomique faible ou dyslexique, a une probabilité supérieure d'être affecté dans un groupe de niveau faible qu'un élève ne présentant pas ces caractéristiques. Muijs (2010) suggère qu'un lien entre le statut socioéconomique et des problèmes de comportement puisse influencer les enseignants ; Connolly (2019) invite les chercheurs à s'intéresser davantage aux élèves borderline (voir ci-dessus) pour comprendre ces surreprésentations qui restent pour l'instant inexpliquées. On retiendra que de toutes façons, la corrélation entre le statut socioéconomique et les résultats scolaires étant largement démontrée, tout le monde sait que constituer des groupes de niveau c'est rassembler des élèves de milieu socioéconomiques similaires, ce qui contrevient aux objectifs de mixité socioculturelle des sociétés modernes. Et cette répartition inégale des élèves impose au regroupement par niveau d'avoir un effet positif sur les compétences et connaissances des élèves[1]. Cette remarque nous amène donc à faire le point sur l'impact que cette organisation scolaire a sur les scores académiques des élèves, tel qu'il a pu être évalué par un siècle de recherches. Deux types de comparaisons ont été conduites. Pour le premier, qui concerne les études américaines, les élèves regroupés par niveau sont comparés à des élèves de niveau similaire non regroupés par niveau. Pour le second, qui concerne les études anglaises, on compare des élèves de niveau similaire regroupés dans des groupes de niveaux différents.

Cinq résultats clés peuvent être retenus :

- Les scores des élèves en moyenne (pour l'ensemble de tous les élèves) ne sont pas impactés par la constitution de groupes de niveau.
- A score initial équivalent, les scores des élèves de niveau élevé sont probablement un peu plus faibles dans les classes hétérogènes que dans les classes homogènes.
- A score initial équivalent, les scores des élèves de niveau faible sont probablement un peu plus élevés dans les classes hétérogènes que dans les classes homogènes.
- A score initial équivalent, les scores des élèves des groupes de niveau élevé sont probablement un peu plus élevés que les scores des élèves des groupes de niveau moyen
- A score initial équivalent, les scores des élèves des groupes de niveau faible sont probablement un peu plus bas que les scores des élèves des groupes de niveau moyen.

Une incertitude caractérise les quatre derniers points ; la différence des opportunités d'apprentissages est également un facteur à prendre en considération. En admettant qu'ils ne soient pas vérifiés, on

[1] Pour une discussion sur la balance équité-efficacité voir la méta-analyse de Terrin (2023)

pourra retenir qu'aucun élément ne vient en faveur des hypothèses contraires. Et entre autres qu'il est peu probable que les élèves des groupes de niveau faible aient profités des regroupements par niveau.

En ce qui concerne l'estime de soi et la satisfaction des élèves, les études montrent qu'elles sont également corrélées au niveau des groupes, et par exemple que les élèves des groupes de niveau élevé ont une estime de soi plus élevée. Mais l'estime de soi étant elle-même corrélée aux résultats scolaires de l'élève, il n'est pas possible d'attribuer cet effet au seul fait d'avoir été affecté à un groupe de niveau élevé. Par contre être regroupé par niveau ou non semble clairement ne pas jouer de rôle ici. En ce qui concerne la satisfaction des élèves, le lien causal est plus facile à mettre en évidence. Sans surprise, les élèves des groupes de niveau faible sont plus nombreux à vouloir changer de niveau que les autres. Les recherches qualitatives montrent que ces élèves savent et regrettent dès leur arrivée dans le secondaire que les opportunités d'apprendre mais aussi les niveaux d'exigence (en termes d'apprentissage, mais également d'attendus concernant aussi bien leurs résultats que leur comportement) soient moins élevés dans les groupes de niveau faible. Ils ont tout à fait conscience des enjeux et des problématiques sous-jacents aux pratiques de regroupement par niveau.

Enfin, le niveau d'expertise des enseignants (évalué pour une discipline donnée) peut également être associé au niveau des groupes, les enseignants les plus experts se trouvant plus souvent affectés aux groupes de niveau élevé, ce qui contrevient aux principes d'égalité de traitement des élèves. *A contrario*, des enseignants particulièrement motivés et parfois spécialement formés ont souhaité enseigner aux élèves les plus faibles ; Mais peu de données confirment ce qui reste pour l'instant à vérifier.

Finalement, pour expliquer pourquoi les élèves de niveau faible réussissent moins bien quand ils sont rassemblés dans un groupe de niveau faible, on revient une dernière fois sur les éléments théoriques et pratiques vus précédemment : les éléments positifs repérés ne suffisent sans doute pas à contrebalancés les effets négatifs (tableau X).

Tableau x : éléments en défaveurs et en faveurs du regroupement des élèves de niveau faible

Éléments défavorables	Éléments favorables
• Effet d'étiquetage • Faibles compétences / expérience des enseignants • Absence d'entrainement des bons élèves • Sentiment d'exclusion sociale • Faibles opportunités d'apprentissage	•Effectifs réduits •Absence du jugement des élèves de bon niveau •Enseignants spécialisés ou motivés

Au bout du compte, bien peu de résultats plaident en faveur du regroupement des élèves par niveau, qui n'a pas montré les effets attendus sur les scores académiques et qui renforce les stratifications sociales dont nos sociétés disent ne plus vouloir. Les enquêtes PISA vont également dans ce sens. Et si on prend en compte l'ensemble des conclusions qui ont traits aux résultats académiques des élèves mais aussi à l'équité des systèmes scolaires, on ne peut que reprendre les propos de Slavin (1990, p.494) et affirmer que la charge de la preuve repose encore et toujours sur le regroupement par niveau homogène, et non sur le regroupement en classes hétérogènes.

3. Conclusion

Nous avons pu le constater dans la troisième partie de cette synthèse : en Angleterre, un socle solide de connaissances issues de recherches financées par les pouvoirs publics n'a clairement pas été pris en compte par ces derniers. Becky Francis qui dirige Education Foundation Endowment depuis 2020 et

dont le nom est associé à cette recherche anglaise, l'explique par la prégnance dans la classe moyenne anglaise d'un discours faisant la part belle au principe d'un 'ordre naturel' des capacités des élèves. Ces dernières sont considérées par une grande partie de la population comme déterminées, voire figées et hiérarchiquement distribuées. Les pouvoirs publics s'appuieraient sur cette vision pour rendre des arbitrages en maintenant de ce fait une hiérarchie sociale traditionnelle (Francis, 2017a).

La France n'est pas l'Angleterre. Les études quantitatives y restent marginales (s'agissant de leur réalisation bien sûr, mais également de leur prise en compte par la communauté scientifique) et n'ont pas les faveurs des chercheurs en sciences de l'éducation[1]. Ce n'est pas le Conseil Scientifique de l'Éducation National mis en place par Jean-Michel Blanquer en 2018 qui démentira cette affirmation : si ses membres se réclament de l'*evidence based education* (que l'on peut traduire par enseignement fondé sur des éléments probants), le manque de financement ne leur permet d'envisager d'autres actions que la publication de quelques synthèses et de recommandations qui n'arrivent pas toujours aux oreilles des enseignants. Mais « la science » reste une valeur sur laquelle il fait bon s'appuyer ; y faire référence s'est donc faciliter l'adhésion du plus grand nombre quand il s'agit de préconiser telle ou telle action, voire de réformer un système. Le ministère de l'Éducation Nationale français l'a bien compris. La synthèse de Dupriez (2010) est l'un des deux documents cités par le dossier de presse publié le 5 décembre qui détaille la réforme Choc des savoirs. Cette citation vient appuyer la constitution de « groupes de niveaux flexibles », qui sont décrits comme des groupes de niveaux disciplinaires dont la composition peut varier. Le vocabulaire n'est pas explicité et aucune autre précision n'est donnée dans le dossier de presse, mais tout laisse croire qu'un même mot peut avoir plusieurs sens, car pour qui se donne la peine de les lire, les conclusions de Dupriez vont à l'encontre de l'organisation retenue par le ministre français (organisation clarifiée dans un décret paru ultérieurement[2]). La seconde publication est une note publiée par IDEE[3] en novembre 2023. On devine en la lisant que les auteurs répondaient à la commande de la mission Exigence des savoirs ; leur très courte synthèse confirme l'inefficacité des regroupements permanents et soutient, comme Dupriez, les regroupements transitoires (à l'intérieur de la classe ou sous forme de tutorat par exemple) qui encore une fois ne correspondent pas à ce que prévoit la réforme ministérielle. Sélectionner des études favorables à une opinion est une chose ; ajouter des références pour soutenir scientifiquement un discours alors que ces références n'y sont pas favorables en est une autre.

Trois limites à cette synthèse doivent être soulignées. La première et la plus conséquente est en lien direct avec la qualité attendue de nos jours d'une synthèse. Cette qualité requiert l'analyse exhaustive et systématique de la littérature, l'évaluation selon des critères prédéfinis des études retenues et l'agrégation de leurs résultats en suivant des méthodes statistiques pertinentes et détaillées dans le cas d'études quantitatives. Ces dernières emploient des procédures complexes qui doivent être précisément étudiées pour en déceler les points forts comme les points faibles. Tous ces éléments sont explicités dans un protocole publié en amont qui cadre précisément le champ d'investigation et les méthodes employées. Ce travail exigeant suppose une collaboration intense de chercheurs de domaines différents avec des moyens financiers conséquents[4]. Cette synthèse ne pouvait pas satisfaire à ces critères exigeants. La seconde est inhérente au caractère prescriptif de certaines des conclusions

[1] Duru-Bellat (2001, p.225) comme Dupriez (2000, p.89) font partie de ceux qui regrettent cet état de chose.

[2] NOR : MENE2407076N _ Note de service du 15-3-2024
https://www.education.gouv.fr/bo/2024/Special2/MENE2407076N

[3] Il s'agit de la seule note de cadrage publiée par IDEE https://www.idee-education.fr/

[4] Le What Works Clearinghouse aux USA ou Education Endowment Foundation en Angleterre sont de bons exemples à consulter.

utilisées par cette synthèse, ce caractère étant lui-même consubstantiel à l'*evidence based education* puisque les études qui s'en réclament ont comme vocation d'apporter des réponses pragmatiques aux décideurs. Et comme cela est écrit à plusieurs reprises sur le site internet d'Education Endouwment Foundation (EEF), les résultats d'études antérieures décrivent « ce qui s'est passé », et ne peuvent prétendre prédire « ce qui va se passer ». Il convient donc d'en circonscrire la portée, ce qui impose aux éléments d'information présentés ici d'être contextualisés. Enfin, on aura beau jeu également de souligner qu'aucune des conclusions reprises ici ne sont issues de données françaises. Elles peuvent néanmoins alerter les chercheurs de ce pays sur les questions qu'il conviendrait de poser lors d'éventuelles études sur l'impact des regroupements par niveau en France.

Cela fait plus d'un siècle que les chercheurs et les politiques en charge de l'organisation de l'enseignement discutent de la question qui nous a préoccupée tout au long de cette synthèse. Mais les choses ont changé depuis la première méta-analyse de Kulik publiée en 1982. Si des questions pouvaient alors se poser sur l'efficacité (ou non) du regroupement des élèves par niveau sur les résultats académiques, elles ont depuis quelques années trouvé une réponse : aucun impact clairement positif sur les élèves n'a été démontré, si ce n'est une probable amélioration des résultats des élèves de niveau élevé quand les élèves sont regroupés par niveau. Faut-il alors poursuivre les recherches en ce sens ? Si les résultats de l'étude EEF en cours et dont nous avons évoqué l'existence (Hodgen, 2019) confirment cette conclusion, peut-être sera-t-il temps de flécher nos ressources vers d'autres questions. Il reste encore des interrogations. Elles concernent par exemple les effets combinés des regroupements par niveaux en mathématiques et en littéracie (on devrait alors se pencher sur les élèves scolarisés dans les groupes de niveau faible dans ces deux disciplines) ; le contenu des enseignements dans les différents groupes de niveau ; les interactions entre la constitution de groupes de niveau et d'autres pratiques (comme les groupes de soutien en dehors de la classe ou le tutorat). Mais peut-être que les réponses peuvent être déduites de ce que l'on sait déjà.

On aura également compris que ce sujet est sensible, qui met en lumière les articulations complexes qui lient l'apprentissage au contexte sociale, culturel et économique, car regrouper ensemble des élèves de niveau faible mais aussi socioéconomiquement plus défavorisés que les autres et souvent issus de minorité ethnique peut être lourd de conséquences. Quand les autorités publiques en charge de l'organisation d'un système scolaire en fixent les règles, elles se doivent d'expliciter tous les objectifs poursuivis qui ne peuvent se limiter aux compétences académiques des élèves ; et la question du regroupement par niveau ne peut se discuter sans faire référence à ce cadre social voire politique.