Education Endowment Foundation

L'enseignement fondé sur des preuves au Royaume-Uni

Nathalie ROQUES

octobre 2022

Table des matières

Résumé

Ce document propose une analyse des textes publiés par Education Endowment Foundation (EEF) sur son site internet.

Dans un premier temps, les synthèses réalisées par EEF à partir d'études publiées ces 40 dernières années sont présentées : les méthodes mises en œuvre par les méta-analyses rassemblées sous la bannière du Toolkit (« boite à outils ») sont explicitées et certains de leurs résultats sont proposés. Puis les documents experts comme les examens de synthèses (à destination des chercheurs) et les rapports d'orientation (à destination des enseignants) sont analysés.

Dans un deuxième temps, les expériences conduites par EEF sur le territoire britannique dans le but d'évaluer l'effet de certaines interventions sur les apprentissages des élèves sont passées en revue : leur procédures sont expliquées et une présentation générale de leurs résultats est proposée.

Un troisième chapitre est consacré aux conclusions que l'on peut tirer de l'ensemble de ces études en ce qui concerne l'enseignement et l'apprentissage des mathématiques.

Enfin, un bilan global et quelques commentaires sur les procédures mais aussi les missions que s'est fixé Education Endowment Foundation sont esquissés.

Avant-propos

C'est en 2011 que le Sutton Trust (une Charity britannique, c'est-à-dire une organisation de bienfaisance comme il en existe beaucoup au Royaume-Uni) lance un nouvel organisme baptisé Education Endowment Foundation (EEF). Une subvention de 125 millions de livres est versée par l'état britannique qui permet alors à EEF de débuter son action. L'objectif annoncé de cette nouvelle structure est de favoriser les apprentissages des élèves et tout particulièrement les plus défavorisés d'entre eux, en identifiant et soutenant les interventions les plus efficaces.

Nous sommes aujourd'hui en 2022, et 11 ans après sa création de nombreux projets ont été menés par Education Endowment Foundation. Deux axes complémentaires peuvent être distingués : le premier concerne la rédaction de recommandations pragmatiques fondées sur des études scientifiques dont les résultats sont publiés dans des revues à comité de lecture internationales. Le second consiste à réaliser des études par comparaison de groupes de grande ampleur sur le territoire britannique pour identifier les actions les plus efficaces.

Ce document va suivre cette partition : le premier chapitre s'attachera à présenter les synthèses rédigées et publiées par l'EEF en direction de publics divers ; ces textes sont le fruit d'une analyse de la littérature scientifique décrivant les résultats d'études quantitatives conduites dans le monde ces 50 dernières années. Le second chapitre sera consacré aux expériences menées par EEF dans les établissements scolaires du Royaume-Uni depuis 10 ans. Dans un troisième temps, les documents et études qui concernent l'enseignement des mathématiques seront analysés ; enfin le dernier chapitre proposera quelques commentaires sur les informations publiées par EEF sur son site internet.

Les mots écrits en **gras** à leur première apparition dans le texte sont définis dans le glossaire (annexe 9).

Les noms des liens actionnables sur le site internet d'EEF sont soulignés.

Pour simplifier la lecture, les documents rédigés et publiés par EEF sont repérés dans le texte et dans la liste des références EEF publiée en annexe 1 par une lettre attribuée par ordre alphabétique en fonction de leur ordre d'apparition dans le texte (le document A est donc le premier document auquel il sera fait références dans ce texte, le document B est le deuxième, et ainsi de suite).

Chapitre 1. Les synthèses publiées par EEF

Les textes dont il sera question ci-dessous ont tous été téléchargés ou copiés à partir du site internet www.educationendowmentfoundation.org. Ce site internet a été totalement refondu en septembre 2021, et, comme cela le sera indiqué par la suite, certains des textes dont il sera question ici ne sont plus actuellement en ligne.

Ces documents sont des fichiers pdf ou des textes non téléchargeables (encore dénommés ci-dessous « pages internet »). Ils vont être répartis en deux catégories.

La première catégorie rassemble les synthèses thématiques publiées dans le Toolkit (que l'on peut traduire par boîte à outil, la dénomination originale a été conservée tout au long de ce texte) dont la vocation est de fournir des informations concises fondées sur des expériences menées sur le terrain dont les résultats ont été publiés dans des revues scientifiques. Ces expériences ont comme objectif d'évaluer l'effet d'une **intervention** sur les apprentissages des élèves en menant des **études par comparaison de groupes**. Les synthèses sont rédigées de manière à être comprises par un large public et portent sur des thèmes la plupart du temps non disciplinaires. Elles se présentent sous la forme de pages internet (sujettes à modifications) et non de fichiers au format pdf[a]. Dans cette catégorie, on trouvera également des guides expliquant succinctement les règles qui encadrent la rédaction des textes précédemment évoqués ainsi que des **protocoles** détaillés, qui tous sont téléchargeables au format pdf[b].

La seconde catégorie rassemble des documents plus fournis, présentés sous la forme classique de rapports tous téléchargeables au format pdf. On distinguera ici les **examens des preuves** (rédigés en direction d'un public d'experts) d'une part, et les **rapports d'orientation** qui regroupent des recommandations pragmatiques destinées aux enseignants d'autre part, les premiers servant de base scientifique aux seconds. Certains des thèmes abordés sont disciplinaires (dans ce cas ils concernent les mathématiques, la lecture ou les sciences) ; d'autres concernent des questions d'ordre pédagogique transdisciplinaires ou des questions relatives à l'organisation de la scolarité.

La genèse et l'organisation de l'ensemble de ces textes ont été synthétisés ci-dessous (tableau 1).

Les synthèses du Toolkit

On va d'abord faire le point sur les protocoles publiés qui détaillent les procédures permettant de mener à bien la synthèse d'études quantitatives colligées pour chacun des thèmes du Toolkit. Dans un second temps, les informations publiées en direction du grand public sont

[a] En tous cas, plus aujourd'hui, car on notera que la version antérieure du site internet permettait de télécharger des fichiers pdf.

[b] Ces protocoles sont toujours téléchargeables sur internet (à partir d'un moteur de recherche), mais il n'est plus possible de les trouver sur le nouveau site EEF.

présentées et les trois indicateurs (coût, effet, niveau de confiance) associés à chacun des thèmes sont expliquées.

Documents associés au Toolkit	**Examens des preuves et rapports d'orientation**
1. Protocoles explicitant les modalités suivies pour la rédaction des pages internet du Toolkit. 2. Guides résumant ces modalités. 3. Pages internet du Toolkit proposant des synthèses thématiques transdisciplinaires en direction du grand public fondées sur une analyse rigoureuse de la littérature scientifique.	1. Rapports publiant les résultats d'examens des preuves qui sont des synthèses quantitatives. 2. Rapports d'orientation basés sur les documents précédents, proposant des recommandations thématiques (parfois disciplinaires), en direction de praticiens.
Sur le site internet Education Evidence puis Teaching and Learning Toolkit[a].	**Sur le site internet** • Education Evidence puis Evidence review pour les premiers • Guidance for Teachers pour les seconds.

Tableau 1 : organisation des synthèses

1. Les protocoles

Un premier protocole (document A) a été publié plutôt tardivement, en 2018, soit 7 ans après le lancement du site internet et bien après la publication des premiers textes du Toolkit. Un an après, un second protocole (document B) a été rédigé, qui a comme objectif de présenter la nouvelle procédure encadrant la rédaction des textes du nouveau Toolkit. Ces deux documents sont présentés ci-dessous (le premier protocole servant de base au second, il est important d'en comprendre les points principaux).

Le premier protocole présente la liste des thèmes (*strands*) identifiés par EEF comme dignes d'intérêt (voir tableau 2 pour quelques exemples et l'annexe 2 pour la liste complète). Cette liste a subi des modifications dans le temps : des thèmes ont été supprimés, d'autres ont fusionné, d'autres encore ont été ajoutés. Son évolution ne semble pas suivre de règles précises, et les thèmes visibles sur internet en avril 2022 (au nombre de 26) ne correspondaient pas à ceux publiés dans le dernier protocole publié en 2019[b]. Ces thèmes sont des caractéristiques pédagogiques (comme l'enseignement collaboratif ou bien l'apprentissage soutenu par les pairs), mais aussi des éléments associés à l'organisation du temps scolaires (comme les devoirs, la réduction du nombre d'élèves par classe, le port de l'uniforme) ou au temps périscolaire (comme l'engagement des parents, les classes d'été). Sur les 26 thèmes déployés en avril 2022, 4 seulement sont disciplinaires : Activités artistiques (*Arts participation*), Phonétique (*Phonics*), Activités physiques (*Physical activity*) et Activités de compréhension en lecture (*Reading comprehension activities*).

[a] Un ensemble de pages dédiées à l'enseignement en maternelle est également publié sur le site internet (Education Evidence puis Early years Toolkit) mais n'a pas été analysé dans ce document.

[b] On notera également que dans le texte de ce protocole il est mentionné que 34 thèmes sont retenus, mais seuls 33 sont recensés dans l'annexe B qui accompagne le texte principal.

Version originale	**Traduction**
• *Behaviour intervention*	• Améliorer les comportements (diminuer la violence, le harcèlement, …)
• *Collaborative learning approaches*	• Apprentissage collaboratif (travail en groupe)
• *Extending school time*	• Allonger la durée d'enseignement (année, journée, …)
• *Feedback*	• Information sur le niveau d'acquisition des élèves
• *Metacognition and self regulation*	• Compétences métacognitives et auto-régulation ou encore « apprendre à apprendre »
• *Reducing class size*	• Réduction des effectifs de la classe
• *Summer school*	• Enseignement durant les vacances

Tableau 2 : 7 thèmes en version originale et leur traduction (exemples)

La suite de ce premier protocole vise à décrire les procédures suivies par EEF pour rédiger ses conclusions pour chacun des thèmes (voir le schéma 1, colonne gauche) : (**a**) des **méta-analyses** sont identifiées et sélectionnées[a], ou en leur absence des **synthèses systématiques**, ou en leur absence des **études primaires**. Quand plusieurs méta-analyses traitant de la même question sont sélectionnées, une méta-analyse secondaire (on parle aussi de **méta-méta analyse**) est conduite et (**b**) une **taille d'effet** globale est calculée, traduite en nombre de mois dont un élève moyen a progressé s'il a bénéficié d'une intervention entrant dans le cadre du thème, par rapport à un même élève moyen qui n'a pas bénéficié de l'intervention (on parlera dans la suite de ce texte plus simplement de « nombre de mois de progrès »).

Le second protocole, publié seulement un an après le premier, décrit la nouvelle procédure qui a permis de réactualiser les conclusions publiées (voir le schéma 1, colonne droite). En effet, ce ne sont plus les tailles d'effet des méta-analyses qui sont utilisées pour calculer la taille d'effet globale, mais les tailles d'effet des études primaires rassemblées par chaque méta-analyse. Cette démarche est plus rigoureuse et permet, entre autres, d'éviter que le résultat d'une même étude, incluses dans plusieurs méta-analyses différentes, ne soit comptabilisé plusieurs fois (*overlapping)*. Les méta-analyses sélectionnées précédemment (selon les procédures décrites dans le protocole en 2018) fournissent maintenant la matière première, c'est-à-dire les études primaires qu'elles ont rassemblées (**c**). Ces études primaires sont alors examinées selon une procédure définie *a priori* avant d'être sélectionnées (ou rejetées) et une nouvelle méta-analyse est conduite par EEF qui aboutit finalement (**d**) au calcul d'une taille d'effet globale. L'essentiel de ce protocole détaille les procédures suivies pour sélectionner les études mais aussi leurs résultats les plus pertinents et qui seront exploités pour calculer cette taille d'effet globale[b]. Ce n'est donc plus une méta-analyse secondaire (ou méta-méta analyse), mais une méta-analyse primaire qui est réalisée.

[a] Certaines peuvent être rejetées pour défaut méthodologique par exemple.

[b] La plupart des études publient plusieurs résultats, par exemple en lecture, ou en maths ou pour des groupes d'élèves différents, ou plus ou moins longtemps après la fin de l'expérience.

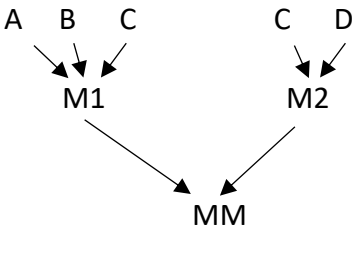
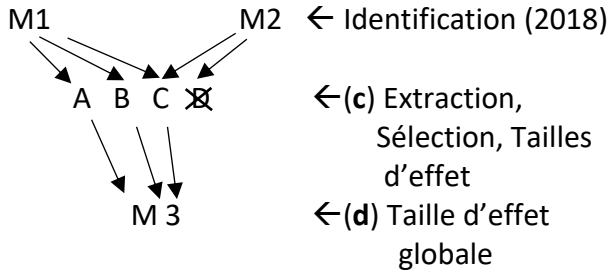
Protocole 2018	Protocole 2019
1. Définition des thèmes transversaux 2. Identification et sélection de méta-analyses 3. Méta-analyse secondaire (méta-méta analyse)  L'étude C est comptabilisée 2 fois.	1. Analyse des études primaires incluses dans les méta-analyses identifiées en 2018 2. Méta-analyse primaire  L'étude C est comptabilisée 1 seule fois. L'étude D a été rejetée.
A, B, C et D sont des études primaires. M : méta-analyse. MM : méta-méta analyse	

Schéma 1 : procédures des protocoles 2018 et 2019

En résumé, on retiendra que ces deux protocoles décrivent les procédures qui permettent de sélectionner (ou non) les données, et expliquent les calculs qui, à partir de ces données, vont permettre de conclure. Les formules permettant de calculer les tailles d'effet sont présentées en annexe 5.

2. Les informations

Cliquer sur Education evidence puis Learning Toolkit permet d'accéder à une première page présentant la liste des thèmes retenus accompagnés pour chacun d'entre eux de trois indicateurs : le cout, l'effet mesuré en nombre de mois de progrès comme nous l'avons signalé auparavant et le niveau de confiance à accorder à l'évaluation de l'effet. Le cout et le niveau de confiance sont des notes entières qui peuvent aller de 1 à 5 (voir figure 1). Ces indicateurs sont associés à un thème mais on notera ici qu'ils concernent les interventions entrant dans le cadre du thème considéré. Ces indicateurs se présentent également sous la forme de filtres permettant de sélectionner des thèmes en fonction de critères choisis par l'internaute. En cliquant sur un thème, on accède à un second niveau spécifique au thème sélectionné, principalement constitué d'un texte explicatif qui suit toujours le même canevas. Enfin, en cliquant sur le lien Technical Appendix, un troisième niveau est proposé et présente un résumé de toutes les études sélectionnées pour le thème donné.

C'est l'ensemble de toutes ces informations qui constitue le Toolkit. Avant de présenter les informations publiées sur ces deux derniers niveaux, nous allons décrire les 3 indicateurs publiés dès la première page du Toolkit. Ces indicateurs constituent l'armature sur laquelle l'internaute est invité à fonder son jugement pour chacun des thèmes proposés. Les informations qui suivent sont publiées dans un guide consultable à partir de la première page du Toolkit (document C) ; le protocole de 2019 (document B) ne fait aucune allusion à ces trois

indicateurs et le protocole de 2018 (document A) en donne des définitions qui ont depuis subies des modifications.

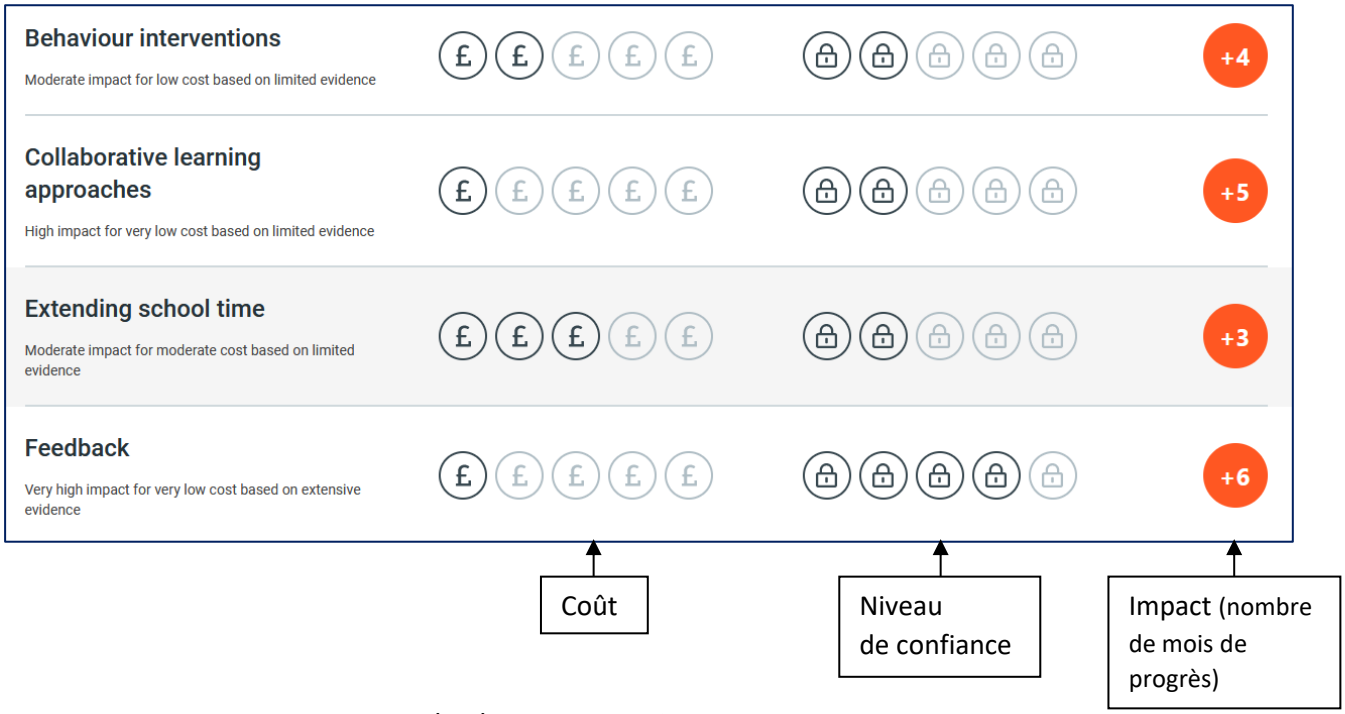


Figure 1 : Capture d'écran (25/08/2022) de 4 thèmes présentés en première page du Toolkit[a]

Les trois indicateurs

➔ Le cout

Le coût moyen de toutes les interventions inclues dans un thème est calculé et présenté sous la forme d'un indicateur qui peut aller de 1 à 5. Plus le cout est élevé, plus l'indicateur est élevé (tableau 2). Par exemple, les interventions qui coutent moins de 2 000 livres (2 350 €) par an et par classe de 25 élèves ou moins de 80 livres (soit environ 95 €) par an et par élève sont classées comme très peu couteuses (c'est le cas pour le thème apprentissage collaboratif, voir la figure 1). La répartition des thèmes par niveau de coût est présentée dans le tableau 3. On remarquera que 80 % des interventions ont un coût faible ou très faible.

Échelon	Niveau (qualificatif)	Tranches (en livres anglaises) par élève et par an
1	Très faible	Inférieur à 80
2	Faible	Entre 80 et 220
3	Modéré	Entre 220 et 700
4	Élevé	Entre 700 et 1200
5	Très élevé	Supérieur à 1200

Tableau 2 : Échelle de couts utilisé dans le Toolkit.

[a] On pourra remarquer que les résultats du thème *Feedback* qui sont donnés comme exemple dans le document C sont différents de ceux qui sont donnés ici, reflétant en cela l'évolution permanente des indicateurs (voir discussion plus loin).

Couts	Nombre de thèmes
1 (très faible)	16
2 (faible)	3
3 (modéré)	5
4 (élevé)	0
5 (très élevé)	2

Tableau 3 : répartition des thèmes en fonction du coût

➔ L'effet

Comme nous l'avons déjà mentionné, l'effet (ou l'impact) est donné en nombre de mois dont un élève moyen du groupe traitement a progressé par rapport à un élève moyen du groupe contrôle[a]. Cet effet est mesuré en calculant dans un premier temps la taille d'effet globale de toutes les études sélectionnées par la méta-analyse puis est transformée dans un second temps en un indicateur plus simple à comprendre pour le grand public. Dans le tableau 4 ci-dessous, le nombre de mois dont un élève a progressé est donné en fonction de la fourchette dans laquelle se trouve la taille d'effet globale calculée[b] et la répartition des thèmes par nombre de mois de progrès évalués est également présentée.

Mois	0	1	2	3	4	5	6	7	8	9	10	11	12
Impact	Très faible	Faible		Modéré		Élevé	Très élevé						
Minimum*	-0,05	0,06	0,10	0,19	0,27	0,36	0,45	0,53	0,62	0,70	0,79	0,88	0,96
Maximum**	0,05	0,09	0,18	0,26	0,35	0,44	0,52	0,61	0,69	0,78	0,87	0,95	1,00
Nombre de thèmes ***	1	2	3	3	6	6	3	1	0	0	0	0	0

Tableau 4 : Effet des interventions sur les apprentissages des élèves.

* : taille d'effet minimum ; ** : taille d'effet maximum ; *** : Le nombre de mois de progrès pour le thème Redoublement est négatif (- 3 mois).

Par exemple, un élève qui a participé à une intervention proposant un apprentissage collaboratif a progressé de 5 mois[c] si on le compare à un élève qui n'a pas été soumis à un tel apprentissage (figure 1). On notera que le nombre de mois de progrès maximum affiché par un thème est de 7 mois[d] (ce qui correspond à une taille d'effet maximale de 0,61) et que presque les 2/3 des thèmes (16 sur 26) présentent un nombre de progrès de 4 mois ou plus.

[a] Et concerne, comme nous l'avons déjà mentionné, toutes les interventions entrant dans le cadre du thème.

[b] Il s'agit du tableau de conversion publié dans le document C. Le document A en proposait une version légèrement différente pour les deux premiers niveaux (voir annexe 4).

[c] La taille d'effet globale calculée et publiée dans le Technical Appendix de ce thème est égale à 0,448 (qui semble avoir été arrondi à 0,44 et non 0,45 puisque dans ce cas un nombre de mois de progrès de 6 mois aurait été proposé).

[d] Il s'agit du thème *Metacognition and self regulation.*

➔ Le niveau de confiance

Plus la confiance que l'on peut accorder à l'évaluation de l'effet est grande, plus le nombre de cadenas (qui représentent ce niveau de confiance) est élevé. Depuis septembre 2021 et l'application des nouvelles procédures, ce niveau est basé principalement sur le nombre d'études primaires sélectionnées pour chaque méta-analyse (tableau 5). Mais ce n'est pas le seul critère, et le niveau de confiance (ou niveau de preuves) est abaissé dans les cas suivants :

- les études ne sont pas récentes,
- de nombreuses études ne sont pas des **essais contrôlés randomisés**,
- de nombreuses interventions ont été menées auprès des élèves par des chercheurs (et non par des enseignants),
- des études ont été menées par des organismes en lien avec l'intervention (conflit d'intérêt possible)
- les résultats sont très variés et leur hétérogénéité inexpliquée.

Par exemple, le niveau de preuve associé au thème apprentissage collaboratif est limité (figure 1). On remarquera que la moitié des thèmes affichent un niveau de confiance supérieur ou égal à 3[a] (voir tableau 5).

Niveau de confiance	Nombre d'études	Nombre de thèmes
5 (très important)	Plus de 70	1
4 (important)	Entre 45 et 69	6
3 (modéré)	Entre 25 et 44	6
2 (limité)	Entre 11 et 24	7
1 (très limité)	Moins de 11	6

Tableau 5 : Répartition des thèmes en fonction du niveau de confiance

Bilan

C'est l'ensemble de ces trois indicateurs (cout, effet et niveau de confiance) qui doit être considéré quand un thème (en fait un ensemble d'interventions) est analysé. Le tableau 6 présente les 9 thèmes qui affichent un cout de mise en œuvre très faible, un niveau de confiance supérieur ou égal à 3 et efficaces (1 mois de progrès au moins). Ils sont d'abord triés par ordre décroissant de niveaux de confiance puis par ordre décroissant de nombres de mois de progrès réalisés[b].

[a] C'est le niveau de confiance minimal exigé par EEF pour ses études primaires (voir chapitre 2).

[b] Le parti pris ici est de d'accorder au niveau de confiance une importance supérieure au nombre de mois de progrès. On remarquera que EEF place cet indicateur après le coût et avant l'effet.

Thème (français)	Thème (anglais)	Confiance	Effet
Phonique	*Phonics*	5	5 mois
Métacognition	*Metacognition*	4	7 mois
Évaluation	*Feedback*	4	6 mois
Langage oral	*Oral langage interventions*	4	6 mois
Compréhension de la lecture	*Reading comprehension strategies*	4	6 mois
Tutorat par les pairs	*Peer tutoring*	4	5 mois
Engagement des parents	*Parental engagement*	4	4 mois
Activités artistiques	*Arts participation*	3	3 mois
Activités physiques	*Physical activity*	3	1 mois

Tableau 6 : Liste des thèmes de coût très faible, efficaces avec un bon niveau de confiance

Le deuxième niveau

Quand on clique sur un thème à partir de la liste évoquée ci-dessus, on accède à une seconde page qui donne quelques détails sur le thème sélectionné, par exemple en distinguant les résultats obtenus en lecture ou en mathématiques, ou en précisant l'impact sur les élèves socialement désavantagés. Le niveau de preuve est également expliqué et les facteurs qui limitent la confiance à accorder à l'effet calculé sont détaillés (un seul thème présente des résultats associés à un niveau de confiance maximal). On trouve en bas de page un rappel des trois indicateurs, accompagnés de la date de mise à jour de la page[a]. A chaque fois, les lecteurs sont invités à considérer les résultats avec précaution, notamment en prenant en compte leur environnement de travail et en adaptant les recommandations à leur situation propre (cette alerte apparait à de nombreuses reprises sur le site internet et les documents publiés). Il est clairement indiqué que les analyses et les résultats montrent que certaines interventions « ont marché », et non pas qu'elles « vont marcher ». Pour chaque thème, une même barre de navigation horizontale (voir figure 2) permet d'accéder plus rapidement à certaines parties de la page. Seul le dernier lien Technical Apprendix ouvre une nouvelle page : c'est le troisième niveau d'information.

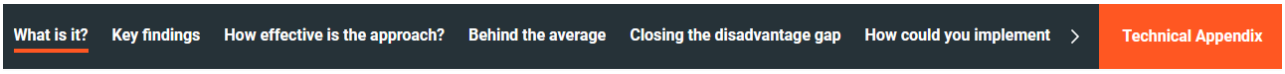


Figure 2 : Capture d'écran au 25/08/2022 de la barre de navigation de la page du second niveau

Le troisième niveau

Pour chacun des thèmes, une troisième page est ouverte quand on clique sur le lien Technical appendix. Il s'agit d'un diagramme en forêt qui présente dans une première colonne la liste

[a] C'est l'une des rares dates publiées sur le site.

des études primaires sélectionnées, et dans une seconde colonne les résultats suivants[a] : la taille d'effet et son intervalle de confiance à 95%, le poids en pourcentage de l'étude et l'erreur standard de la taille d'effet. Des résultats globaux sont affichés en bas de page : taille d'effet globale, intervalle de confiance à 95 % ; ainsi que des indicateurs d'hétérogénéité : variance inter-étude Tau², Q et I². Les études peuvent être filtrées en fonction de la nature des tests (tous les tests, ou bien tests en maths ou bien tests en lecture) et des niveaux scolaires (tous les niveaux, primaire ou secondaire). Une fois les études filtrées, les résultats globaux ne sont plus affichés[b]. Enfin les études peuvent être triées par ordre croissant ou décroissant en fonction du nom des auteurs[c], de la date de publication, de la taille d'effet, de l'intervalle de confiance ou du poids. On ne peut pas copier/coller ce texte. Si on clique sur une étude en particulier, on ouvre une fenêtre qui donne à nouveau ces résultats, mais aussi les références complètes et le résumé de l'article tel qu'on peut le consulter sur internet (voir l'analyse du thème apprentissages collaboratifs dans le domaine des mathématiques pour plus de détails).

Même s'il est possible d'extraire du Toolkit des informations sur l'efficacité d'une méthode pédagogique en lecture ou en mathématiques, l'objectif principal est ici de proposer une analyse globale et donc transdisciplinaire. Il en va autrement pour les recommandations dont il sera question ci-dessous.

Examens des preuves et rapports d'orientation

Comme on l'a déjà souligné au début de ce texte, des documents plus étoffés et détaillés destinés aux enseignants, aux formateurs et au personnel de direction ont été rédigés, et peuvent être considérés comme des documents experts. Deux types de documents sont ainsi proposés sur des sujets reconnus par EEF comme pertinents : des examens des preuves[d] (*evidence review*) qui suivent les principes des synthèses systématiques d'une part, et des rapports d'orientation (*guidance report*) qui proposent des recommandations concrètes s'appuyant sur les conclusions des premiers, d'autre part. Ces documents sont tous téléchargeables au format pdf.

1. Examens des preuves

En cliquant sur Evidence Review, on ouvre une page recensant tous les examens des preuves répartis dans 8 catégories. Ces documents sont rédigés par des experts reconnus[e] qui suivent les grandes lignes suivantes : (1) des critères sont définis *a priori* qui permettent de procéder

[a] Pour plus de détails sur ces résultats, consulter *Mesurer l'effet d'un traitement. Les méta-analyses en sciences de l'éducation* (ROQUES, 2021, www.mathadoc.fr)

[b] Les intitulés de ces résultats apparaissent toujours, mais des valeurs nulles sont affichées, à l'exception du thème *Collaborative learning approaches* quand les tests maths sont sélectionnés (voir plus loin).

[c] En fait, du premier auteur, seul nom communiqué à ce niveau.

[d] Parfois dénommés rapports d'évaluation (*evaluation report*) dans certains documents quand ils font référence à ces textes experts.

[e] Ce sont le plus souvent des universitaires.

à une sélection systématique des études quantitatives concernées, (2) une analyse qualitative de ces études est menée et dans certains cas (3) une analyse quantitative de leurs résultats (des tailles d'effet globales pouvant alors être calculées). Certaines synthèses sélectionnent des études primaires (c'est le cas par exemple pour le rapport portant sur la rétroaction (*feedback*), document D), d'autres des méta-analyses (c'est le cas pour le rapport portant sur les interventions en mathématiques, document E dont il sera question plus tard). L'objectif poursuivi par EEF est de faire le point, pour un sujet donné, sur la recherche internationale en considérant pour l'essentiel des études quantitatives basées sur des comparaisons de groupes afin de repérer les interventions les plus efficaces. Ces textes n'ont pas vocation à proposer des recommandations mais de répondre de façon rigoureuse à une (ou plusieurs) question de recherche.

Aucun document n'a été trouvé qui expliciterait les procédures suivies par les experts pour réaliser ces examens des preuves. Pour les examens les plus récents, un protocole spécifique a été rédigé : c'est par exemple le cas pour l'examen portant sur la rétroaction qui est accompagné de son protocole (document D) et la rédaction de ce dernier semble être recommandé de nos jours par EEF. Pour les autres examens, c'est à l'intérieur de chacun de ces documents que l'on trouvera des informations sur leur genèse. Deux modèles (des fichiers word) peuvent être téléchargés par les internautes : un modèle pour la rédaction de protocole (document F), et un modèle permettant de rédiger le rapport de l'examen (document G).

En avril 2022, 36 rapports étaient publiés sur le site internet d'EEF réparties dans 8 catégories (tableau 7).

Catégories	Lire, écrire	Maths	Sciences	Évaluations	Apprendre	Compétences de la vie	Leadership, formation	Covid 19
Nombre de rapports	9	2	2	3	6	4	10	2

Tableau 7 : Nombre d'examens des preuves par catégories

C'est à partir de ces examens des preuves que des rapports d'orientation sont alors rédigés en direction des praticiens.

2. Rapports d'orientation

Basés sur les documents précédemment décris et répartis en 7 catégories (aucun document concernant l'épidémie du Covid 19 n'a été publié), 17 rapports d'orientation proposent des recommandations pragmatiques en direction des professionnels de l'enseignement et des personnels des établissements scolaires (tableau 8). Le rapport complet est le plus souvent accompagné d'un résumé voire d'autres documents annexes (poster, auto-évaluation, …). Ces rapports ne sont pas datés (les seules dates de publication disponibles figurent sur le site internet) et sont parfois spécifiques à un niveau d'étude. Dans le système britannique comme en France les élèves changent de niveau chaque année, et on distingue également l'école

primaire, le collège et le lycée. Mais il existe également quatre autres niveaux, les *Key stages* (voir annexe 3).

Là encore aucun protocole général n'a été rédigé qui expliciterait les procédures suivies pour la réalisation de ces rapports d'orientation, et c'est à l'intérieur de chacun de ces documents que l'on trouvera quelques informations à ce sujet.

Catégories	Lire, écrire	Maths	Sciences	Évaluations	Apprendre	Compétences de la vie	Leadership, formation
Nombre de rapports	4	2	1	1	4	1	4

Tableau 8 : Nombre de rapports d'orientation par catégories

Chapitre 2. Les études conduites par EEF

C'est très certainement la partie la plus importante des travaux d'EEF, tant au niveau de la quantité d'informations que l'on peut en retirer que de l'importance que l'on peut leur accorder. De très nombreuses études primaires ont été menées en l'espace de 10 ans sur le territoire britannique par cet organisme (elles seront dénommées « études EEF » par la suite) afin de déterminer l'effet d'une intervention donnée sur les apprentissages des élèves. Ces études par comparaison de groupes sont la plupart du temps des essais contrôlés randomisés.

Quand une intervention est identifiée comme susceptible d'améliorer les apprentissages des élèves, les étapes suivantes seront suivies qui correspondent à différents types d'études :

Étape 1. Une étude pilote (*pilot study*) est conduite sur un petit nombre d'établissements pour explorer son impact. La recherche est de type qualitatif et vise à préciser la faisabilité d'une intervention à plus grande échelle. Si cette étape est concluante, alors EEF passe à l'étape suivante.

Étape 2. Une étude d'efficacité (*efficacy trial*) est menée sur un plus grand nombre d'établissements (généralement une cinquantaine) par les développeurs de l'intervention dans des conditions idéales. Une évaluation quantitative de l'effet de l'intervention sur les apprentissages des élèves est réalisée (une taille d'effet et un nombre de mois de progrès sont notamment calculés). La mise en œuvre de l'intervention est également évaluée et un cout indicatif de l'élargissement de cette intervention à l'ensemble du territoire est calculé. Si cette étude aboutit au calcul d'un effet négatif ou très faible (inférieur ou égal à 1 mois de progrès), le projet est interrompu ; dans le cas contraire, EEF passe à l'étape suivante.

Étape 3. Une étude d'efficacité en conditions réelles (*effectiveness trial*) est menée dans des conditions ordinaires de mise en œuvre : les développeurs ne sont plus directement impliqués dans les processus d'apprentissage et le nombre d'établissements est généralement d'une centaine, répartis dans au moins 3 régions du Royaume-Uni. Une évaluation quantitative de l'effet et une évaluation de la mise en œuvre de l'intervention sont conduites dans les mêmes conditions que pour une étude d'efficacité et le coût de l'intervention est calculé.

Étape 4. Certaines interventions ayant franchies les étapes précédentes sont évaluées par une étude à plus grande échelle (*scale-up*) sur un large territoire.

Pour chaque étude, un rapport d'évaluation (*evaluation report*) est publié. Il inclue une évaluation de l'effet (*impact evaluation*), c'est-à-dire un ensemble de résultats issus d'une analyse statistique permettant de mesurer l'effet de l'intervention sur les apprentissages des élèves, et une évaluation de la mise en œuvre et du processus (*implementation and process evaluation*) dont l'objectif est de comprendre comment et pourquoi une intervention a été efficace (ou non). Un protocole d'évaluation (*protocol report*) est également publié, et pour les études les plus récentes (et à l'exception des études pilote qui ne sont pas concernées), un plan d'analyse statistique (*statistical plan*). Ces deux documents explicitent les méthodes

définies *a priori* et sont publiés avant que les résultats ne soient connus. Il est important de souligner ici que les évaluations des interventions sont conduites par des chercheurs indépendants, c'est-à-dire qu'ils ne sont pas en charge de la mise en œuvre de l'intervention évaluée.

Plusieurs documents définissent les procédures générales mises en œuvre par EEF pour réaliser leurs études comme pour rédiger les textes évoqués ci-dessus ; pour certains de ces documents, un résumé est proposé en annexe 6. Des modèles sont également disponibles : ce sont des fichiers word qui peuvent être utilisés par les évaluateurs et qui parfois donnent des informations complémentaires comme on le verra plus loin. Une présentation des éléments saillants de ces procédures est proposée ci-dessous puis une description globale des études est présentée.

Les procédures générales

Tous les textes permettant de comprendre les procédures mises en œuvre par EEF pour mener leurs études sur le terrain sont à télécharger en cliquant sur Project and evaluation, Evaluation puis Evaluation guidance and resources.

De la même façon que pour les thèmes du Toolkit, trois indicateurs vont être calculés pour chaque intervention étudiée : son effet (en nombre de mois de progrès), le niveau de confiance que l'on peut lui associer (sur une échelle allant de 1 à 5) et le coût de l'intervention (sur une échelle allant de 1 à 5).

1. L'effet de l'intervention

Les procédures qui encadrent le calcul de la taille d'effet sont publiées dans le document H (consulter l'annexe 6 pour plus de détails). La taille d'effet calculée est le g de Hedges et cette fois, contrairement aux règles en vigueur pour les méta-analyses du Toolkit, le niveau initial des élèves est pris en compte. Les analyses doivent être menées en **intention de traiter**. Le fait que des classes (et non des élèves) soient aléatoirement distribuées dans les groupes traitement et contrôle doit également être pris en compte. Une attention particulière est portée à l'**attrition** des données et à la multiplicité des mesures qui risque d'augmenter le risque de première espèce (et donc de faux positif). Aucune information sur la transformation de la taille d'effet en nombre de mois de progrès n'est donnée dans le document H ; une grille est proposée dans le modèle permettant de rédiger le rapport d'évaluation (fichier word, document I). Cette grille reprend les mêmes seuils que ceux utilisés pour le Toolkit à l'exception des deux premiers échelons (voir annexe 4 pour plus de précisions).

La taille d'effet doit être accompagnée d'une estimation de son incertitude. On notera qu'EEF préconise aujourd'hui d'abandonner la présentation dichotomique habituelle : les tailles d'effet ne doivent plus être présentées comme soit « statistiquement significatives » soit « non statistiquement significatives ». Dans le même ordre d'idée, l'intervalle de confiance

laisse la place à l'intervalle de compatibilité[a] et la valeur-p qui accompagne la taille d'effet est considérée comme une valeur continue. En fait, toutes les décisions basées sur des seuils fixés arbitrairement (comme la valeur 0,05 traditionnellement utilisée dans ce type d'analyse pour caractériser une valeur-p calculée) n'ont pas lieu d'être. C'est dans un document publié en février 2020 que ces mises au point sont formulées (document J). Elles font suite à un premier texte publié en 2018 (document K) qui s'interrogeait sur le nombre non négligeable d'études EEF qui, suivant les procédures *ad hoc* d'un essai contrôlé randomisé et dimensionnées de façon à repérer des tailles d'effet positives et statistiquement significatives (comme il était d'usage de le dire alors), aboutissaient à des résultats positifs modestes (1 mois de progrès par exemple) avec un bon niveau de confiance (supérieur ou égal à 3) mais associés à une valeur-p supérieure à 0,05.

On terminera sur ce sujet en évoquant la différence des variances du groupe traitement et du groupe contrôle comme potentiel résultat permettant d'évaluer l'effet d'une intervention. En effet, l'objectif déclaré d'EEF est de réduire les inégalités entre les élèves de niveaux socioéconomiques différents ; cette réduction devrait en toute logique s'accompagner d'une diminution de la variance pour les élèves ayant subi l'intervention (le groupe traitement). Une étude a été menée sous l'égide d'EEF (document L) mais pour l'instant, face au manque de connaissances théoriques, les chercheurs proposent de conserver la taille d'effet comme résultat principale de l'effet d'une intervention, tout en l'accompagnant d'une comparaison de l'évolution dans le temps (avant/après l'intervention) des variances pour chacun des groupes (traitement et contrôle). On notera tout de même que sur les 16 études EEF analysées par les auteurs de ce document, à chaque fois qu'une variation statistiquement significative était identifiée il s'agissait d'une augmentation de la variance du groupe intervention après l'expérience.

2. Le niveau de confiance

Ce niveau de confiance est associé au calcul de l'effet et (comme pour les méta-analyses du Toolkit) permet de classer l'étude sur une échelle allant de 1 à 5. Quatre critères sont considérés (document M, consulter l'annexe 6 pour plus de détails) :

1. Le design de l'étude : les essais contrôlés randomisés constituent ici la référence ; en permettant une attribution aléatoire des élèves dans les groupes contrôle et traitement, ce type d'étude permet de se libérer de la question des facteurs de confusion.
2. La taille d'effet minimal détectable : c'est la taille d'effet minimale que l'étude est susceptible de détecter (cela revient à fixer une taille de l'échantillon minimale).
3. L'attrition : c'est le niveau global de la perte de données (groupe contrôle et groupe traitement confondus) qui doit être mesuré au niveau des élèves, quel que soit le niveau de l'attribution aléatoire.

[a] Seul le nom change, les calculs restent identiques.

4. Les menaces à la validité interne : 7 menaces sont examinées (facteurs de confusion, présence d'interventions concurrentes associées à l'intervention évaluée, effet Hawthorne[a], fidélité de la mise en œuvre, données manquantes, questions liées aux mesures des résultats, publication sélective des résultats).

Le niveau de confiance minimum attendu pour une étude EEF est de 3, et les résultats des études qui n'atteignent pas ce niveau de confiance doivent être considérés avec beaucoup de précautions.

3. Le cout de l'intervention

Plusieurs principes doivent être suivis qui permettent d'évaluer le cout d'une intervention et sont détaillés dans le document N (consulter l'annexe 6 pour plus de détails). On soulignera entre autres que tous les couts liés à l'intervention doivent être inclus, comme la formation des enseignant, les coûts liés au recrutement, l'achat de matériel nécessaire ; ils seront divisés en cout de démarrage (correspondant à la mise en place de l'intervention) et couts récurrents sur la base d'une durée d'implémentation de 3 années ; ils doivent être présentés de manière à tenir compte de l'inflation pour permettre une comparaison entre plusieurs interventions évaluées à des dates différentes. La grille qui permet de classer le cout d'une intervention sur une échelle de 1 à 5 n'est pas donnée dans le document N ; cette grille est proposée dans le modèle permettant de rédiger le rapport d'évaluation (fichier word, document I)[b]. On notera que le seuil qui sépare les niveaux 2 et 3 est fixé à 200 £ ici, alors que pour le Toolkit ce seuil est égal à de 220 £[c].

4. La mise en œuvre de l'intervention

Comme nous l'avons évoqué précédemment, EEF ne se contente pas d'évaluer l'effet d'une intervention (et le niveau de confiance que l'on peut accorder à cet évaluation) et son coût, mais s'attache également à évaluer sa mise en œuvre. L'objectif est alors de déterminer les contextes ou les circonstances qui peuvent influer sur l'effet que peut avoir une intervention (par exemple en précisant son effet sur des élèves de niveaux scolaires différents ou en identifiant des freins à sa mise en œuvre). On retiendra notamment que pour chaque évaluation de la mise en œuvre d'une intervention, un protocole spécifique à l'intervention doit être rédigé (les méthodes générales et les mesures doivent donc être précisées avant l'évaluation à proprement parlé) ; que la fidélité (l'intervention a-t-elle été délivrée telle que prévu ?), la conformité (les participants ont-ils reçu l'intervention tel que prévu ?) et les pratiques usuelles (ce sont les pratiques observées dans le groupe contrôle) doivent être analysées et décrites (document O). Cette évaluation doit être établie conjointement à

[a] Cet effet est aux études en sciences de l'éducation ce que l'effet placebo est aux études pharmacologiques.

[b] On la retrouve également en annexe dans les rapports d'intervention les plus récents.

[c] Le guide proposant ce seuil de 220 £ pour les méta-analyses du Toolkit a été rédigé en septembre 2021, soit 2 ans après le document H.

l'évaluation de l'effet, et ce sont ces deux analyses qui constituent le cœur du rapport d'évaluation de l'intervention.

Les études publiées

Quand on clique sur Project and evaluation puis Projects on visualise une première page qui présente l'ensemble des études sous forme de vignettes avec les informations suivantes[a] : le type d'étude (étude pilote, étude d'efficacité, …), son niveau de complétion (en cours ou achevée), le nom de l'intervention, une très courte description de l'étude, le thème de l'étude (à distinguer des thèmes définis pour le Toolkit, ils caractérisent ici l'intervention), le nombre de mois de progrès calculé et les conditions dans lesquelles ce nombre a été calculé (ce sont les meilleures conditions possibles pour les études d'efficacité et les conditions réelles pour les études d'efficacité en conditions réelles). Ni le niveau de confiance ni le cout ne sont indiqués à cette étape.

Sur les 242 études comptabilisées le 6 aout 2022, 172 étaient achevées, 56 en cours, 12 en phase de recrutement et 2 annulées. On notera que plusieurs études peuvent analyser une même intervention (celle-ci ayant fait l'objet d'une étude pilote puis d'une étude d'efficacité par exemple).

Les études peuvent être filtrées selon 7 critères, dont le niveau scolaire, le type d'étude, le thème et le sujet abordé (c'est la nature des résultats utilisés pour quantifier l'effet de l'intervention, par exemple des scores en mathématiques[b]). La répartition des études en fonction des sujets est proposée dans le tableau 9. On notera que ni le niveau de confiance, ni l'effet (c'est le nombre de mois de progrès), ni le coût ne sont des critères de sélection. La répartition des études d'efficacité en condition réelles achevées en fonction de l'effet estimé (en nombre de mois de progrès) est proposée dans le tableau 10. On remarquera que pour les trois quarts de ces études (26 études sur 33) l'effet calculé est inférieur ou égal à 1 mois et qu'aucune n'a abouti à un effet supérieur à 3 mois de progrès.

Sujets	Art	Anglais	Lecture[c]	Maths	Numératie	École	Sciences	**Total**
Nombre total	2	9	72	39	1	41	16	**180***
Nombre études achevées	1	5	58	27***	0	37	9	**138***
dont *Études d'efficacité*	0	3	40	11	0	17	3	**74**
dont *Études d'efficacité en conditions réelles*	0	1	14**	10***	0	6	2	**33**

Tableau 9 : répartition des études EEF par sujet et par type d'étude

[a] Visibles en juin 2022, certaines informations avaient disparues ou étaient modifiées fin aout 2022.

[b] Aucune définition des critères Thèmes et Sujet n'a été trouvée, il s'agit donc ici d'une interprétation personnelle. Aucune procédure concernant le codage des études (c'est-à-dire l'association d'une étude à telle ou telle catégorie) n'a été trouvée également.

[c] *Literacy* ou encore compétences associées à la lecture et la rédaction a été traduit plus simplement par lecture ici.

* : les totaux sont inférieurs à 242 et 172, certaines études n'ayant pas été associées à un sujet (par exemple les études sur l'impact de l'épidémie du covid 19). ** : par erreur, une des interventions (*Grammar for writting*) est codées comme étude d'efficacité en conditions réelles alors qu'il s'agit d'une étude d'efficacité. *** : par erreur une étude non achevée (*Research Into Teacher Training*) a été comptabilisée. Les nombres proposés ici sont les nombres corrigés.

Sujets		Anglais	Lecture	Maths	École	Sciences	**Total**
Nombre total d'études		1	14	10	6	2	**33**
Nombre de mois de progrès	**0 mois ou négatifs**	0	10	5	5	2	**22**
	1 mois	0	2	2	0	0	**4**
	2 mois	1	1	2	1	0	**5**
	3 mois	0	1	1	0	0	**2**

Tableau 10 : répartition des études d'efficacité en conditions réelles achevées selon leur effet

A partir de cette liste, on accède à un second niveau d'information en cliquant sur une étude donnée, et c'est à ce niveau que l'on trouve notamment le niveau de confiance à accorder à l'effet calculé. La répartition des études d'efficacité en conditions réelles achevées en fonction du niveau de confiance est présentée dans le tableau 11. On notera que presque 90% des études (29 sur 33) ont un niveau de confiance au moins égal à 3. Si on se limite aux études d'efficacité en condition réelles achevées avec un niveau de confiance supérieur ou égal à 3, à très faible coût et présentant un effet positif non nul, on obtient alors les six interventions présentées dans le tableau 12. Les rapports d'évaluation, protocoles et plan d'analyses sont également téléchargeables sur cette page.

Sujets		**Anglais**	**Lecture**	**Maths**	**École**	**Sciences**	**Total**
Nombre total d'études		1	14	10	6	2	**33**
Niveau de confiance	**1**	0	0	0	1	0	**1**
	2	0	0	2	1	0	**3**
	3	0	5	3	2	1	**11**
	4	1	3	4	1	1	**10**
	5	0	6	1	1	0	**8**

Tableau 11 : niveau de confiance des études d'efficacité en conditions réelles achevées

Interventions	**Sujet**	**Effet**	**Niveau de confiance**
Nuffield early langage intervention	Lecture	3 mois	5
Embedding formative assessment	École	2 mois	5
1st class@number	Maths	2 mois	4
Mathematical reasoning	Maths	1 mois	4
IPEELL: using self-regulation to improve writing	Anglais	2 mois	4
Success for All	Lecture	1 mois	3

Tableau 12 : Interventions de coût très faible, efficaces avec un niveau de confiance supérieur ou égal à 3

Les études conduites par EEF depuis 2011 ont fait l'objet de plusieurs méta-analyses (document P) dans le but de répondre aux trois questions de recherche suivantes :

1. Les interventions ont-elles conduit à une amélioration des compétences et connaissances en anglais et en mathématiques des élèves éligibles au *Free School Meals* (FSM)[a] ?
2. Quelles sont les catégories d'études ou d'interventions associées à une amélioration des compétences et connaissances en anglais et en mathématiques toujours pour ces élèves en particulier ?
3. Les élèves éligibles au FSM ont-ils plus ou moins progressé que les élèves non éligibles au FSM ?

Contrairement aux méta-analyses du Toolkit, les analyses ont d'emblée été séparées en fonction des compétences explorées, soit la lecture et rédaction, soit les mathématiques[b]. A partir des 82 études EEF éligibles (certaines études n'ont pu être exploitées, car par exemple n'avaient pas abouti à des résultats ni en anglais ni en maths), trois méthodes statistiques permettant de conduire des méta-analyses ont été explorées. Les deux premières suivent un modèle classique en deux étapes (calculs des tailles d'effet pour chacune des études sélectionnées puis calcul d'une taille d'effet globale). La troisième, qui a été retenue, est une méta-analyse sur les données individuelles des participants : il s'agit alors de calculer en une étape les tailles d'effet des études et une taille d'effet globale à partir de l'ensemble des scores des élèves de toutes les études sélectionnées[c].

Des tailles d'effet globales ont également été calculées qui ont permis de conduire des analyses de sous-groupes.

[a] C'est le critère retenu par EEF pour identifier les élèves de milieux socio-économiques défavorisés.

[b] Aucun résultat n'est donné « toutes disciplines confondues ».

[c] On notera que le modèle retenu est un modèle simplifié qui finalement revient à calculer une taille d'effet globale en deux étapes (les équations pour chacun des modèles sont données dans le document P) ; les tailles d'effet de chacune des études sont présentées de façon classique par un graphique en forêt (voir annexe 5 pour les méthodes de calcul).

Chapitre 3. Les apprentissages en mathématiques

Aucun thème spécifique aux mathématiques n'est traité dans le Toolkit. Cependant, il est possible pour chacune des méta-analyses conduites sur un thème donné de sélectionner les tailles d'effet calculées à partir de scores obtenus par les élèves en mathématiques. Une analyse des résultats obtenus pour le thème enseignement collaboratif est proposé ci-dessous. Ce thème a été retenu ici car souvent étudié par les auteurs de méta-analyses récentes[a].

Deux rapports d'orientation concernent les mathématiques : l'un en direction des écoles maternelles (non analysé ici), et l'autre en direction des élèves de primaire et du collège (document Q). Ce dernier s'est appuyé sur un examen des preuves (document E). Une présentation rapide de ces deux documents est proposée en seconde partie.

Enfin, plusieurs études comparatives ont été menées par EEF sur le territoire britannique qui concernent les apprentissages en mathématiques. Les études d'efficacité en conditions réelles sont décrites dans la dernière partie de ce chapitre ainsi que les principaux résultats des méta-analyses conduites par EEF pour les élèves éligibles au FSM (document P).

On peut accéder à ces documents en cliquant sur Guidance for Teachers puis Mathematics.

Enseignement collaboratif (Toolkit)

1. Présentation générale des résultats EEF

La méta-analyse réalisée par EEF sur le thème apprentissage coopératif (*cooperative learning approaches*) du Learning Toolkit a sélectionné plusieurs études pour lesquelles 64 tailles d'effet ont été calculées à partir de scores obtenus par les élèves en mathématiques[b]. Sous le graphique en forêt, on trouve également la taille d'effet globale ES, les bornes de son intervalle de confiance et les indicateurs d'hétérogénéité[c] Q, T^2 et I^2. Toutes ces tailles d'effet ont été rassemblées dans un fichier excel et les résultats globaux ont été recalculés (tableau 13). Certaines tailles d'effet peuvent être considérées comme anormalement élevées : dans le domaine des sciences de l'éducation, il est en effet très rare d'obtenir des tailles d'effet supérieures à 1 pour des études menées dans de bonnes conditions (par exemple avec des échantillons de tailles importantes[d]). Deux autres calculs ont été menés en excluant d'abord la taille d'effet égale à 3,511, puis en excluant toutes les tailles d'effet supérieures à 1. Ces

[a] Voir par exemple *Comment aider les élèves en difficulté en mathématiques ? Les réponses de Campbell et du What Works Clearinghouse (WWC),* aout 2021, www.mathadoc.fr

[b] Dans le Technical Appendix de ce thème, sélectionner comme sujet Maths. Il ne s'agit pas du nombre d'études, car plusieurs études ont abouti au calcul de plusieurs tailles d'effet.

[c] En septembre 2021, seul la taille d'effet globale était calculée ; les autres résultats ont été publiés après requête personnelle.

[d] Voir par exemple la figure 45 p.113 dans *Mesurer l'effet d'un traitement. Les méta-analyses en sciences de l'éducation* (ROQUES, 2021).

calculs conduisent bien évidemment à des tailles d'effet globale inférieures à celle publiées par EEF (tableau 13).

		ES	Intervalle de confiance	I^2	Q	T^2
Recalculés à partir des 64 tailles d'effet	**Sans exception**	0,368	[0,244 - 0,492]	0,968	1967,199	0,211
	En retirant ES = 3,511	0,333	[0,208 - 0,458]	0,969	1978,464	0,212
	En retirant les ES >1	0,226	[0,097 - 0,354]	0,970	1831,846	0,200
Publiés par EEF (pour les 64 tailles d'effet)		0,371	[0,236 - 0,507]	0,973	1754,211	0,259

Tableau 13 : Résultats pour les tailles d'effet associées aux mathématiques

On avait noté que le thème apprentissage collaboratif proposait un effet évalué (toutes disciplines confondues) à 5 mois de progrès. Si on ne considère que les 64 tailles d'effet associées aux mathématiques, la taille d'effet globale correspond toujours à ce même nombre ; par contre quand on enlève l'étude présentant une taille d'effet égale à 3,511, le nombre de mois de progrès tombe à 4, et ce n'est plus que 3 mois quand les tailles d'effet supérieures à 1 sont toutes exclues.

On notera que les tailles d'effet calculés au niveau des études primaires ne prennent pas en compte les scores prétests et que la taille d'effet globale est calculée en suivant le modèle des effets aléatoires et en considérant que toutes les tailles d'effet sont indépendantes (voir le dernier chapitre pour une discussion à ce sujet).

Quand on trie les tailles d'effet par ordre croissant sur l'année de la publication (figure 3), on remarque que 31 d'entre elles (soit un peu moins de la moitié) proviennent d'études dont les résultats ont été collectés il y a 30 ans ou plus[a].

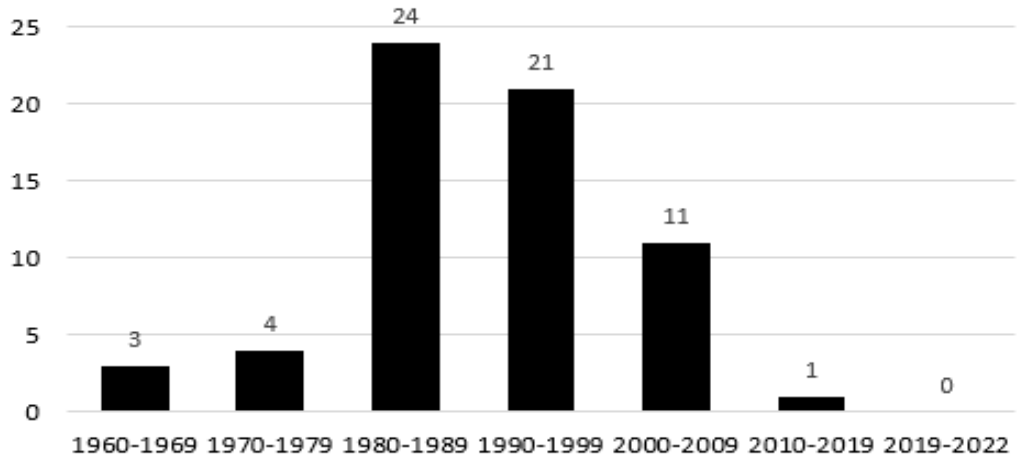


Figure 3 : nombre de tailles d'effet par année de publication

[a] Aucune limite concernant les dates de publication n'a pu être trouvée dans les documents publiés par EEF ; c'est pourtant un critère fréquent des méta-analyses qui souhaitent écarter les études jugées trop anciennes.

Le chercheur américain Robert SLAVIN a beaucoup travaillé sur les questions concernant l'apprentissage des élèves (son attention s'est notamment portée sur l'apprentissage collaboratif) en conduisant de nombreuses études primaires aux Etats-Unis mais également en réalisant plusieurs méta-analyses. L'intégration de ses résultats dans le Toolkit est présentée ci-dessous.

2. Les études de SLAVIN sélectionnées par EEF

Parmi les 64 tailles d'effet calculées par EEF pour le thème apprentissage collaboratif et associées au mathématiques, 13 sont issues de 10 articles publiés entre 1979 et 1985 par Robert SLAVIN et son équipe de chercheurs (voir résumé et principaux résultats calculés par EEF en annexe 7). La question de l'indépendance de ces tailles d'effet se pose alors, avec évidence pour les tailles d'effet calculées à partir des résultats d'une même étude[a], mais également pour des tailles d'effet calculées à partir des résultats de publications dont on peut se demander s'ils ne sont pas, au moins en partie, issus d'une seule et unique expérience. Par exemple, deux articles publiés l'un en 1983 (SLAVIN, 1983) et l'autre en 1984 (SLAVIN, 1984 b) font état d'expériences s'étant déroulées dans les écoles publiques du Comté de Howard dans le Maryland[b].

3. La méta-analyse de Robert SLAVIN

Robert SLAVIN a conduit une méta-analyse sur l'enseignement des mathématiques au secondaire[c] en incluant des études primaires selon des critères très restrictifs[d]. Plusieurs de ces études sélectionnées étaient associées à l'apprentissage collaboratif.[e]. Parmi ces études, quatre ont également été sélectionnées par EEF dans la méta-analyse du Toolkit (voir annexe 8 pour un résumé de ces études) concernant l'apprentissage collaboratif. Les tailles d'effet de ces 4 études ont été recalculées en utilisant la méthode mise en œuvre par EEF (voir annexe 5 pour les méthodes de calcul et le fichier excel), c'est-à-dire sans tenir compte des scores prétests, et les résultats ont été comparés (tableau 14).

[a] Pour trois études, deux tailles d'effet sont calculées.

[b] les tailles des échantillons sont également très proches.

[c] Slavin, R.E., Lake, C., et Groff, C. (2009). Effective programs in middle and high school mathematics : A best-evidence synthesis. *Review of Educational Research, 79 (2)*, 839-911.

[d] Pour être sélectionnée par la méta-analyse de SLAVIN, l'intervention devait être mise en œuvre au moins 12 semaines et c'est ce qui explique que certaines des études primaires de SLAVIN inclues dans la méta-analyse de EEF aient été exclues de la méta-analyse de SLAVIN.

[e] Cette méta-analyse a déjà fait l'objet d'une étude commentée (voir *Comment enseigner les maths ? La réponse du Center for Research and Reform in Education,* juin 2021, www.mathadoc.fr). Les tailles d'effet calculées par SLAVIN sont des Δ de Glass et tiennent compte des scores prétests.

	Publiées par EEF		Recalculées		Publiées par SLAVIN	
	ES	**erreur**	**ES**	**erreur**	**ES**	**erreur***
Barbato (2000)	0,841	0,145	0,841	0,144	1,09	0,1587
Nichols (1996)	0,164	0,224	Aucun résultat ne correspond		0,2	0,2381
Reid (1992)	0,649	0,291	0,649	0,286	0,38	0,2881
Slavin (1984 a)**	0,248	0,117	0,248	0,116	0,18	0,1166

Tableau 14 : Résultats publiés par EEF ainsi que par SLAVIN et recalculés pour 4 études

* : non publié par SLAVIN et calculé en utilisant les tailles d'échantillon publiées par SLAVIN.

** : SLAVIN publie trois tailles d'effet ; le résultat proposé par EEF semble correspondre à la comparaison du groupe Coopératif (sans Maitrise) versus groupe contrôle.

Sur les 4 études, un seul recalcul ne donne aucun résultat correspondant à ceux publiés par EEF (NICHOLS, 1996). Sans surprise, les résultats de EEF diffèrent de ceux publiés par SLAVIN, car les méthodes de calcul ne sont pas les mêmes.

On notera que trois autres articles sélectionnés par SLAVIN dans sa méta-analyse et associés à l'apprentissage collaboratif[a] ne sont pas sélectionnés par EEF. Cela vient du fait que la méta-analyse de SLAVIN elle-même n'a pas été sélectionnée par le Toolkit Collaborative de EEF (voir le document R pour avoir la liste des méta-analyses sélectionnées[b]), car le mot « *collaborative* » ne figure ni dans le titre ni dans le résumé de l'article publié par SLAVIN.

L'examen des preuves

L'examen des preuves (document E) publié en 2018 par EEF avait comme finalité de synthétiser l'ensemble des preuves disponibles permettant d'évaluer l'efficacité de stratégies déployées par les établissements scolaires pour faciliter l'apprentissage des mathématiques par les élèves à l'école primaire (*key stage 2*) et au collège (*key stage 3*). Les auteurs de cet examen ont formulé 24 questions de recherche qui correspondent entre autres à des caractéristiques pédagogiques, des domaines mathématiques et des outils ou supports utilisés par l'enseignant et les élèves[c]. Afin d'assurer un niveau de preuve suffisant, les résultats collectés et synthétisés lors de cet examen devaient être issus d'expériences contrôlés randomisées ou d'expériences quasi-expérimentales. Le nombre d'études primaires quantitatives concernées par ces 24 questions étant jugé trop important, les chercheurs ont

[a] CALHOON (2003), KRAMARSKI (2001) et MEVARECH (1997), références complètes dans *Comment enseigner les maths ? La réponse du Center for Research and Reform in Education,* juin 2021, www.mathadoc.fr.

[b] Les méta-analyses sélectionnées par ce document ne concernent pas uniquement les apprentissages en mathématiques.

[c] D'autres thèmes, comme les devoirs ou l'implication des parents ont été également analysés.

choisi d'utiliser en priorité les résultats publiés par des méta-analyses. Ce document a déjà fait l'objet d'une analyse[a] et seuls les points saillants sont repris ici.

Parmi les 24 questions de recherches, 10 d'entre elles ont soit exploité au moins 6 méta-analyses (et semblent donc les modules *a priori* les mieux documentés), soit affiché un niveau de preuve élevé[b]. Une description synthétique de ces 10 thèmes est proposée dans le tableau 15 et s'appuie sur les conclusions formulées par les auteurs de l'examen des preuves EEF.

Thèmes[c]	Efficacité[d]	Nombre de méta-analyses exploitées	Niveau de preuve
Technologie	Peu ou pas efficace	13	Outils : faible ; EAO* : modérée
Tutorat	**Efficace**	**10**	**Modéré**
Métacognition	Efficace avec réserve	8	Modéré
Problèmes	Efficace avec réserve	8	Investigation : faible ; résolution : modéré
Explicite	Efficace avec réserve	8	Modéré
Collaboratif	**Efficace avec réserve**	**6**	**Élevé**
Manipulation	Efficace avec réserve	5	Manipulation : élevé ; représentation : modéré
Feedback	**Efficace avec réserve**	**4**	**Élevé**
Calculatrices	**Efficace**	**2**	**Élevé**
Manuels	Aucune efficacité	2	Élevé

Tableau 15 : les modules les mieux documentés ou avec un niveau de preuve élevé

* : enseignement assisté par ordinateur ; en **gras** les questions de recherche efficaces ou efficaces avec réserve avec un niveau de preuve élevé.

D'après ce tableau synthétique (qui n'est pas proposé par l'examen des preuves qui ne propose aucune synthèse des résultats), on soulignera l'importance que l'on pourrait accorder aux pratiques soutenant le tutorat entre élèves, les apprentissages collaboratifs, les rétroactions auprès des élèves et l'usage de calculatrices.

En ce qui concerne le module apprentissage collaboratif, on notera que parmi les 6 méta-analyses recensées par cet examen des preuves, seules 2 ont été retenues en 2018 par la méta-analyse du Toolkit concernant le même thème[e] (document R).

[a] *Comment améliorer le niveau des élèves en mathématiques ? La réponse de l'Education Endowment Foundation*, septembre 2021, www.mathadoc.fr

[b] Le niveau de preuve ou niveau de confiance d'une synthèse thématique est un nombre compris entre 1 (le niveau le plus faible) et 3 (le niveau le plus élevé).

[c] Ce sont les titres abrégés des questions de recherche.

[d] Cette efficacité est une interprétation personnelle issue de la lecture de cet examen des preuves.

[e] Les méthodes suivies pour identifier les méta-analyses susceptibles d'être sélectionnées diffèrent : pour le Toolkit, le terme *collaborative* devait apparaitre dans le titre ou le résumé de la méta-analyse ce qui n'était pas le cas pour la méta-analyse de cet examen des preuves. Le Toolkit ne publie pas de listes des méta-analyses rejetées (ce que fait l'examen des preuves) : une méta-analyse absente peut donc ne pas avoir été repérée ou avoir été rejetée.

Le rapport d'orientation

Ce rapport (document Q) s'est appuyé sur l'examen des preuves précédemment évoqué et a été intégralement traduit en français[a]. Il propose 8 recommandations à l'adresse des enseignants et des personnels des établissements scolaires dont les titres sont les suivants :

1. S'appuyer sur les évaluations pour construire un enseignement fondé sur les connaissances et les compétences des élèves.

2. Utiliser des représentations concrètes et semi-concrètes.

3. Enseigner aux élèves des stratégies leur permettant de résoudre les problèmes.

4. Permettre aux élèves de développer un riche réseau de connaissances mathématiques

5. Développer l'autonomie et la motivation des élèves (améliorer leurs compétences métacognitives).

6. Exploiter des exercices et des ressources pédagogiques pour stimuler les élèves et leur permettre d'améliorer leurs compétences mathématiques.

7. Mettre en œuvre des interventions pour proposer un soutien supplémentaire aux élèves en difficulté en proposant notamment un enseignement explicite.

8. Faciliter la transition des élèves de l'école primaire vers le collège.

Pour chaque recommandation on trouve un très court résumé qualitatif des éléments probants soutenant la recommandation. Il s'agit bien évidemment des méta-analyses de l'examen des preuves (dont les références ne sont pas communiquées), mais aussi parfois des méta-analyses publiées par le What Works Clearinghouse aux USA. On notera l'absence de toute référence directe aux méthodes pédagogiques associées à l'apprentissage coopératif ainsi qu'au tutorat pourtant considérées très favorablement pas l'examen des preuves. *A contrario*, certaines recommandations (comme celle évoquant les connexions entre des faits et des concepts mathématiques, mais aussi le soutien aux élèves en difficulté ainsi que la question de la transition école – collège) sont absentes de l'examen des preuves.

Les études EEF portant sur les mathématiques

On va s'intéresser dans un premier temps aux études d'efficacité en conditions réelles car ce sont les études les plus exigeantes menées par EEF et que leurs conclusions en sont d'autant

[a] *Favoriser un enseignement efficace des mathématiques à l'école primaire et au collège. Guide d'orientation* www.mathadoc.fr. Non daté, on peut considérer qu'il a été probablement publié en 2018 peu après l'examen des preuves.

plus fortes. Ensuite, les principaux résultats obtenus par la méta-analyses conduite par EEF en 2021 (document P) seront présentés.

1. Les études EEF

Onze interventions (où projets tels qu'ils sont également dénommés sur le site internet d'EEF) ont comme sujet les mathématiques[a] et ont été l'objet d'une étude d'efficacité en conditions réelles. L'intervention *Research into Teacher Training* qui apparait dans cette liste n'a pas été menée à terme et n'a donc pas été retenue ici. On notera que l'étude menée pour évaluer l'intervention *Affordable Maths Tuition* devrait plutôt être considérée selon ses auteurs comme une étude d'efficacité se déroulant dans les meilleures conditions possibles (elle a tout de même été conservée dans la liste ci-dessous, tableau 16).
On sait déjà (chapitre précédent) que le niveau de confiance minimum attendu par EEF pour une étude de ce type est le niveau 3 : on peut donc considérer les résultats des études évaluant les interventions *Catch up numeracy* et *Maths Champions* comme peu fiables. Enfin le rapport d'évaluation de l'intervention *Shared Maths* n'est pas disponible (achevée en 2014, c'est l'une des premières interventions évaluées par EEF) et aucune information sur la précision des tailles d'effet calculées n'est donnée. Les informations présentées dans le tableau 17 sont issues des rapports d'évaluation de ces 10 interventions (à l'exclusion de l'intervention *Shared Maths* pour les raisons évoquées précédemment) ; on y trouve notamment les effectifs totaux (pour l'échantillon global mais aussi pour les élèves bénéficiant de l'aide alimentaire), les tailles d'effet, les intervalles de compatibilité (ou de confiance[b]) et les valeurs-p publiés. Si on s'intéresse plus particulièrement à la précision des estimations des tailles d'effet, on remarquera qu'aucune d'entre elles, à l'exception de la taille d'effet globale calculée pour l'intervention *Ark Mathematics Mastery*[c], n'est statistiquement significative au sens classique du terme pour un niveau de confiance égal à 0,95[d].
On notera enfin que les tailles d'effet calculées pour les élèves éligibles aux FSM (*Free School Meals,* c'est-à-dire bénéficiant d'une aide alimentaire) soit ne sont pas associées à un niveau de confiance, soit ont des niveaux de confiance inférieurs à ceux de l'étude menée sur l'échantillon total[e].

[a] C'est-à-dire que les élèves ont été évalués sur des sujets mathématiques.
[b] On a vu précédemment que ces deux dénominations différentes d'un même intervalle correspondent à des façons différentes de considérer les conclusions des tests d'hypothèse
[c] Cette taille d'effet a été calculé en suivant le modèle d'une méta-analyse à partir des deux tailles d'effet calculées pour les niveaux KS1 et KS2.
[d] Pour ces 9 études, l'intervalle de confiance inclue la valeur nulle ou bien la valeur-p est supérieure à 0,05.
[e] Cela vient du fait que la taille de l'échantillon des élèves éligibles aux FSM est bien sûr inférieure à celle de l'échantillon total.

Intervention	Date de publication*	Thème / Niveau	Effet	Niveau de confiance	Cout
1stClass@Number	Juillet 2018	Feedback Assesment / KS1	2	4	1
Pas d'amélioration pour le test de fin de KS1 ni pour les élèves éligibles au FSM (faible niveau de confiance). Une analyse de sous-groupe post-hoc montre que l'intervention a un effet significatif pour les élèves faibles.					
Affordable Maths Tuition	Juillet 2016	Maths / KS2	0 (-1 ancienne version)	3	3
Devrait être considérée comme une étude d'efficacité selon EEF (car évaluée dans des conditions idéales).					
Ark Mathematics Mastery	Février 2015	Learning Behavior/ year 1 et year 7	1	3	2
2 expériences, l'une en primaire (+2 mois, niveau de confiance 3) l'autre au secondaire (year 7, +1 mois, niveau de confiance 4). Ces deux résultats ont été globalisés (méta-analyse).					
Catch up numeracy	Fevrier 2019	Maths / KS1	0	2	1
- 2 mois pour les élèves éligibles au FSM.					
Chess in primary school	Juillet 2016	Learning Behavior/ KS1	0	5	1
Mesuré par le test en fin de KS2 un an après. Même résultat pour les élèves éligibles au FSM					
Increasing Competence and Confidence in Algebra and Multiplicative Structures (ICCAMS).	Décembre 2021	Feedback Assesment / KS3 (year 7 et 8)	0	3	1
+1 mois de progrès pour les élèves éligibles au FSM					
Mathematical Reasoning	Décembre 2018	Learning Behavior/ KS1	1	4	1
Suite à l'étude d'efficacité sur Improving Literacy and Numeracy in KS1 (+ 3 mois). Même résultat pour les élèves éligibles au FSM					
Maths Champions	Juillet 2018	Early year / KS1	2	2	1
Intervention en maternelle. Beaucoup d'enfants n'ont pas participé aux tests finaux.					
Shared Maths	2014	Learning Behavior/ KS1	0	4	1
Mise en œuvre difficile avec les élèves de niveau faible.					
Tutor Trust Affordable Tutoring	Novembre 2018	Maths / KS1 (year 6)	3	4	2
Fait suite à une étude d'efficacité. Même résultat pour les élèves éligibles au FSM					

Tableau 16 : premiers résultats pour les études d'efficacité en conditions réelles

FSM : free school meals ; * : ce n'est pas la date annoncée sur la page internet qui elle correspond à la fin de l'expérience (et qui est donc antérieure). En **gras** les interventions avec effet supérieur ou égal à 1 et un niveau de confiance supérieur ou égale à 3.

	Échantillon total		Élèves éligibles au FSM	
Intervention	**N**	**ES, [CI], valeur-p**	**N**	**ES, [CI], valeur-p**
1stClass@Number	491	0,18 [-0,08 ;0,43] p=0,09	149	-0,03 [-0,30 ; 0,36] p=0,92
Affordable Maths Tuition	578	-0,03 [-0,35 ; 0,28]	184	-0,08 [-1,23 ; 0,74]
Ark Mathematics Mastery	4176 (KS1)	0,10 [-0,01 ; 0,21] p<0,10	Pas d'informations	
	5938 (KS2)	0,06 [-0,04 ; 0,15]		
	Tout	**0,07 [0,00 ; 0,14] p<0,05**		
Catch up numeracy	1481	-0,04 [-0,21 ; 0,13]	551	-0,14 [-0,33 ; 0,09]
Chess in primary school	3865	0,01 [-0,15 ; 0,16] p=0,9	1321	0,01 [-0,18 ; 0,19] p=0,95
ICCAMS	18052	0,04 [-0,07 ; 0,15] p=0,51	4981	0,06 [-0,04 ; 0,16] p=0,22
Mathematical Reasoning	6353	0,08 [-0,03 ; 0,18] p=0,16	1323	0,09 [-0,07 ; 0,25] p=0,29
Maths Champions	628	0,10 [-0,13 ; 0,33] p=0,41	Pas d'informations	
Shared Maths*	3305 (year 3)	0,01	?	-0,05
	3167 (year 5)	0,02	?	0,05
Tutor Trust Affordable Tutoring	1201	0,19 [-0,05 ; 0,44] p=0,10	576	0,25 [-0,02 ; 0,51] p=0,06

Tableau 17 : résultats détaillés pour les études d'efficacité en conditions réelles
N = traitement + contrôle ; ES = taille d'effet ; CI = intervalle de confiance ou de compatibilité ; p = valeur-p ; FSM =Free School Meals ; * : les résultats sont uniquement visibles sur la page internet et aucune précision sur les estimations n'est communiquée.

2. Les méta-analyses conduites à partir des études EEF

En 2021, EEF publie dans un même document (document P) les résultats de plusieurs méta-analyses portant soit sur les scores en lecture/rédaction (*litteracy*) soit sur les scores en mathématiques des élèves éligibles au FSM. Toutes les étudesayant fourni des résultats dans ces deux disciplines ont été considérées par EEF (c'est-à-dire qu'aucune sélection ni prise en compte de la qualité des études ne sont effectuées).

Parmi elles, 48 résultats[a] concernaient les mathématiques et pour chaque résultat, une taille d'effet a été calculée. Conformément aux recommandations du rapport sur la présentation de la précision des estimations (document J), les « intervalles de confiance » sont devenus des « intervalles de compatibilité »[b] et il n'est plus mentionné qu'un résultat est (ou non) statistiquement significatif[c]. Les auteurs ont classé les tailles d'effet par ordre décroissant ce qui leur a permis d'en caractériser quatre comme prometteuses[d] (tableau 18).

Interventions	ES	Intervalle de compatibilité
Powerful Learning Conversations	0,31	[-0,25 ; 0,98]
Dialogue Teaching	0,16	[0,03 ;0,29]
Improv Num and Lit KS 2	0,13	[-0,10 ; 0,36]
Act, Sing, Play 1	0,12	[-0,10 ; 0,35]

Tableau 18 : taille d'effet et intervalles de compatibilité des 4 interventions prometteuses (élèves éligibles au FSM).

[a] Là encore il s'agit d'analyses (et non d'études en tant que publications, plusieurs résultats ayant pu être calculés pour une même étude).
[b] Mis à part quelques apparitions en tout début de texte.
[c] Mis à part une apparition p.39.
[d] Avant 2021, certaines interventions étaient labellisées comme étant « prometteuses ». La liste de ces interventions n'est plus publiée sur la nouvelle version du site internet.

On notera que ces calculs ne correspondent pas aux résultats calculés par les auteurs des études originales (chaque taille d'effet a été recalculée par les auteurs de la méta-analyse en utilisant un modèle de calcul particulier, voir annexe 5 pour plus de détails). Par exemple, dans le rapport de l'intervention *Powerful Learning Conversations* (qui est une étude pilote dont les résultats quantitatifs ne sont donc pas publiés sur internet), les auteurs de l'étude initiale calculent pour les 84 élèves éligibles au FSM une taille d'effet égale à 0,97 [0,09 ; 1,86].
Des tailles d'effet globales ont également été calculées. Quand on considère l'ensemble des 48 analyses donnant des informations sur les scores des élèves en mathématiques, la taille d'effet globale est égale à 0,00 avec comme intervalle de compatibilité [-0,03 ; 0,04]. D'autres résultats concernent les études regroupées par niveau d'étude[a], par type d'intervention[b], par design d'études[c] et par type de scores[d]. Toutes ces tailles d'effet sont proches de zéro et leur intervalle de compatibilité inclut la valeur nulle (voir annexe 6).
Enfin, la différence entre l'effet des interventions sur les élèves éligibles au FSM d'une part et sur les élèves non éligibles au FSM d'autre part a été mesurée en calculant là aussi une taille d'effet. En ce qui concerne les compétences mathématiques, elle est égale à -0,01 et son intervalle de compatibilité est [-0,04 ; 0,02]. Les auteurs concluent en affirmant que bien que l'écart estimé entre ces deux groupes d'élèves soit négatif et que cette analyse n'ait pas permis de montrer qu'il ait diminué (ce qui reste l'objectif principal d'EEF rappelons-le), il n'en reste pas moins qu'aucun élément probant ne vient appuyer la thèse contraire[e].
D'une façon générale, on retiendra qu'aucune interprétation de l'ampleur de ces tailles d'effet globale n'est proposée[f] et qu'elles ne sont pas traduites en nombre de mois de progrès comme c'est habituel pour les analyses EEF. Si cela avait été le cas, ce nombre aurait été égal à 0 mois pour toutes les méta-analyses. Les seuls commentaires concernent le signe (soit positif, soit négatif) de ces estimations.

[a] KS1, KS2, KS3 et KS4
[b] Intervention individuelle, en petits groupes, en classe entière ou concernant l'école entière
[c] Études multisites ou études par randomisation de clusters
[d] Score primaire (de première importance) et score secondaire. Parfois des interventions en écriture / lecture ont pu donner lieu à une évaluation secondaire en mathématiques par exemple.
[e] Cette remarque est écrite sans explicitation ni interprétation claire des résultats calculés.
[f] Cette remarque concerne également les tailles d'effet calculées pour chacune des études.

Chapitre 4. Bilan

On se doit pour commencer cette partie de souligner à nouveau l'importance des textes mis en ligne sur le site internet d'EEF, que ce soit en termes de volume mais aussi en termes de qualité. Ces textes, soit des pages internet, soit des fichiers téléchargeables, tous librement accessibles, sont le fruit d'un ensemble de travaux de recherches probablement unique au monde. On peut partager ces travaux en deux catégories : la réalisation de méta-analyses à partir d'études internationales tout d'abord, mais aussi (et surtout) la conduite d'un grand nombre d'expériences par comparaison de groupes menées sur le territoire britannique. Si des méta-analyses ont déjà été conduites par le passé par d'autres organismes dans le domaine des Sciences de l'éducation (on peut bien entendu citer le What Works Clearinghouse et le Center for Research and Reform in Education aux Etats-Unis mais aussi le réseau international Campbell), la réalisation de plus d'une centaine d'études primaires en une dizaine d'années, toutes réalisées dans des conditions similaires par des équipes qui, bien qu'indépendantes, partagent les mêmes procédures, est tout simplement exceptionnel. La rédaction de documents ressources à destination des évaluateurs et des chercheurs est également à mettre au crédit d'EEF.

Vient ensuite la volonté de mettre à la portée de tous dans un langage clair et accessible les informations scientifiques évoquées ci-dessus. On trouvera dans cette catégorie de textes les pages internet du Toolkit bien sûr, les guides d'orientation rédigés à l'attention des enseignants, mais aussi des guides s'intéressant à la mise en œuvre de recommandations concrètes (par exemple le document S). On citera également ici le guide *Moving forwards, making a difference. A planning guide for schools* publié en mai 2022 (document T) qui reprend à son compte les recommandations publiées dans les guides d'orientation concernant l'enseignement des mathématiques (par exemple le document Q) et l'enseignement de la lecture et de la rédaction, mais qui donne également de nombreux éléments de réflexion sur leurs mises en œuvre dans les établissements scolaires.

Publier autant d'informations sur un même site relève de la gageure et même si le site internet présente parfois quelques lacunes sur lesquelles nous allons revenir, la nouvelle mouture présentée à l'automne 2021 relève plutôt bien la plupart des défis (lisibilité, navigabilité, simplicité) associés à de tels projets. Le seul véritable reproche que l'on pourrait faire concerne la disparition de textes qui étaient publiés dans la version antérieure du site[a] : un dossier d'archives aurait permis de garder une trace de ce qui a existé pour rendre compte de l'évolution des publications et au-delà de l'évolution des résultats de la recherche. Parmi les détails qui pourraient être améliorés, certains sont presque inhérents au média utilisé. On notera par exemple que peu d'informations sont datées (presque toutes les pages internet sont concernées, mais aussi certains documents de niveau expert). Mais également que l'emploi d'une taille de police très importante ne permet pas à l'internaute d'avoir une vision

[a] Ainsi, la liste des méta-analyses retenues en 2018 a disparue du site internet (on peut la trouver dans le document U).

globale et synthétique des informations. Enfin plusieurs chemins d'accès aboutissant à une même page, il est facile de se perdre dans les méandres de ce site gigantesque.

Les commentaires qui suivent vont porter sur 4 axes. Le premier est en lien direct avec l'importance du nombre de documents publiés sur le site internet d'EEF et la multiplicité des procédures mises en œuvre qui en découle. Le second sera plus technique, car il s'agira de poser quelques questions sur le traitement statistique des données. Les deux derniers nous engagerons sur une réflexion plus profonde sur la communication d'une part et sur les missions mêmes d'EEF d'autre part.

Multiplicité des procédures

Les documents publiés sur le site internet d'EEF sont les fruits de multiples analyses, de plusieurs dizaines de travaux de recherche appliquée, ou encore de plus de 150 expériences de terrain qui ont été conçus, pilotés et analysés par plusieurs dizaines de personnes (chercheurs, responsables administratifs, enseignants, directeur d'établissements, etc.). Cette multiplicité d'intervenants notamment dans les équipes de recherche, se traduit par une multiplicité des façons de procéder, de rédiger et de conclure, et cela même si des procédures explicites encadrent la réalisation et la publication des analyses comme des expériences. C'est d'autant plus vrai que, en fonction du type d'analyse ou des questions posées, les règles ne sont pas toujours exactement les mêmes. On va prendre comme exemple le repérage et la sélection des études inclues dans une méta-analyse. On sait que cette étape est cruciale et qu'elle a une grande influence sur les conclusions tirées. Ainsi, quand pour un même sujet, on trouve à la fois une synthèse conduite dans le cadre du Toolkit, et une autre menée par des chercheurs indépendants dans le cadre d'un examen des preuves, des résultats et des conclusions très différents peuvent être observés. C'est par exemple le cas du thème rétroaction (*feedback*), qui affiche une taille d'effet globale égale à 0,481 (soit 6 mois de progrès) dans le Toolkit, alors que cette taille d'effet n'est plus que de 0,17 dans l'examen des preuves[a] (document D). Ces deux méta-analyses utilisent les mêmes critères de sélection à une exception près : la date de publication des études. Cette dernière devait être postérieure au 1er janvier 2000 pour l'examen des preuves, alors que les auteurs de la méta-analyse du Toolkit ont considéré tous les articles publiés après le 1er janvier 1960. Ils ont de ce fait intégrés 155 études dans leur synthèse, alors que l'examen des preuves n'en compte que 51.

Ces différences dans les procédures peuvent parfois s'expliquer par des différences entre les questions de recherche elles-mêmes. Par exemple, la méta-analyse menée par EEF sur ses propres études primaires a d'emblé considéré l'ensemble des études publiées entre 2011 et 2019, c'est-à-dire qu'aucune sélection n'a été opérée et que la qualité des études n'a pas été

[a] Ce qui correspondrait à 2 mois de progrès pour les élèves.

prise en compte. On sait pourtant que la force des résultats issus de méta-analyses repose en grande partie sur la qualité des études primaires sélectionnées[a].

Dans le même ordre d'idée, on notera que seuls les examens des preuves publient la liste des études (ou des méta-analyses le cas échéant) rejetées[b]. Et que les listes complètes des études retenues par les méta-analyses du Toolkit ne sont pas réellement publiées (les références affichées au troisième niveau d'information ne sont pas complètes, et aucun fichier pdf n'étant publié, il est pratiquement impossible de reconstituer une liste exhaustive).

On terminera en remarquant que les conclusions des examens de synthèses ne sont pas toujours reprises par les rapports d'orientation, même quand ces derniers affirment le contraire. C'est le cas nous l'avons vu pour le rapport publié sur l'enseignement des mathématiques qui ne fait allusion à pratiquement aucun des thèmes mis en avant par l'examen de preuves associé. Là encore, des regards différents sont posés sur une même question, qui ne peuvent sans doute pas produire une seule et unique réponse. Et les divergences dans les conclusions sont tout simplement le reflets de postures professionnelles différentes.

Les analyses statistiques

Des différences peuvent également être notées quant aux méthodes de calculs de résultats statistiques. On a déjà noté dans les chapitres précédents que les tailles d'effet des études primaires ont pu être calculées de plusieurs façons. Ainsi, les méta-analyses du Toolkit ne prennent pas les scores prétests des élèves en compte contrairement aux calculs menés dans le cadre des études EEF. Cependant, dans ces deux cas de figure, les tailles d'effet sont des différences standardisées de moyennes alors que les méta-analyses conduites par EEF sur ses propres études utilisent les scores des élèves comme donnée de base (et non des moyennes). Ces différences de méthodes perdurent quand on considère cette fois le calcul des tailles d'effet globales (voir annexe 5).

On serait pourtant en droit d'attendre d'un organisme se posant comme promoteur de l'*evidence based education*[c] qu'il impose une méthode d'analyse (et une seule), méthode qui serait alors reconnue comme étant la plus satisfaisante pour obtenir des réponses claires et indiscutables aux nombreuses questions posées. Ou, si ce n'était pas le cas, que les différences entre les méthodes employées soient au moins reconnues et discutées. Là encore, malgré la publication de plusieurs textes cadres, la multiplicité des intervenants mais peut-être aussi

[a] On a remarqué précédemment que des études pilotes ont pu être sélectionnées dans cette méta-analyse, alors même qu'EEF ne considère leurs résultats quantitatifs que comme des indications (ils ne sont par exemple associés à aucun niveau de confiance).

[b] Cette liste est souvent accompagnée d'une courte explication sur les motifs d'exclusion.

[c] Pour les anglo-saxons, ce titre regroupe les études qui exploitent les résultats numériques d'expériences de terrain ou d'études observationnelles à grande échelle.

des objectifs, conduit à un manque de cohérence qui affaiblit la force des conclusions et prête le flanc à la critique.

Deux autres points techniques doivent également être évoqués.

Le premier concerne le calcul des tailles d'effets globales obtenues à partir de données qui ne sont vraisemblablement pas toutes indépendantes. Les méta méta-analyses autrefois mises en œuvre pour le Toolkit, ont été abandonnées en 2019 entre autres pour cette raison : les méta-analyses de premier niveau sélectionnées avaient très souvent des études en commun (on parle de recouvrement ou *overlapping*), et elles n'étaient donc pas indépendantes les unes des autres ; on sait que cette indépendance des données conditionne l'utilisation de certains calculs statistiques. Malheureusement ce défaut perdure encore aujourd'hui. Nous avons vu en ce qui concerne le thème « apprentissage collaboratif » du Toollkit que des résultats non indépendants (issus d'une même étude ou issus d'études différentes mais probablement dépendantes) sont traités exactement comme s'ils étaient indépendants. Cette question retient l'attention des chercheurs actuellement et deux méthodes sont souvent proposées. La première consiste à ne sélectionner dans la méta-analyse que des études jugées indépendantes entre elles, puis à ne sélectionner qu'un seul résultat par étude. Cela revient donc à mettre de côté certaines des études repérées et certains des résultats publiés. C'est la méthode retenue par le What Works Clearinghouse par exemple. La deuxième consiste à inclure tous les résultats, y compris ceux pour lesquels un lien de dépendance existe, et à tenir compte de cette dépendance au niveau des calculs statistiques, par exemple en utilisant la méthode de l'estimation robuste de la variance (*Robust Variance Estimation*, RVB). C'est la méthode retenue par la collaboration Campbell par exemple.

Le second point concerne l'interprétation de la précision des estimations que sont les tailles d'effet calculées. On a vu en effet que EEF recommande aujourd'hui pour ses expériences menées sur le terrain britannique de ne plus caractériser un résultat (c'est-à-dire une taille d'effet) comme étant (ou n'étant pas) « statistiquement significatif ». Et comme on l'a vu dans la méta-analyses menées en 2019 (document P), plus aucun commentaire n'accompagne les intervalles de compatibilité. La présentation dichotomique habituelle des résultats est rejetée par EEF qui la juge trop réductrice car souvent mal interprétée[a]. Cette position peut paraitre surprenante et pour l'instant elle ne semble pas être partagée par les autres organismes cités ci-dessus. Comme nous l'avons déjà évoqué, on peut imaginer que la faible ampleur d'un grand nombre des résultats obtenus par EEF est au moins partiellement à l'origine de cette prise de position[b]. Et même si ce n'était pas le cas, il n'en reste pas moins que les concepts et procédures des tests d'hypothèses largement utilisés par la communauté scientifique de nos jours, s'appuient sur des modèles statistiques qui conduisent inévitablement à des conclusions dichotomiques : après avoir choisi *a priori* un niveau de confiance (en général 95%), les

[a] Le document J explique clairement par exemple ce qu'est une valeur-p et ce qu'elle n'est pas.

[b] Les auteurs du document K ne sont pas loin de l'admettre et l'absence de commentaires sur les tailles d'effet très faibles calculées dans les méta-analyses conduites sur les études primaires EEF ne fait que renforcer cette impression.

chercheurs calculent une estimation ponctuelle à partir des mesures prises sur un échantillon et concluent en rejetant ou non ce qu'il est convenu d'appeler l'hypothèse nulle. Cette interprétation dichotomique habituelle des résultats est tout simplement liée à la question de recherche elle-même : cette intervention est-elle, oui ou non, efficace ? Convient-il, oui ou non, d'élargir sa mise en œuvre à tout un territoire[a] ? On voit ici que cette position singulière d'EEF, bien qu'imperceptible pour le grand public, n'est pas un détail et mériterait d'être discutée.

On terminera ici en notant que, finalement, les critiques les plus importantes concernent avant tout les procédures encadrant la rédaction des recommandations du Toolkit, qui constituent la partie visible de l'iceberg qu'est le site EEF. Il serait donc souhaitable que ces procédures soient revues de manière à ce que les conclusions soient basées sur des preuves réellement irréfutables (voir encadré ci-dessous).

Toolkit : les points à améliorer

- Prendre en compte les scores prétests des élèves
- Prendre en compte la non indépendances des tailles d'effet dans le calcul de l'effet global
- Fixer une date limite pour éviter de sélectionner des études trop datées
- Renforcer les critères de sélection des études (des tailles d'effet supérieures à 1 sont suspectes)
- Publier les listes des études sélectionnées et des études non-sélectionnées (avec leur motif d'exclusion)
- Publier tous les résultats statistiques quand les études sont filtrées par sujet.

La lecture des textes publiés par EEF alimente bien évidemment notre réflexion sur un thème extrêmement exposé, que ce soit d'un point de vue médiatique mais aussi politique. La mission que s'est fixée Education Endowment Foundation est claire : trouver et sélectionner des interventions dans le but d'élever le niveau général des élèves mais également réduire les écarts entre les élèves de groupes socioéconomiques différents[b]. D'autres pays poursuivent bien entendu ces mêmes objectifs. On peut légitimement se demander si les résultats publiés par EEF en dix années d'existence ne viendraient pas finalement remettre en question la faisabilité de cette double mission qui s'apparente fort à une injonction paradoxale. Mais Education Endowment Foundation est également un organisme financé par l'État britannique

[a] On peut rapprocher cette discussion du débat qui oppose deux approches différentes quant à la mise en œuvre de tests statistiques : l'approche de Fisher (recommandée par EEF) et l'approche de Neyman et Pearson (ici dénommée approche « habituelle »). L'inférence bayésienne est mise en œuvre dans l'analyse statistiques des résultats de l'intervention *Catch up Numeracy,* mais aucune référence explicite à ce type d'inférence n'a été trouvé dans les documents publiés par EEF.

[b] « *What we are is a charity with a moral imperative – to support teachers and senior leaders to raise attainment and close the disadvantage gap – which roots its response to this educational challenge in the best available evidence* » sur la première page après avoir cliqué sur Who we are

et donc peu ou prou partie prenante d'un discours politique. Et se faisant, on se doit également de considérer la diffusion des informations comme faisant partie d'une communication politique en direction du grand public.

Communiquer

Avant de revenir sur le premier point qui concerne directement la mission d'EEF, nous allons prendre un exemple pour montrer que les informations présentées sur le site EEF peuvent ne pas toujours respecter les règles transparentes d'une publication scientifique. Ainsi, sur la page permettant d'accéder aux documents et études EEF concernant l'enseignement des mathématiques[a], six interventions[b] sont présentées comme favorables à l'apprentissages des mathématiques. On trouve en première place l'intervention *Mathematical reasoning* ; cette étude d'efficacité en conditions réelles a pourtant montré un nombre de mois de progrès de 1 mois seulement et on notera que le lien internet proposé à ce niveau conduit sur la page de l'étude d'efficacité, plus favorable puisque le nombre de mois de progrès est de 3. Vient ensuite l'intervention *Tutor trust Affordable Tutoring.* Deux études ont été menées par EEF : une première étude d'efficacité qui a calculé un nombre de mois de progrès de – 1 mois mais sans pouvoir lui associer de niveau de confiance (lié en partie à des difficultés de recrutement) et vers laquelle pointe le lien présent au niveau du nom de l'intervention, et une étude d'efficacité en conditions réelles qui elle a montré un résultat positif de 3 mois de progrès. C'est d'ailleurs cette étude qui est mise en lien sur l'une des 4 vignettes que l'on trouve sous les descriptions des deux interventions sus-citées (capture d'écran, figure 4). On reconnait ici l'intervention dont on vient de parler mais aussi *Affordable Maths Tuition* qui a elle aussi été analysée par une étude d'efficacité en conditions réelles. Les deux autres interventions ont été analysées par des études d'efficacité (mises en œuvre dans les meilleures conditions possibles donc) ; pour *onebillion* (étude publiée en 2019) l'effet sur l'ensemble des élèves est de 3 mois de progrès[c] (avec un niveau de confiance maximal de 5). Mais là aussi, on apprend en lisant le rapport, qu'il est de – 2 mois[d] sur les élèves éligibles au FSM (avec un niveau de confiance non calculé). Et enfin, pour *ReflectED Metacognition*[e] (étude publiée en 2016) l'effet est de 4 mois de progrès pour les scores en mathématiques[f] avec un niveau de confiance de 4. Si on prend la peine de lire le rapport, on apprend que cet effet est seulement de 2 mois pour les élèves éligibles au FSM[g], ce qui implique que les écarts entre les élèves de groupes

[a] Guidance for teachers puis Mathematics
[b] En fait il n'y en a que 5, mais l'une d'elle (*Tutor trust Affordable Tutoring)* a fait l'objet de deux études.
[c] ES = 0,24 [0,12 ; 0,36] p = 0,000078
[d] ES = -0,10 [-0,33 ; 0,14] p = 0,43
[e] La dernière intervention *ReflectED Metacognition* n'est pas incluse dans la catégorie des études ayant comme sujet les mathématiques.
[f] ES = 0,30 [-0,04 ; 0,63] p = 0,08
[g] ES = 0,14 [-0,26 ; 0,53] p = 0,5

socioéconomiques différentes ont augmenté. Pour les 3 premières interventions citées ici, des informations sont rassemblées dans les tableaux 16 et 17.

On notera que les trois premières vignettes montrent les trois meilleurs résultats pour le sujet des mathématiques.

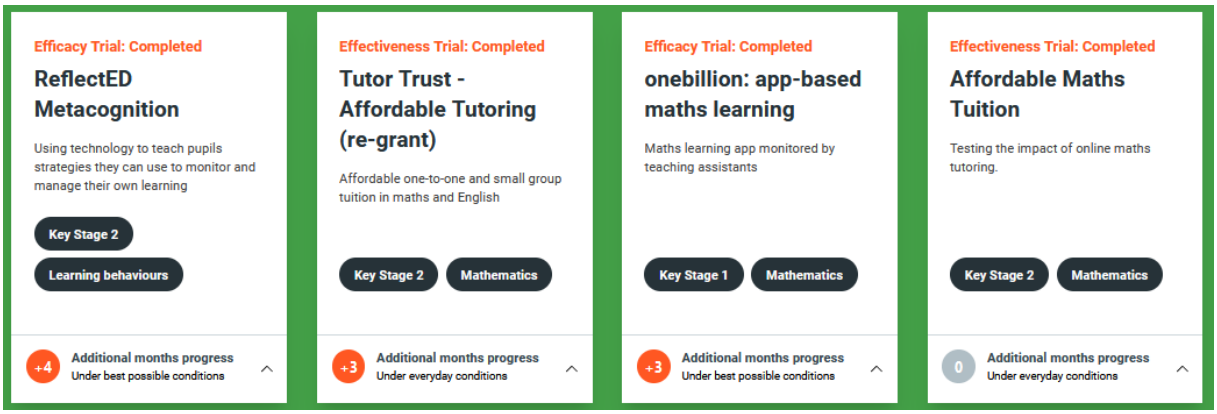


Figure 4 : capture d'écran des 4 interventions mises en avant par EEF pour les mathématiques

Cette présentation d'une toute petite partie du site internet est l'occasion de souligner l'importance qu'il convient d'accorder à la sélection des résultats proposés en première page au grand public. C'est une question cruciale qui mériterait d'être discutée afin que des règles claires, transparentes et connues de tous soient fixées.

Missions impossibles ?

Pour revenir au premier point qui concerne la mission d'EEF, deux constats peuvent d'ores et déjà être posés : (1) si on se base sur les études de très bonnes qualité conduites par EEF, peu d'interventions semblent réellement efficaces au niveau de l'ensemble des élèves d'une part, (2) et trouver une intervention susceptible de réduire les écarts entre les publics de niveaux socio-économiques différents tient de la gageure d'autre part.

On a en effet déjà montré que les résultats des études d'efficacité menées sur le terrain dans des conditions proches de la réalité (les *effectiveness trial*) montrent que pour une grande majorité des interventions, le nombre de mois de progrès est inférieur ou égal à 1 mois et ne dépassent jamais 3 mois (tableau 10). La liste des interventions prometteuses autrefois publiée a aujourd'hui disparue du site internet, et bien que cela ne soit qu'une supposition (aucune information n'étant communiquée sur les changements opérés), il est possible que les résultats modestes aient invités les rédacteurs du site à plus de prudence. En ce qui concerne la réduction attendue des écarts entre les élèves, plusieurs éléments peuvent être considérés. Tout d'abord, quand on s'intéresse plus particulièrement aux scores des élèves éligibles au FSM, on constate que parfois leur évolution après une intervention est soit moins favorable que celle des tous les élèves (éligibles et non éligibles au FSM) et parfois même

négative (on l'a vu par exemple avec l'intervention *onebillion*, voir aussi les tableaux 16 et 17). Les résultats de la méta-analyse menée par EEF sur ses propres études en 2019 (document P) vont également dans ce sens : toutes les tailles d'effet globales calculées à partir des scores d'élèves éligibles au FSM sont très proches de zéro[a] et ce quelque soit le regroupement choisi ; de plus, les élèves éligibles au FSM n'ont pas mieux réussi que les élèves non éligibles au FSM quand on considère l'ensemble des interventions. On ajoutera que les premiers éléments de réflexion qui proviennent de la comparaison de l'évolution des variances entre les groupes traitement et contrôle (document L) ont montré sur les 16 études EEF analysées que, quand des évolutions des variances significatives ont été observées, elles étaient toutes défavorables au groupe traitement. En d'autres termes, les écarts entre les élèves étaient plus importants après l'intervention dans le groupe traitement que dans le groupe contrôle.

Cela fait 10 ans maintenant que EEF produit des analyses sur des interventions censées améliorer les connaissances et compétences des élèves, et notamment des élèves issus de milieux défavorisés. Pour rassembler des preuves et trouver les chemins à suivre, de nombreux chercheurs se sont mobilisés qui ont pu mener à bien leurs travaux grâce au soutien financier conséquent de l'État britannique. Les questions abordées dans cette dernière partie semblent montrer que le temps est venu aujourd'hui de tirer un premier bilan afin d'ajuster les objectifs et d'affiner les questions de recherche en conséquence.

[a] Et pour reprendre un vocabulaire habituellement employé, aucune des tailles d'effet globales calculées n'est statistiquement et significativement différente de zéro.

Les annexes

Annexe 1. Les documents EEF

Les titres en **gras** signalent qu'un résumé est disponible dans l'annexe 6.

A- Sutton Trust-EEF Teaching and Learning Toolkit & EEF Early Years Toolkit. Technical appendix and process manual (Working document v.01)
Sans auteurs
Juillet 2018
Ce texte peut encore être trouvé sur internet, mais il n'est plus en lien avec le site EEF.

B- Education Endowment Foundation Evidence Database : Protocol and Analysis Plan
Auteurs : Steve Higgins, Alaidde Berenice Villanueva Aguilera, Emma Dobson, Louise Gascoine, Maria Katsipataki, Taha Rajab
Version 1.1, juin 2019
https://educationendowmentfoundation.org.uk/public/files/Toolkit/EEF_Evidence_Database_Protocol_and_Analysis_Plan_June2019.pdf
Pas accessible depuis le nouveau site (chemin d'accès introuvable)

C- Teaching and learning. Early years toolkit guide
Sans auteurs
Septembre 2021
https://educationendowmentfoundation.org.uk/education-evidence/using-the-toolkits

D - The impact of Feedback on student attainment: a systematic review
Auteurs : Newman, M., Kwan, I., Shucan Bird, K., Hoo, H.T.
2021
https://educationendowmentfoundation.org.uk/public/files/Publications/Feedback/Teacher_Feedback_to_Improve_Pupil_Learning.pdf

E- Improving Mathematics in Key Stages Two and Three: Evidence Review
Auteurs : Jeremy Hodgen, Colin Foster, Rachel Marks, Margaret Brown
Mars 2018
https://educationendowmentfoundation.org.uk/education-evidence/evidence-reviews/mathematics-in-key-stages-2-and-3

F - Evidence review protocole template
Sans auteurs
Dernière mise à jour : février 2020
https://educationendowmentfoundation.org.uk/education-evidence/evidence-reviews/review-resources

G - Evidence review template
Sans auteurs
Dernière mise à jour : février 2020
https://educationendowmentfoundation.org.uk/education-evidence/evidence-reviews/review-resources

H - Statistical analysis guidance for EEF evaluations
Sans auteurs
Mars 2018
https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/evaluation-design

I - Impact evaluation report template
Sans auteurs
Sans date
Modèle word pour les rapports d'évaluation (études primaires)
https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/reporting-templates

J - Statement on statistical significance and uncertainty of impact estimates for EEF evaluations
Sans auteurs
Février 2020
https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/evaluation-design

K - Statistical uncertainty in Randomised Controlled Trials
Sans auteurs
Date probable (indiquée sur le site internet) : 2018
https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/evaluation-design

L – Research paper n°2_Standard deviation as an outcome on interventions : a methodological investigation
Auteurs : Peter Tymms, Adetayo Kasim
Février 2018
https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/eef-evaluation-reports-and-research-papers/methodological-research-and-innovations/standard-deviation-as-an-outcome-on-interventions

M - Classification of the security of findings from EEF evaluations
Sans auteurs
Version 2.0 – juillet 2019
https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/evaluation-design

N - Cost evaluation guidance for EEF evaluations
Sans auteurs
Décembre 2019
https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/evaluation-design

O - Implementation and process evaluation guidance for EEF evaluations
Sans auteurs
Aout 2019
https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/evaluation-design

P - Individual participant data meta-analysis of the impact of EEF trials on the educational attainment of pupils on Free School Meals: 2011 - 2019
Auteurs : Bilal Ashraf, Akansha Singh, Germaine Uwimpuhwe, Tahani Coolen-Maturi, Jochen Einbeck, Steve Higgins, Adetayo Kasim
Mai 2021
https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/eef-evaluation-reports-and-research-papers/syntheses-of-eef-evaluations/meta-analysis-of-the-impact-of-eef-trials-on-fsm-pupils-2011-2019

Q - Improving mathematics in key stages two and three. Guidance Report
Auteurs : Peter Henderson, Jeremy Hodgen, Colin Foster, Dietmar Kuchemann
Date probable : 2018
https://educationendowmentfoundation.org.uk/education-evidence/guidance-reports/maths-ks-2-3#nav-download-the-guidance-report-and-poster

R - Collaborative learning Teaching & Learning Toolkit
Sans auteurs
13 Novembre 2018
Ce texte n'est plus publié sur internet.

S - Putting evidence to work : a school's guide to implementation. Guidance report
Auteurs : Jonathan Sharples, Bianca Albers, Stephen Fraser, Stuart Kime
Décembre 2019
https://educationendowmentfoundation.org.uk/education-evidence/guidance-reports/implementation

T - Moving forwards, making a difference. A planning guide for schools
Sans auteur
Mai 2022
https://educationendowmentfoundation.org.uk/support-for-schools/school-planning-support

U – The Sutton Trust-Education Endowment Foundation Teaching and Learning Toolkit
Auteurs : Higgins, S., Katsipataki, M., Villanueva-Aguilera, A.B., Coleman, R., Henderson, P., Major, L.E., Coe, R., Mason, D.
Décembre 2016
Ce texte peut encore être trouvé sur internet, mais il n'est plus en lien avec le site EEF.

Annexe 2. Liste des thèmes du Toolkit[a]

	Mai 2011[b]	Janvier/juillet 2013[b]	Février/octobre 2014 et juillet 2018[b]	Juin 2019[c]	Avril 2022[d]
Ability grouping	O	X			
After school programmes	O	X			
Arts participation	O				
Assessment for learning	O	X			
Aspiration interventions		O			X
Behaviour interventions		O			
Block scheduling	O				X
Built environment				O	
Collaborative learning		O			
Digital technology (ICT)	O				X
Early years intervention	O				X
Extending school time		O			
Feedback	O				
Homework	O	X		O	
Homework (Primary)		O		X	
Homework (Secondary)		O		X	
Individualised instruction	O				
Learning styles	O				X
Mastery learning		O			
Mentoring		O			
Metacognition and self-regulation	O				
One to one tuition	O				
Oral language interventions			O		
Outdoor adventure learning		O			X
Parental engagement	O				
Peer tutoring	O				
Performance pay	O				
Phonics		O			
Physical activity		O		X	O
Reading comprehension strategies			O		
Reducing class size	O				
Repeating a year		O			
School uniform	O				X
Setting or streaming		O			
Small group tuition		O			
Social and emotional learning		O			
Sports participation	O				X
Summer schools	O				
Teaching assistants	O				
Within class grouping			O	X	O

O : ajouté ; X : supprimé ou fusionné ; en **gras** les thèmes présents en avril 2022 sur le site internet.

[a] Les intitulés ont pu connaitre de légères modifications au fil des ans.

[b] Document A

[c] Document B

[d] Consultation du site internet, avril 2022

Annexe 3. Systèmes scolaires français, britannique et américain

Âges	**France**		**Royaume-Uni**		**USA**
			Years (Y)	Key Stage (KS)	
3 - 4 ans	Petite section	Cycle 1	Nursery		Pre school
4 – 5 ans	Moyenne section		Reception class		Pre-K
5 - 6 ans	Grande section		year 1	KS 1	Kindergarten
6 -7 ans	CP	Cycle 2	year 2 (examen KS1)		1er grade
7 – 8 ans	CE1		Year 3	KS 2	2ème grade
8 – 9 ans	CE2		Year 4		3ème grade
9 – 10 ans	CM1	Cycle 3	Year 5		4ème grade
10 – 11 ans	CM2		Year 6 (examen KS2)		5ème grade
11 -12 ans	6ème		year 7 (examen KS2)	KS 3	6ème grade
12 – 13 ans	5ème	Cycle 4	year 8 (examen GCSE*)		7ème grade
13 – 14 ans	4ème		year 9		8ème grade
14 – 15 ans	3ème		year 10	KS 4	9ème grade
15 - 16 ans	2nde		Year 11		10ème grade

* : *General certificate of secondary school*

Annexe 4. Nombre de mois de progrès et tailles d'effet

Fourchettes des tailles d'effet (ES) pour chaque nombre de mois de progrès des élèves.
Les nombres de mois de progrès négatifs ne sont pas explicités dans les documents I et P.

Nombre de mois de progrès	**ES minimum et maximum**	**Références**
- 3 mois	-0,26 à -0,19	Toolkit 2018 (document B)
- 2 mois	-0,18 à -0,10	Toolkit 2018 (document B)
- 1 mois	-0,09 à -0,02	Toolkit 2018 (document B)
0 mois	-0,01 à 0,01	Toolkit 2018 (document B)
	-0,05 à 0,05	Toolkit 2021 (document I)
	-0,04 à 0,04	Études EEF (document P)
1 mois	0,02 à 0,09	Toolkit 2018 (document B)
	0,06 à 0,09	Toolkit 2021 (document I)
	0,05 à 0,09	Études EEF (document P)
2 mois	0,10 à 0,18	Même seuils pour tous les documents.
3 mois	0,19 à 0,26	
4 mois	0,27 à 0,35	
5 mois	0,36 à 0,44	
6 mois	0,45 à 0,52	
7 mois	0,53 à 0,61	
8 mois	0,62 à 0,69	
9 mois	0,70 à 0,78	
10 mois	0,79 à 0,87	
11 mois	0,88 à 0,95	
12 mois	0,96 à 1,00	

Annexe 5. Calculs de taille d'effet (*Effect Size, ES*)

Les calculs sont présentés ici d'une manière simplifiée ; pour plus de détails consulter *Mesurer l'effet d'un traitement. Les méta-analyses en sciences de l'éducation* (ROQUES, 2021, www.mathadoc.fr).

Taille d'effet d'une étude primaire

Toolkit après 2019

La taille d'effet (ES) et son erreur standard (SE) sont calculées à partir des scores post-test, sans tenir compte des scores pré-test (document F). Il s'agit du g de Hedges avec les moyennes post-test (ω est un facteur correctif pour les échantillons de petite taille).

$$ES = \frac{m_t - m_c}{s} \times \omega \qquad (1)$$

$$\mathrm{s} = \sqrt{\frac{(n_t - 1){s_t}^2 + (n_c - 1){s_c}^2}{n_t + n_c - 2}}$$

$$\omega = 1 - \frac{3}{4(n_t + n_c) - 9}$$

$$SE^2 = \frac{n_t + n_c}{n_t \times n_c} \times \omega^2 + \frac{g^2}{2(n_t + n_c)} \qquad (2)$$

m_t= moyenne des scores post-test du groupe traitement

m_c= moyenne des scores post-test du groupe contrôle

s = écart-type groupé (calculé à partir des écarts-types des groupes traitement et contrôle).

n_t= taille de l'échantillon traitement

n_c= taille de l'échantillon contrôle

s_t= écart-type des scores de l'échantillon traitement

s_c= écart-type des scores de taille de l'échantillon contrôle

Études EEF

Dans les rapports d'intervention publiées pour chaque étude EEF, la taille d'effet (ES) et son erreur standard (SE) sont calculées à partir des scores post-test en tenant compte des scores pré-test (document N). Il s'agit également du g de Hedges mais ici les moyennes post-test sont ajustées aux scores prétest après avoir menée une ANCOVA. Dans l'équation (1) on remplace donc m_t et m_c par ces moyennes ajustées. Un modèle linéaire est également utilisé (voir ci-dessous).

Méta-analyses sur les études EEF

Une méta-analyse a été conduite sur les études EEF antérieures à 2020 et les tailles d'effet de chaque étude ont été recalculées en utilisant un modèle simplifié de méta-analyse sur les données individuelles des participants (document T). Le score de l'élève i qui est scolarisé dans l'école j est utilisé par l'analyse[a] k. La taille d'effet d'une étude est calculée en utilisant l'équation (3) basée sur un

[a] Certaines études publient les résultats de plusieurs analyses.

modèle de régression linéaire multiple. Pour pouvoir utiliser ce modèle simplifié, les scores post-tests et prétests doivent être standardisés.

$$Y_{ijk} = \beta_{0k} + \beta_{1k} * Pret_{ijk} + \beta_{2k} * T_{ijk} + S_{jk} + \varepsilon_{ijk} \qquad (3)$$

Y_{ijk} : score posttest standardisé de l'élève i scolarisé dans l'école j utilisé par l'analyse k

$Pret_{ijk}$: score prétest standardisé de l'élève i scolarisé dans l'école j utilisé par l'analyse k

T_{ijk} = 1 si l'élève est dans le groupe traitement, = 0 si l'élève est dans le groupe contrôle.

β_{0k} : intercept

β_{1k} : gradient entre scores posts test et prétest

β_{2k} : effet moyen de l'analyse k.

$S_{jk} \sim N(0, \omega_{sk}^2)$: variation entre les écoles dans l'analyse k

$\varepsilon_{ijk} \sim N(0, \sigma_k^2)$: variation entre les élèves dans l'analyse k

Pour une description de ces calculs avec les données de l'intervention *Firstclass@numbers* voir l'annexe 5 de *La régression linéaire multiple dans les études par comparaison de groupes en milieu scolaire* (ROQUES, 2022, www.mathadoc.fr)

Taille d'effet globale d'une méta-analyse

On note ES la taille d'effet globale estimée.

Toolkit avant 2019

Avant 2019, pour chaque thème du Toolkit des méta-méta-analyses (ou méta-analyses secondaires) étaient menées par EEF en sélectionnant des méta-analyses publiées. C'est le modèle des effets fixes qui est utilisé. Soit i le rang de la méta-analyse sélectionnée, on a alors (4) :

$$ES = \frac{\sum \frac{ES_i}{var_i}}{\sum \frac{1}{var_i}} \qquad (4)$$

ES_i : taille d'effet globale de la méta-analyse i, non calculée par EEF (ce résultat est publié par la méta-analyse sélectionnée)

var_i : variance de la taille d'effet globale de la méta-analyse i, non calculée par EEF (ce résultat est publié par la méta-analyse sélectionnée)

Toolkit après 2019

Pour chaque thème du Toolkit, une méta-analyse est conduite à partir des études sélectionnées (et identifiées grâce aux méta-analyses sélectionnées avant 2019). C'est le modèle des effets aléatoires qui est utilisé. On a alors (5) :

$$ES = \frac{\sum \frac{ES_i}{var_i^*}}{\sum \frac{1}{var_i^*}} \qquad (5)$$

Avec

ES_i : taille d'effet de l'analyse i calculée par EEF (voir équation (1))

var_i^* : variance de la taille d'effet de l'étude i (équation (2)) à laquelle on ajoute la variance inter-étude τ^2, calculée par EEF. Donc $var_i^* = {SE_i}^2 + \tau^2$

Méta-analyses sur les études EEF

Après avoir calculé les coefficients de régression β_{2k} (voir équation (3)) et les variances ω_{sk}^2 et σ_k^2, on calcule la taille d'effet globale ES à l'aide de l'équation (6) :

$$ES = \frac{\sum \frac{\beta_{2k}}{\omega_{sk}^2 + \sigma_k^2}}{\sum \frac{1}{\omega_{sk}^2 + \sigma_k^2}} \qquad (6)$$

Annexe 6. Résumés de quelques documents EEF

H - Statistical analysis guidance for EEF evaluations

Sans auteurs.
Mars 2018
https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/evaluation-design

Ce document explicite les règles utilisées par EEF pour évaluer quantitativement l'effet d'une intervention (*impact evaluation*).
Les résultats des études ne sont pas analysés individuellement mais bien de façon plus globale et doivent, par exemple, pouvoir être comparés d'une étude à l'autre. C'est avec cet objectif qu'a été rédigé le guide sur l'analyse statistique des résultats des évaluations conduites par EEF.
Pour toutes les évaluations, un plan d'analyse statistique (*statistical analysis plan, SAP*) est publié conjointement au protocole de l'évaluation (au plus tard 3 mois après la randomisation). Les principes clés sont les suivants :

1. Les analyses doivent **refléter le design de l'étude et la procédure de randomisation** (les caractéristiques initiales des élèves de chacun des groupes conduisant parfois à certains traitements des données particuliers).
2. Les analyses doivent être menées sur la base d'une **procédure « en intention de traiter »** (tous les élèves inclus au départ dans l'étude doivent être inclus dans l'analyse, quelque soit leur statut en fin d'expérience), ce qui conduit à une estimation de l'effet de l'intervention la plus conservative possible.
3. Le **niveau initial des élèves doit être pris en compte** dans un modèle de régression (par exemple une ANCOVA) en tenant compte des clusters, ce qui a comme effet d'augmenter la puissance et la précision de l'étude ; pour faciliter la comparaison des études entre elles, les évaluations ne doivent pas utiliser d'autres covariables que les scores pre tests, le statut du groupe et les caractéristiques du design de l'étude.
4. Il faut **tenir compte des clusters**, par exemple en utilisant une modélisation linéaire hiérarchique (*hierarchical linear modelling, HLM*) comme cela est conseillé par la What Works Clearinghouse (WWC) ; ne pas en tenir compte n'influe pas sur l'estimation ponctuelle qui reste juste, mais minore les erreurs standards des tailles d'effet ce qui augmenterait injustement le poids des études dans d'éventuelles méta-analyses.
5. Déterminer **la taille d'effet en utilisant la variance totale** en calculant le *g* de Hedges (quotient de la différence des moyennes ajustées aux scores pré test divisé par l'écart-type groupé de l'échantillon total).
6. L'**incertitude statistique** de chaque estimation de la taille d'effet doit être déterminée.
7. Calculer les **coefficients intra classes** pour les essais randomisés par classes et pour les scores post test (et les scores pré tests si possible).

La multiplicité des mesures et des inférences pouvant augmenter le risque de faux positifs, il est préférable pour les études de définir un résultat principal (*primary outcome*). Cela minimise également le risque de faux négatif en permettant d'estimer la taille d'échantillon nécessaire pour une étude suffisamment puissante. Des mesures composites (par exemple incluant des scores en maths et lecture) sont déconseillées. Dans les études d'efficacité, si le modèle logique et des preuves antérieures le soutiennent, les évaluateurs peuvent avoir besoin de plusieurs résultats pour évaluer l'impact d'une intervention. Par contre, pour les études d'efficacité en condition réelles, l'utilisation d'un résultat principal est vivement recommandée. Pour éviter de sous estimer les effets positifs d'une intervention en cas de non-conformité de l'expérience, des analyses complémentaires doivent être menées (par exemple si une intervention comprend une formation des enseignants et l'utilisation d'un

logiciel, la conformité de ces deux éléments doit être évaluée). L'attrition des données, inévitables dans toute étude contrôlée randomisée, doit être prise en compte dans les analyses.
On remarquera qu'aucune information sur la traduction des tailles d'effet en nombre de mois de progrès n'est donnée dans ce document. Cette information a été trouvée dans un modèle (*EEF evaluation report template 2019*, document I) : la même grille que celle utilisée pour le Toolkit est utilisée, à l'exception des deux premiers échelons (voir ci-dessous).

		0 mois de progrès	**1 mois de progrès**
Toolkit	**2018 (document A)**	-0,01 à 0,01	0,02 à 0,09
	2021 (document C)	-0,05 à 0,05	0,06 à 0,09
Etudes EEF (document I)		-0,04 à 0,04	0,05 à 0,09

J - Statement on statistical significance and uncertainty of impact estimates for EEF evaluations

Sans auteurs
Février 2020
https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/evaluation-design

Ce document établit des recommandations concernant l'évaluation et l'interprétation de la signification statistique qu'il convient d'associer à l'estimation de la taille d'effet calculée lors des études EEF et plus particulièrement des essais contrôlés randomisés. L'évaluation de la validité interne d'une étude (avec l'établissement d'un niveau de confiance mesuré tel que précisé dans le document M) est aussi évoquée. Deux sources d'incertitude[a] (*uncertainty*) sont décrites : l'incertitude liée à l'échantillonnage (l'échantillon est prélevé aléatoirement dans le but de représenter une population) et l'incertitude liée à la répartition aléatoire des élèves entre les groupes traitements et contrôle (la répartition aléatoire assurant que seule l'intervention différencie les deux groupes). La structure des tests d'hypothèses utilisés en statistique et les concepts qui la fonde sont rappelés, ainsi que la définition de la valeur-p (ainsi que les interprétations erronées qui en sont parfois faites).
Sept recommandations sont proposées :

1. L'**incertitude des résultats doit être reconnue et publiée** et la **présentation simplificatrice dichotomique des valeurs-p doit être abandonnée**.
2. Se concentrer sur une **interprétation scientifique et pratique** ; la définition d'un seuil (souvent la valeur 0,05) ne permet pas d'aller dans ce sens et c'est la taille d'effet qui doit rester le centre d'intérêt des chercheurs.
3. Les **résultats doivent être interprétés en considérant plusieurs informations** comme la validité interne, l'incertitude statistique, la force de preuves existantes, la plausibilité du mécanisme causal, la preuve de la qualité de la mise en œuvre de l'intervention, et des considérations contextuelles entre autres.
4. Utiliser un **langage précis et clair** qui prenne en considération les conditions à partir desquelles les statistiques sont calculées puis interprétées.
5. Calculer et **interpréter la valeur-p comme une valeur continue** et éviter les décisions basées sur un seuil arbitrairement défini (par exemple 0,05) et ne plus employer des phrases comme « il n'y a pas de différence entre les deux groupes ».
6. Changer la dénomination « intervalles de confiance » par « **intervalle de compatibilité »** (car compatible avec le modèle statistique utilisé), qui ne doivent pas être interprétés de façon dichotomique.
7. Utiliser d'**autres statistiques** qui permettraient d'interpréter les résultats (en évitant encore une fois une interprétation dichotomique), comme par exemple les valeurs-p permutées

[a] Peut-être à traduire plutôt par « erreur », voire même hasard.

(*permuted p-value*) qui ne tiennent pas compte de l'incertitude liée à l'échantillonnage et qui donc ne permettent pas d'inférer au-delà de l'échantillon.

Pour toutes ces raisons, EEF ne décrit pas les résultats comme étant statistiquement significatif ou non.

Remarque : ce document traite essentiellement du cas de l'absence de significativité statistique (qui semble être le réel sujet, voir document K).

K - Statistical uncertainty in Randomised Controlled Trials

Sans auteurs
Date probable (indiquée sur le site internet) : 2018
https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/evaluation-design

Ce document fait état des interrogations de EEF sur le traitement et l'importance à accorder à ce que l'on nomme traditionnellement la significativité statistique associée à une estimation (ici une taille d'effet calculée à partir d'un échantillon après une intervention dans le cadre classique d'un essai contrôlé randomisé, ou ECR).
« *Les ECR conduits par EEF sont généralement suffisamment dimensionnés (taille d'échantillon), et donc lorsque le système d'évaluation du niveau de confiance a été développé, on s'attendait à ce qu'il soit rare qu'une étude EEF obtienne une évaluation élevée de son niveau de confiance, une taille d'effet d'une ampleur importante, mais que celle-ci ne soit pas statistiquement significative telle que traditionnellement énoncé en statistiques. Maintenant que nous avons publié les résultats d'un grand nombre d'essais contrôlés randomisés, nous avons constaté qu'il existe un certain nombre d'études qui aboutissent à un résultat positif d'un mois de progrès ou plus (ce qui pourrait être une ampleur importante selon le contexte), un niveau de confiance au moins égal à 3 (sur 5), mais pour lesquels l'estimation de la taille de l'effet n'est pas statistiquement significative (en utilisant le seuil de valeur p le plus populaire de 5 %).* »

En 2020, EEF publie finalement ses recommandations sur cette question (document J).

L – Research paper n°2_Standard deviation as an outcome on interventions : a methodological investigation

Auteurs : Peter Tymms, Adetayo Kasim
Février 2018
https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/eef-evaluation-reports-and-research-papers/methodological-research-and-innovations/standard-deviation-as-an-outcome-on-interventions

Ce document interroge l'utilisation de la différence de la variance entre deux groupes (le groupe traitement et le groupe contrôle) comme potentiel résultat permettant d'évaluer l'effet d'une intervention. Deux raisons sont mises en avant : la première tient à l'objectif même d'une intervention qui est de réduire les inégalités, cette réduction devant logiquement conduire à une diminution de la variance ; la seconde concerne l'identification des causes permettant d'expliquer et de détailler l'effet d'une intervention (qui peut être plus efficace sur un groupe d'élèves que sur un autre).

Deux approches sont analysées : l'ANCOVA (dont l'utilisation pour comparer des moyennes a déjà bien fait ses preuves) et l'analyse de la différence des différences. Ces deux méthodes ont été formalisées puis testées sur un ensemble de 16 études conduites par EEF. Certains résultats diffèrent selon la méthode employée, mais à chaque fois qu'un résultat statistiquement significatif est apparent, cela

est dû à une augmentation de la variance du groupe intervention en fin d'expérience quand on la compare à celle du groupe contrôle.

Devant le manque de connaissances théoriques, les chercheurs proposent en l'état actuel des connaissances de publier comme mesure principale de l'effet d'une intervention la différence des moyennes des groupes (taille d'effet) mais tout de même de l'accompagner d'une comparaison des deux différences de variances pour chacun des groupes (traitement et contrôle) sous la forme d'un pourcentage d'augmentation (ou de réduction) en utilisant une ANCOVA.

M - Classification of the security of findings from EEF evaluations

Sans auteurs.
Version 2.0 – Juillet 2019
https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/evaluation-design

Pour chaque évaluation (donc chaque étude primaire), ce système permettant de classer sur une échelle allant de 1 (le plus mauvais classement) à 5 la confiance associée aux résultats (la plupart du temps une taille d'effet convertie en nombre de mois de progrès réalisés). Il évalue avant tout la qualité interne de l'étude (force de la preuve telle que mesurée par l'étude) plutôt que sa qualité externe (possibilité d'inférer au-delà de l'échantillon analysé). Cette classification concerne le résultat principal uniquement.
Quatre critères vont être considérés :

1. **Le design de l'étude** : seuls les essais contrôlés randomisés sont classés au niveau 5. En permettant une attribution aléatoire des élèves dans les groupes contrôle et traitement, ce type d'étude permet de se libérer de la question des facteurs de confusion. Un classement supérieur ou égal à 3, correspondant à des études de comparaison de groupes qui tiennent compte de tous les facteurs de confusion observables est requis pour les études EEF (sauf circonstances extraordinaires).
2. **La taille d'effet minimal détectable** : c'est la taille d'effet minimale que l'étude est susceptible de détecter (cela revient à fixer une taille de l'échantillon minimum). Elle doit être inférieure ou égale à 0,2 pour le niveau 5.
3. **L'attrition** : le niveau global de la perte de données (groupe contrôle et groupe traitement confondus) est mesuré au niveau des élèves, quel que soit le niveau de l'attribution aléatoire. Elle doit être inférieure à 10 % pour le niveau 5.
4. **Menaces à la validité interne** : 7 menaces sont examinées (facteurs de confusion, présence d'interventions concurrentes associées à l'intervention évaluée, effet Hawthorne[a], fidélité de la mise en œuvre, données manquantes, questions liées aux mesures des résultats, publication sélective des résultats).

Deux étapes sont suivies pour définir le niveau de confiance d'une étude.
Première étape : les trois premiers critères sont évalués indépendamment sur une échelle de 1 à 5, et le niveau global est alors la plus faible valeur de ces trois évaluations.
Seconde étape : le 4ème critère est alors considéré, et le niveau global précédemment défini peut être diminué de deux échelons au plus si des menaces importantes sont observées.

Dans les rapports d'intervention récents, un bilan est donné en annexe qui permet de comprendre rapidement le niveau de confiance calculé pour une étude. Le tableau de l'intervention *Increasing Competence and Confidence in Algebra and Multiplicative Structures (ICCAMS)* est donné comme exemple ci-dessous. Le niveau initial (obtenu après l'étape 1) a été diminué de 1 niveau pour des questions liées aux groupes contrôles (voir capture d'écran sous le tableau).

[a] Cet effet est aux études en sciences de l'éducation ce que l'effet placebo est aux études pharmacologiques.

Rating	Criteria for rating			Initial score		Adjust		Final score
	Design	MDES	Attrition					
5	Randomised design	<= 0.2	0-10%					
4	Design for comparison that considers some type of selection on unobservable characteristics (e.g. RDD, Diff-in-Diffs, Matched Diff-in-Diffs)	0.21 - 0.29	11-20%	4				
3	Design for comparison that considers selection on all relevant observable confounders (e.g. Matching or Regression Analysis with variables descriptive of the selection mechanism)	0.30 - 0.39	21-30%			Adjustment for threats to internal validity -1		3
2	Design for comparison that considers selection only on some relevant confounders	0.40 - 0.49	31-40%					
1	Design for comparison that does not consider selection on any relevant confounders	0.50 - 0.59	41-50%					
0	No comparator	>=0.6	>50%					

Threat 2: Concurrent Interventions	Moderate	Data from a subset of control schools (n=24) suggests that many schools engaged with other maths professional development approaches, including Mastery, which may have weakened experimental contrast. No information was collected from intervention schools about other programmes implemented before or alongside ICCAMS.

N - Cost evaluation guidance for EEF evaluations
Sans auteurs.
Décembre 2019
https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/evaluation-design

Dix principes doivent être suivis pour évaluer le coût d'une intervention :

1. Le coût doit être estimé en considérant que les établissements sont les décideurs. Les coûts supportés par d'autres parties (les parents, les bénévoles) seront reportés séparément.
2. Le coût doit être estimé pour l'intervention telle qu'elle a été mise en œuvre dans l'évaluation car ils sont associés à l'impact tel qu'il a été mesuré dans l'évaluation.
3. Le coût sera estimé en comparaison avec le groupe contrôle.
4. Les ressources allouées à la formation, préparation et délivrance de l'intervention devront être estimées (toujours en comparaison avec le groupe contrôle).
5. Ces coûts doivent inclure toutes les dépenses, comme les couts de recrutement, les assurances, …
6. Toutes les ressources doivent être évaluées au prix du marché.
7. Les couts doivent être présentés de manière à pouvoir tenir compte de l'inflation de façon à ce que les personnes intéressées puissent comparer les couts de plusieurs interventions évaluées à des moments différents.
8. Les coûts seront divisés en prérequis, coût de démarrage et coûts récurrents pour permettre aux établissements d'évaluer les coûts à chaque étape de la mise en œuvre.

9. Les coûts par élève seront calculés sur la base d'une intervention sur une durée de 3 ans.
10. La variabilité des coûts sera explorée par des analyses de sensibilité.

La classification des couts sur une échelle de 1 à 5 (1 pour les couts les plus faibles) n'est pas donnée dans ce document, les seuils étant sujets à modification d'une année sur l'autre (pour tenir compte de l'inflation et permettre de comparer plusieurs interventions entre elles). Le modèle du rapport de l'évaluation daté de 2019 également (document I, annexe A) donne cette grille :

Niveau (échelon)	Niveau (qualificatif)	Tranches (en livres anglaises) par élève et par an
1	Très faible	Inférieur à 80
2	Faible	Entre 80 et 200
3	Modéré	Entre 200 et 700
4	Élevé	Entre 700 et 1200
5	Très élevé	Supérieur à 1200

On notera que le seuil fixé ici à 200 £ pour les études primaires, et qui sépare les niveaux 2 et 3, est fixé à 220 £ pour le Toolkit (document C publié en 2021, soit 2 ans plus tard).

O - Implementation and process evaluation guidance for EEF evaluations
Sans auteurs.
Aout 2019
https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/evaluation-design

Ce document fait référence au Handbook (voir document suivant) et recommande un ensemble de principes pour les évaluateurs EEF quand ils planifient, conduisent et rapportent l'**évaluation de la mise en œuvre et du processus** (*implementation and process evaluation, IPE*) d'une étude EEF portant elle-même sur l'évaluation d'une intervention en milieu scolaire. Cette évaluation de la mise en œuvre se fonde sur les pratiques recommandées pour des recherches pragmatiques s'appuyant sur des méthodes mixtes (avec des méthodes d'évaluation quantitatives et qualitatives). Les 5 principes importants suivants doivent être suivis :

1. **Intégration** : les évaluations de la mise en œuvre et du processus d'une part, et les **évaluations d'impact** (*impact evaluation, IE*) d'autre part, doivent être complémentaires et intégrées.
2. **Modèle logique** : le design de ces évaluations et leur cohérence et complémentarité avec les évaluations d'impact doit s'appuyer sur un modèle logique qui inclue des éléments clés spécifiques, est examiné à la lumière des données de l'évaluation et doit être révisé si besoin.
3. **Conformité, fidélité et pratique usuelle** : toutes les évaluations d'impact devraient inclure des mesures sur la conformité, la fidélité et les pratiques usuelles en plus de toutes autres dimensions pertinentes concernant la mise en œuvre.
4. **Pertinence** : le design d'une évaluation de la mise en œuvre et du processus doit être adapté à chaque étude en fonction notamment du type de l'étude (étude pilote, étude d'efficacité ou étude d'efficacité en conditions réelles).
5. **Pré-spécification** : le design de l'évaluation de la mise en œuvre et du processus devrait être planifié et pré-spécifié de façon transparente (rédaction d'un protocole d'évaluation qui doit préciser les méthodes et mesures employées, donner des informations sur les participants et les échantillons analysés, les approches visant à limiter les biais).

Modèle logique : représentation visuelle des entrées (ressources nécessaires) du programme ou de l'intervention, des activités (actions, processus, ressources), des sorties (résultat direct des activités), des résultats (intentionnels ou non) à court ou long terme (changement spécifique dans les

connaissances, attitudes, …) et des mécanismes causals (comprenant aussi les médiateurs) sous-jacents. Dans le cas d'un résultat négatif ou nul, un bon modèle logique devrait permettre aux évaluateurs de faire la distinction entre un échec de la théorie, une mise en œuvre défaillante et une méthodologie (méthode d'évaluation) inadéquates.

Conformité : mesure l'intensité avec laquelle les ingrédients critiques d'une intervention sont délivrés à, et reçus par les participants ciblés.

Fidélité : mesure l'intensité avec laquelle une intervention est délivrée telle que prévue ou prescrite.

Remarque : la conformité met l'accent sur les sujets (les élèves) recevant l'intervention, la fidélité met l'accent sur l'intervention elle-même.

Le document suivant n'est pas cité dans la liste des référence (annexe 1), mais peut utilement compléter le document O.
Implementation and process evaluation (IPE) for interventions in education settings : An introductory handbook
Auteurs : Neil Humphrey, Ann Lendrum, Emma Ashworth, Kirsty Frearson, Robert Buck and Kirstin Kerr
Date probable : 2016 (donnée en référence dans le document N ; sur le site internet : 2019).
https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/evaluation-design

L'évaluation de l'impact, et plus généralement les essais contrôlés randomisés, cherchent à mesurer l'effet d'une intervention, alors que l'**évaluation de la mise en œuvre et du processus** (*Implementation and process evaluation, IPE***)** permet de comprendre comment, pourquoi et pour qui les interventions sont efficaces (ou non).
L'objectif ici est de développer une compréhension plus complète non seulement de « ce qui marche », mais aussi comment, pourquoi, dans quels contextes ou circonstances, et pour qui une intervention est (ou non) efficace. L'IPE peut nous aider à répondre à ces questions importantes en fournissant aux chercheurs des outils théoriques, méthodologiques et analytiques qui permettent de mieux comprendre les processus et les mécanismes qui sous-tendent l'impact (ou son absence) des interventions éducatives.

Huit dimensions caractérisent une mise en œuvre :

1. La fidélité / adhésion : mesure dans laquelle les enseignants adhèrent à l'intervention
2. Le dosage : évaluation quantitative (dose et fréquence) de l'intervention telle qu'elle a été donnée ou reçue
3. Qualité : les différents composants de l'intervention ont-ils été correctement délivrés ?
4. Portée : définie par le nombre et la nature des participants ciblés
5. Réactivité : mesure dans laquelle les participants ciblés se sont engagés dans l'intervention
6. Spécificité de l'intervention : mesure dans laquelle l'intervention peut être distincte des autres pratiques existantes
7. Analyse des groupes contrôle/traitement : caractérisation du groupe contrôle (quelles pratiques mises en œuvre en l'absence de l'intervention)
8. Adaptation : la nature et la mesure des changements apportés à l'intervention durant l'expérience

Toutes ses dimensions doivent être prises en compte pour éviter les erreurs de type 3 (une cause est attribuée de façon erronée à un résultat). Cependant, il peut arriver que certaines de ces dimensions soient à considérer comme prioritaires.

Plusieurs facteurs affectent la mise en œuvre :

- **Besoins et caractéristiques initiaux** : quel est le niveau de besoin, de préparation et de capacité au changement de la cible (au sens large) de l'intervention ?
- **Système de soutien pour la mise en œuvre** : quelles stratégies et pratiques sont utilisées pour soutenir une mise en œuvre de bonne qualité ?
- **Environnement de la mise en œuvre** : quelles sont les caractéristiques contextuelles influentes associées à la cible (au sens large) de l'intervention ?
- **Les caractéristiques des enseignants** : quelles sont leurs perceptions de l'intervention, leurs attitudes et leurs caractéristiques psychosociales susceptibles de jouer un rôle dans la mise en œuvre de l'intervention ?
- **Les caractéristiques des interventions** : quelle forme prend l'intervention ?

De même que pour les dimensions, certains facteurs peuvent être considérés comme prioritaires.

Les principes suivants devraient être respectés par les évaluateurs EEF :

1. Utiliser la checklist des dimensions caractérisant la mise en œuvre.
2. Utiliser la checklist des facteurs affectant la mise en œuvre.
3. Utiliser des méthodes mixtes (recherche qualitative et quantitative) pour les IPE à moins qu'il existe une solide raison pour n'utiliser qu'une seule des deux approches.
4. Intégrer l'IPE et l'évaluation de l'impact (IE) et non pas les considérer comme des processus complétement dissociés.
5. Utiliser le modèle (document externe à EEF) proposant un cadre pour les descriptions et réplications d'interventions (aussi bien pour la mise en place que pour la rédaction du rapport).
6. Développer un modèle logique détaillé en partenariat avec l'équipe mettant en œuvre l'intervention
7. Après les réunions préparatoires à la mise en place, organiser un atelier d'analyse et d'évaluation de l'application de l'intervention avec l'équipe mettant en œuvre l'intervention.
8. Des données descriptives détaillées sur la mise en œuvre doivent être fournies et considérées comme indispensable pour la rédaction du rapport (afin d'assurer la validité des résultats).
9. Le développement, le pilotage et les propriétés psychométriques (statistiques de fiabilité inter-évaluateurs par exemple) de mesures quantitatives sur la mise en œuvre doivent être publiées (en annexe).
10. **Études pilotes** : cartographier les potentiels indicateurs probants associés à la théorie à chaque étape de l'intervention et utiliser l'évaluation de l'étude pilote pour collecter des données dans le but d'évaluer sa véracité.
11. **Essais randomisés** : les pratiques habituelles doivent être analysées et évaluées dans toutes les écoles de l'essai avant la répartition aléatoire et au moins encore une fois après l'intervention (par exemple au post test).
12. **Essais randomisés** : les données concernant les pratiques habituelles doivent être analysées de manière à repérer des évolutions qui auraient lieu durant l'intervention et qui pourraient affecter l'interprétation des résultats de l'étude.
13. **Essais randomisés** : les adaptations de l'intervention et la mesure dans laquelle elles sont conformes à l'intervention telle qu'initialement prévue doivent être examinées.
14. **Essais randomisés** : les analyses de sous-groupes devraient être spécifiées a priori et le choix des sous-groupes soutenu par la théorie sous-jacente à l'intervention.
15. **Essais randomisés** : les plans d'analyses quantitatives doivent être clairement précisés dans le protocole de l'étude.
16. **Essais randomisés** : quand des données quantitatives sur la mise en œuvre sont disponibles, des analyses devraient être conduite pour examiner la relation entre la variabilité de la mise en œuvre et les résultats de l'intervention.

17. **Études d'efficacité** : quand cela est pertinent et faisable, des analyses par protocole doivent accompagner les analyses en intention de traiter.
18. **Études d'efficacité en conditions réelles** : les variations contextuelles introduites dans ces études doivent être clairement documentées et quand cela est pertinent et faisable, des analyses pour explorer les associations contexte – mise en œuvre et/ou contexte – mise en œuvre – résultats doivent être mis en œuvre.

P - Individual participant data meta-analysis of the impact of EEF trials on the educational attainment of pupils on Free School Meals: 2011 - 2019
Auteurs : Bilal Ashraf, Akansha Singh, Germaine Uwimpuhwe, Tahani Coolen-Maturi, Jochen Einbeck, Steve Higgins and Adetayo Kasim
Mai 2021
https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/eef-evaluation-reports-and-research-papers/syntheses-of-eef-evaluations/meta-analysis-of-the-impact-of-eef-trials-on-fsm-pupils-2011-2019

Les études conduites par EEF depuis 2011 ont fait l'objet de plusieurs méta-analyses dans le but de répondre à trois questions de recherche :

1. Les interventions ont-elles conduit à une amélioration des compétences et connaissances en anglais et en mathématiques des élèves éligibles au FSM[a] ?
2. Quelles sont les catégories d'études ou d'interventions associées à une amélioration des compétences et connaissances en anglais et en mathématiques des élèves éligibles au FSM ?
3. Les élèves éligibles au FSM ont-ils plus ou moins progressé que les élèves non éligibles au FSM ?

Trois méthodes permettant de conduire des méta-analyses à partir des 82 études EEF éligibles (certaines études n'ont pu être exploitées, car par exemple n'avaient pas abouti à des résultats en anglais ni en maths) ont été explorées. Les deux premières suivent un modèle classique en deux étapes avec (1) le calcul de tailles d'effet pour chaque étude, suivi du (2) calcul d'une taille d'effet globale, en utilisant soit le modèle des effets fixes soit le modèle des effets aléatoires. On notera cependant que les auteurs utilisent un modèle de régression linéaire multiples avec comme données de départ les scores des élèves (et non des moyennes de groupes comme c'est le cas habituel quand on calcule un g de Hedges par exemple). La troisième, qui a été retenue par les auteurs, est une méta-analyse sur les données individuelles des participants ; il s'agit alors de calculer en une étape une taille d'effet globale à partir de l'ensemble des scores des élèves de toutes les études sélectionnées[b] (voir annexe 5 pour une présentation des principales méthodes de calcul). Pour chacun des sous-groupes d'études considérés, les indicateurs Q et I^2 évaluant l'hétérogénéité des tailles d'effet globales sont également calculés.
Une comparaison des résultats obtenus à partir des scores des élèves éligibles au FSM pour chacune de ces trois méthodes est présentées dans le tableau 6 (les scores ont été standardisés ; d'autres résultats pour des scores bruts sont donnés dans le tableau 2 en annexe 2 de ce document P).

[a] Free School Meals
[b] On notera que finalement le modèle retenu est un modèle simplifié qui revient dans les faits à calculer une taille d'effet globale en deux étapes (les équations pour chacun des 3 modèles sont données dans le document) ; les tailles d'effet de chacune des études sont présentées de façon classique par un graphique en forêt.

Table 6: Overview of pooled effect size from IPD meta-analysis and two-stage fixed-effect (FE) and random-effect (RE) models using standardised outcome data.

	IPD	Two-stage Fixed-effect	Two-stage Random-effect
Outcome	Pooled ES	Pooled ES	Pooled ES
Literacy	0.06 (0.03, 0.08)	0.02 (0.00, 0.04)	0.03 (0.01, 0.06)
Maths	0.00 (-0.03,0.04)	0.01 (-0.01, 0.03)	0.00 (-0.02, 0.02)

Conformément aux recommandations du rapport sur la présentation de la précision des estimations (document J), les « intervalles de confiance » sont devenus des « intervalles de compatibilité»[a] et il n'est plus mentionné qu'un résultat est (ou non) statistiquement significatif[b]. On ne peut pas ne pas remarquer que les tailles d'effet globales calculées sont toutes très faibles (inférieures ou égales à 0,05) et que pratiquement aucune n'est (si en reprend la terminologie usuelle) statistiquement significative. Aucune interprétation de l'ampleur de ces tailles d'effet globale n'est proposée (elles ne sont pas traduites en nombre de mois de progrès) et les seuls commentaires concernent leur signe (soit positif, soit négatif). Seul le classement par ordre décroissant des tailles d'effet pour chacun des sous-groupes concernés permet aux auteurs de comparer les interventions les unes aux autres et d'en caractériser certaines comme prometteuses.
Pour les 48 tailles d'effet calculées à partir de scores obtenus par les élèves en mathématiques, on obtient une taille d'effet globale de 0,00 [-0,03 ; 0,04]. Les auteurs ont classé les tailles d'effet par ordre décroissant ce qui leur a permis d'en caractériser quatre comme prometteuses (tableau 1).

Interventions	**ES**	**Intervalle de** compatibilité
Powerful Learning Conversations	0,31	[-0,25 ; 0,98]
Dialogue Teaching	0,16	[0,03 ;0,29]
Improv Num and Lit KS 2	0,13	[-0,10 ; 0,36]
Act, Sing, Play 1	0,12	[-0,10 ; 0,35]

Tableau 1 : taille d'effet et intervalles de compatibilité des 4 interventions prometteuses.

Aucune des tailles d'effet globale calculées pour des sous-groupes d'études (par niveau d'étude[c], par type d'intervention[d], par design d'études[e], par type de scores[f]) en utilisant des résultats obtenus par les élèves en mathématiques n'est statistiquement significative. Aucune information n'est donnée sur la définition de ces catégories ni sur la manière dont les études ont été codées (c'est-à-dire répartie dans les différentes catégories). Les résultats sont rassemblés dans le tableau 2.

Scores primaires 0,01 [-0,02 ; 0,05]		**Scores secondaires** -0,07 [-0,15 ; 0,00]	
KS1 0,02 [-0,07 ; 0,11)	**KS2** -0,01 [-0,04 ; 0,03]	**KS3** 0,01 [-0,09 ; 0,12]	**KS4** 0,02 [-0,03 ; 0,07]

[a] Mis à part quelques apparitions en tout début de texte.
[b] Mis à part là encore d'une apparition p.39.
[c] KS1, KS2, KS3 et KS4
[d] Intervention individuelle, en petit groupe, sur la classe entière, sur l'école entière
[e] Etudes multisites ou études par randomisation de clusters
[f] Score primaire (de première importance) et score secondaire. Parfois des interventions en écriture / lecture ont pu donner lieu à une évaluation secondaire en mathématiques par exemple.

Individuel (*one-to-one*) 0,04 [-0,04 ; 0,12]	**Petits groupes (*small group*)** -0,04 [-0,11 ; 0,03]	**Classe entière (*whole class*)** -0,01 [-0,06 ; 0,05]	**Ecole entière (*whole school*)** 0,02 [-0,02 ; 0,07]
Etude contrôlée randomisée par grappes (*Cluster RCT ou CRT*) 0,01 [-0,03 ; 0,05]		**Etude multisite (*mutlisite trial, MST*)** 0,00 [-0,06 ; 0,06]	

Tableau 2 : analyses de sous-groupes pour les 48 études EEF sur les mathématiques (tailles d'effet et intervalles de compatibilité).

Les auteurs concluent à nouveau en repérant les interventions ayant les plus grandes tailles d'effet dans chaque sous-groupe. Enfin l'indice d'hétérogénéité I^2 calculé pour chacun des sous-groupes n'est jamais évoqué dans le texte. La valeur de cet indice, toujours égale à zéro, laisse à penser que les variations inter-études ne représentent qu'une très faible part de la variation totale.
Enfin, la différence entre l'effet de l'intervention sur les élèves éligibles au FSM d'une part et sur les élèves non éligibles au FSM d'autre part a été mesurée en calculant là aussi une taille d'effet. En ce qui concerne les compétences mathématiques, elle est égale à -0,01 et son intervalle de compatibilité est [-0,04 ; 0,02]. Les auteurs concluent en affirmant que bien que l'écart estimé entre ces deux groupes d'élèves soit négatif et que cette analyse n'ait pas permis de montrer qu'il ait diminué (ce qui reste l'objectif principal d'EEF), il n'en reste pas moins qu'aucun élément probant ne vient appuyer la thèse contraire.

Annexe 7. Résumés des études correspondants aux 13 tailles d'effet calculées à partir des résultats de SLAVIN

Copié/collé des vignettes des études de SLAVIN citées dans le Toolkit, collaborative teaching, sujet : maths. Les références ne sont pas complètes (il manque la liste des auteurs, les numéros de volume, de pages).
Note : *Standard error* = SE= (*upper confidence interval – lower confidence interval*)/(2*1,96)
TAI = *team assisted individualization*
Voir l'annexe 4 pour convertir les niveaux américains en niveaux français.

Slavin (1979)
An Extended Cooperative Learning Experience in Elementary School.
NA
Student Team Learning instructional techniques (Teams-Games-Tournament, Student Teams-Achievement Divisions, and Jigsaw II) have been extensively researched in six-to-twelve-week classroom experiments and found to have many positive effects. This study examines the effects of the techniques when they are used as the primary instructional method for a full semester in language arts, math, and social studies. The subjects were 388 fourth- and fifth-grade students in five elementary schools, two of which served as experimental schools and three as controls. The experimental groups scored significantly higher on four of seven Comprehensive Test of Basic Skills (CTBS) subscales than did the control groups. No significant differences were found for affective variables measured, except that the experimental groups named significantly more "friends in school." (Author/MP)

Effect Size 0.022
Lower Confidence Interval -0.004
Upper Confidence Interval 0.048
Weight 1.837

Notes : non référencé par SLAVIN

Slavin (1981)
Cognitive and Affective Outcomes of an Intensive Student Team Learning Experience
The Journal of Experimental Education
Three student team learning methods, Student Teams-Achievement Divisions (STAD), Teams-Games-Tournament (TGT), and Jigsaw, have been found to have positive effects on such student outcomes as achievement, race relations, mutual concern, and self-esteem. These techniques involve students working in small teams to master academic materials. However, the three methods have always been evaluated one at a time, for only small parts of the school day. This study evaluated use of all three methods together, covering most of students' instructional day, to discover whether student team learning methods can be used to replace traditional methods. Fourth- and fifth-grade students were assigned to experimental or control treatments for a semester. Results indicated that the intensive use of student team learning methods was feasible and produced positive outcomes on student friendships, liking of school, self-esteem, and language and reading achievement.

Effect Size 0.098
Lower Confidence Interval -0.108
Upper Confidence Interval 0.304
Weight 1.763

Notes : non référencé par SLAVIN

Slavin (1982) 1_1
Student Teams and Mastery Learning: A Factorial Experiment in Urban Math Nine Classes.
NA
Mastery Learning and Student Team Learning are two widely used instructional methods designed to confront the problem of student diversity in group-paced instruction. This study evaluated Mastery Learning, Student Team Learning, and a combination, in 43 inner-city math nine classes over a full school year. Results indicated greater achievement on a standardized test for Team classes than non-Team classes, but no greater achievement in Mastery than non-Mastery classes. Differences in time use were suggested to explain treatment differences in student achievement. (Author/MP)

Effect Size 0.15
Lower Confidence Interval -0.072
Upper Confidence Interval 0.373
Weight 1.751

Notes : 43 classes de grade 9. Non référencé par SLAVIN

Slavin (1982) 1_2
Student Teams and Mastery Learning: A Factorial Experiment in Urban Math Nine Classes.
NA
Mastery Learning and Student Team Learning are two widely used instructional methods designed to confront the problem of student diversity in group-paced instruction. This study evaluated Mastery Learning, Student Team Learning, and a combination, in 43 inner-city math nine classes over a full school year. Results indicated greater achievement on a standardized test for Team classes than non-Team classes, but no greater achievement in Mastery than non-Mastery classes. Differences in time use were suggested to explain treatment differences in student achievement. (Author/MP)

Effect Size 0.178
Lower Confidence Interval -0.058
Upper Confidence Interval 0.414
Weight 1.741

Notes : 43 classes de grade 9. Non référencé par SLAVIN

Slavin (1983)
Combining Student Teams and Individualized Instruction in Mathematics: An Extended Evaluation.
NA
This study evaluated the achievement effects of the Team-Assisted Individualization (TAI) mathematics program over a 24-week period. Involved were 1,317 students in grades 3, 4, and 5, with 700 students in 31 classes receiving TAI instruction and a control group of 617 students in 30 classes receiving other mathematics instruction on the same objectives. Analysis of covariance was used to analyze the data, with achievement measured by the Mathematics Concepts and Applications and the Mathematics Computation subtests of the Comprehensive Test of Basic Skills. TAI classes gained more than control classes on each test at each grade level. The differences were statistically significant for grades 3 and 5 on the Computation subtest. On the Concepts and Applications subtest, differences were statistically significant for grade 4 and marginally significant for grade 5. In overall analyses, the TAI classes significantly exceeded control classes on both tests. (Author/MNS)

Effect Size 0.2
Lower Confidence Interval 0.117
Upper Confidence Interval 0.283
Weight 1.826

Notes : 1317 élèves du Maryland, grades 3-5 (primaire) dont 700 dans 31 classes traitement TAI et 617 dans 30 classes contrôle. Non référencé par SLAVIN

Slavin (1984)
Combining Cooperative Learning and Individualized Instruction: Effects on Student Mathematics Achievement, Attitudes, and Behaviors
The Elementary School Journal
While programmed instruction has not generally been found to increase mathematics achievement, the problems appear to lie more in managerial and motivational difficulties rather than in the theory of individualizing instruction. This study evaluated programmed instruction in mathematics using a system designed to solve these problems by having students work in heterogeneous teams and do all scoring themselves. This cooperative individualized program, called Team-Assisted Individualization, or TAI, was assessed in two field experiments in elementary schools. Students in the TAI classes in both experiments scored higher than control students (controlling for pretest and grade) on a standardized mathematics test, ***but not higher than a group that used the materials and student management but not teams***. Attitude and behavioral rating results followed the same general pattern.
Effect Size 0.195
Lower Confidence Interval -0.038
Upper Confidence Interval 0.427
Weight 1.744

Notes : TAI école élémentaire ; 2 expériences ; **rejetée par SLAVIN** (méta-analyse en primaire) car durée intervention < 12 semaines.

Slavin (1984)
Effects of Cooperative Learning and Individualized Instruction on Mainstreamed Students
Exceptional Children
This study examines the effects on mainstreamed academically handicapped students of an instructional method, Team Assisted Individualization (TAI), that combined cooperative learning with individualized instruction in mathematics. Eighteen classes (grades 3–5) in six schools were randomly assigned to one of three conditions TAI individualized instruction (II) without cooperative teams or control. The 117 academically handicapped students in these classes served as the subjects. The TAI and II methods both had significantly positive effects on the social acceptance of academically handicapped students by their nonhandicapped classmates, on their attitudes toward math, and on teacher ratings of their behavior. No achievement differences were found, although students as a whole (handicapped and nonhandicapped) in TAI and II classes achieved more than control students.
Effect Size 0.139
Lower Confidence Interval -0.381
Upper Confidence Interval 0.66
Weight 1.444

Notes : 18 classes de grade 3 – 5 dans 6 écoles ; 117 élèves handicapés ; intervention : TAI ; **rejetée par SLAVIN** (méta-analyse en primaire) car durée intervention < 12 semaines.

Slavin (1984)
Mastery Learning and Student Teams: A Factorial Experiment in Urban General Mathematics Classes
American Educational Research Journal
The mathematics achievement effects of principal components of mastery learning and student team learning were evaluated in a year-long randomized experiment in urban ninth grade general mathematics classes. The components were formative tests, corrective instruction for nonmasters, summative tests (Mastery), and practice in four-member heterogeneous teams and team rewards (Teams). A 2 X 2 (Mastery by Teams) factorial experiment compared Mastery, Teams, Teams + Mastery, and a control treatment. All methods used the same materials and schedule of teaching, worksheets, and test. A nested analysis of covariance indicated significant achievement main effects for Teams but not for Mastery. No Mastery by Teams or pretest by treatment interactions were found.
Effect Size 0.248
Lower Confidence Interval 0.019
Upper Confidence Interval 0.476
Weight 1.747

Notes : 588 élèves avec des données (1092 au départ) de grade 9 dans 16 écoles ; **sélectionnée par SLAVIN** (méta-analyse en secondaire, SLAVIN calcule ES = 0,23 pour STAD + mastery ; ES = 0,18 pour STAD seul et ES = 0,21 comme moyenne). ES calculée pour *cooperatif* versus contrôle avec g de Hedges sans tenir compte des scores prétests (recalculée dans le fichier Excel reprenant 4 études analysées par SLAVIN) ; p < 0,03 selon SLAVIN et ici avec les données EEF, on a z = g/SE = 2,13 donc p = 0.0334. STAD = *Student Teams-Achievement Divisions*

Slavin (1984)
Effects of team assisted individualization on the mathematics achievement of academically handicapped and nonhandicapped students.
Journal of Educational Psychology
Evaluated the achievement effects of an individualized mathematics program designed to solve problems of management, direct instruction, and incentives common to earlier attempts at individualization of instruction. Team Assisted Individualization (TAI), a mathematics program that combines individualized instruction, cooperative learning teams, and direct instruction, was compared to control methods in a 24-wk experiment involving 1,371 students in 59 3rd-, 4th-, and 5th-grade classrooms. Standardized mathematics computations and concepts and application scales served as dependent variables and covariates. Random-effects nested analyses of covariance indicated statistically significant treatment effects favoring TAI for mathematics computations but not concepts and applications. However, individual-level analyses found significant treatment effects for both variables for the full sample and for a subsample of 63 academically handicapped Ss. No Treatment by Handicap interactions were found. (26 ref) (PsycINFO Database Record (c) 2016 APA, all rights reserved)

Effect Size 0.151 SE = (0,257- 0,044)/2*1.96 = 0,054
Lower Confidence Interval 0.044
Upper Confidence Interval 0.257
Weight 1.818

Notes : 1371 élèves du Maryland, grades 3-5 (primaire) dont 717 en traitement TAI et 646 en contrôle dans 59 classes. Taille d'effet et erreur standard recalculées pour les mesures post test (sans tenir compte des scores prétest) en calculs (*computation*) donnent les mêmes résultats que EEF. **Sélectionnée par SLAVIN** (méta-analyse en primaire).

Slavin (1984) 1_1

Team Assisted Individualization: Cooperative Learning and Individualized Instruction in the Mainstreamed Classroom

Remedial and Special Education

This paper summarizes research on a mathematics program that combines cooperative learning and individualized instruction to accommodate the social and academic needs of mainstreamed academically handicapped students and their nonhandicapped classmates. This program, Team Assisted Individualization (TAI), was found to improve the social and academic behavior of these students, and in one of two studies, to increase their mathematics achievement more than traditional methods. Positive effects on the achievement and behavior of nonhandicapped students were also found. The implications of these findings for instruction in mainstreamed classrooms are discussed.

Effect Size 0.275
Lower Confidence Interval 0.042
Upper Confidence Interval 0.508
Weight 1.743

Notes : on ne sait pas pour quels élèves est calculée la taille d'effet (élèves non handicapés ou élèves handicapés ?). Pas d'information sur les grades dans le résumé. Non référencé par SLAVIN

Slavin (1984) 1_2

Team Assisted Individualization: Cooperative Learning and Individualized Instruction in the Mainstreamed Classroom

Remedial and Special Education

This paper summarizes research on a mathematics program that combines cooperative learning and individualized instruction to accommodate the social and academic needs of mainstreamed academically handicapped students and their nonhandicapped classmates. This program, Team Assisted Individualization (TAI), was found to improve the social and academic behavior of these students, and in one of two studies, to increase their mathematics achievement more than traditional methods. Positive effects on the achievement and behavior of nonhandicapped students were also found. The implications of these findings for instruction in mainstreamed classrooms are discussed.

Effect Size 0.151
Lower Confidence Interval 0.044
Upper Confidence Interval 0.257
Weight 1.818

Notes : on ne sait pas pour quels élèves est calculée la taille d'effet (élèves non handicapés ou élèves handicapés ?). Pas d'information sur les grades dans le résumé. Non référencé par SLAVIN

Slavin (1985) 1_1

Effects of Whole Class, Ability Grouped, and Individualized Instruction on Mathematics Achievement

American Educational Research Journal

Achievement and attitudinal effects of three mathematics instruction methods directed in varying degrees toward accommodating diversity in student performance levels were compared in two randomized field experiments. Treatments included an individualized model, Team Assisted Individualization (TAI) an ability grouped model, Ability Grouped Active Teaching (AGAT) a group-paced model, the Missouri Mathematics Program (MMP) and, in Experiment 2 only, untreated Control classes. Analysis of Comprehensive Test of Basic Skills (CTBS) Computations scores adjusted

for pretests indicated that in both experiments, TAI and AGAT exceeded MMP. TAI, AGAT, and MMP also exceeded Control. No effects on CTBS Concepts and Applications were found, and there were no treatment by prior achievement interactions on either scale. Effects on Liking of Math Class and Self-Concept in Math generally favored TAI.
Effect Size -0.789
Lower Confidence Interval -1.073
Upper Confidence Interval -0.505
Weight 1.701

Notes : **sélectionnées par SLAVIN** (méta-analyse primaire). 345 élèves dans 16 classes. Ce résultat ne correspond à aucun des résultats proposés par SLAVIN.

Slavin (1985) 1_2
Effects of Whole Class, Ability Grouped, and Individualized Instruction on Mathematics Achievement
American Educational Research Journal
Achievement and attitudinal effects of three mathematics instruction methods directed in varying degrees toward accommodating diversity in student performance levels were compared in two randomized field experiments. Treatments included an individualized model, Team Assisted Individualization (TAI) an ability grouped model, Ability Grouped Active Teaching (AGAT) a group-paced model, the Missouri Mathematics Program (MMP) and, in Experiment 2 only, untreated Control classes. Analysis of Comprehensive Test of Basic Skills (CTBS) Computations scores adjusted for pretests indicated that in both experiments, TAI and AGAT exceeded MMP. TAI, AGAT, and MMP also exceeded Control. No effects on CTBS Concepts and Applications were found, and there were no treatment by prior achievement interactions on either scale. Effects on Liking of Math Class and Self-Concept in Math generally favored TAI.
Effect Size 0.669
Lower Confidence Interval 0.396
Upper Confidence Interval 0.942
Weight 1.71

Notes : **sélectionnées par SLAVIN** (méta-analyse primaire). 345 élèves dans 16 classes. Ce résultat ne correspond à aucun des résultats proposés par SLAVIN.

Annexe 8. Résumés des études sélectionnées par SLAVIN et EEF

Résumés des études sélectionnées par la méta-analyse de EEF pour le thème apprentissage collaboratif (*collaborative learning*) du Toolkit et également par la méta-analyse sur l'enseignement des mathématiques au secondaire de SLAVIN.

Barbato (2000)

Policy implications of cooperative learning on the achievement and attitudes of secondary school mathematics students

ProQuest Dissertations and Theses

This study compared the effects of traditional and cooperative methods of teaching on the mathematical achievement, attitudes, and course enrollment plans of 10th-grade students. Further, an attempt was made to ascertain whether mathematics achievement test scores, attitude toward mathematics, and plans for enrollment in mathematics courses differed for gender. Two hundred eight subjects from a suburban high school participated in the study. Approximately half were assigned to a mathematics course taught using traditional lecture, including a question-and-answer period and individualized assignments. The same instructor taught the other half using a cooperative learning method. This involved engaging students in interactive groups to discuss ideas, test conjectures, and solve problems. All classes were heterogeneously grouped for ability and mathematical achievement. Students completed a pretest and posttest measure of mathematical achievement using the New York State Integrated Math Test for Course I and II, respectively. Attitudes toward mathematics as a subject were measured at the beginning and end of the course using The Motivational Survey. At the beginning and at the end of the course, students identified the math courses they planned to take in their junior and senior years. The analysis of posttest data suggested that the class taught cooperatively had significantly higher mathematics achievement and showed more positive attitudes toward mathematics than the traditional group. There were no gender differences for mathematics achievements. Males did score higher than females for two subscales on The Motivational Survey, extrinsic motivation and self-efficacy motivation. No significant difference between males and females existed for intrinsic motivation. Regarding plans to enroll in advanced mathematics courses, a chi-square analysis of the pretest data revealed significant differences. More males than females planned to enroll in one of the highest level courses offered to junior and senior students. However, on the posttest no significant differences were found for gender. This was due to the shift in enrollment plans among females in the cooperative group. On the posttest, females in the cooperative group indicated a desire to enroll in higher level courses compared to their pretest selections. The results of this study support recent evidence that the achievement gap between males and females is narrowing, yet gender differences continue to exist in terms of course enrollment plans and attitudes toward mathematics.

Effect Size 0.841
Lower Confidence Interval 0.557
Upper Confidence Interval 1.125
Weight 0.523

Nichols (1996)

The Effects of Cooperative Learning on Student Achievement and Motivation in a High School Geometry Class

Contemporary Educational Psychology

In this study, the effects of a form of cooperative group instruction (Student Teams Achievement Divisions) on student motivation and achievement in a high school geometry class were examined. Eighty students were randomly assigned to either a control group receiving traditional instruction or one of two treatment groups receiving cooperative learning instruction. Geometry achievement was assessed using scores from the IOWA Test of Basic Skills and teacher-made exams. An 83-item

questionnaire was used as a pretest, posttest, and post-posttest assessment of efficacy, intrinsic valuing, goal orientation, and cognitive processing. Students in the cooperative treatment groups exhibited significantly greater gains than the control group in geometry achievement, efficacy, intrinsic valuing of geometry, learning goal orientation, and reported uses of deep processing strategies. The implications for cooperative group structures and motivation theory are discussed.
Effect Size 0.164
Lower Confidence Interval -0.275
Upper Confidence Interval 0.603
Weight 0.488

Reid (1992)
The Effects of Cooperative Learning with Intergroup Competition on the Math Achievement of Seventh Grade Students.
NA
There has been a large amount of research on the relative effects of cooperative, competitive, and individualistic learning on achievement and productivity. This paper reports a study designed to determine the effect of cooperative learning strategies on mathematics achievement of 7th graders. Based on school records indicating whether a student had participated in the cooperative learning strategies group or had received individualized/competitive instruction, 50 seventh-grade students were selected from an elementary school located in a low socioeconomic neighborhood in Chicago made up of all minority students. Scores on the mathematics portion of the Iowa Test of Basic Skills administered in the Spring of 1991 and 1992 were utilized to compare the mathematics achievement of the students in the two groups. T-tests performed on the pre-test and post-test data indicated that no differences existed in the groups prior to instruction, but that the cooperative learning groups performed significantly higher on the post-test. These findings confirmed results of similar studies. The paper concluded that cooperative group learning strategies are more effective in promoting mathematics achievement. (MDH)
Effect Size 0.649
Lower Confidence Interval 0.079
Upper Confidence Interval 1.219
Weight 0.452

Annexe 9. Glossaire

Attrition : Perte de sujets au cours de l'expérience qui entraîne une perte de données.

Étude par comparaison de groupes : étude expérimentale qui évalue quantitativement l'effet d'une intervention (ou traitement) en comparant deux groupes d'élèves : le groupe traitement et le groupe contrôle (qui ne subi pas l'intervention).

Essai contrôlé randomisé : étude expérimentale par comparaison de groupes dans laquelle les groupes sont constitués de façon aléatoire, soit au niveau des élèves, soit au niveau des classes répartis.

Étude primaire : études expérimentales qui évalue l'effet d'une intervention (ou traitement) sur des élèves scolarisés. Elles peuvent être incluses dans une méta-analyse.

Examen des preuves : synthèse systématique quantitative sur une problématique déclinée en plusieurs questions de recherche.

Intention de traiter (analyse en) : tous les élèves inclus au départ de l'étude doivent être inclus dans l'analyse statistique, y compris quand ils ont quitté l'établissement par exemple.

Intervention : Programme scolaire ou ensemble de pratiques qualifiés souvent de méthode pédagogique ou produit commercial (comme une application informatique), dont l'objectif est d'améliorer les compétences des élèves.

Méta-analyse : Synthèse systématique d'études quantitatives indépendantes répondant à une même question de recherche qui se traduit notamment par le calcul d'indicateurs mesurant l'effet d'une intervention avec comme objectif principal d'aboutir à une conclusion plus fiable ou plus solide que celle pouvant être tirée d'études distinctes. Une méta-analyse est une étude secondaire rétrospective ; les études incluses dans la méta-analyse sont des études primaires prospectives.

Méta méta-analyse : Synthèse systématique de méta-analyses indépendantes répondant à une même question de recherche ; on parle aussi de méta-analyse de second ordre ou secondaire.

Protocole : Document définissant à un stade précoce les méthodes suivies par les chercheurs pour réaliser une étude ou une expérience.

Rapport d'orientation : document à destination des praticiens qui donne des recommandations basées sur des éléments probants (résultats de méta-analyses notamment).

Synthèse systématique : compilation sur un sujet donné de toutes les études antérieures détectables (publiées ou non) puis évaluation de ces études au regard de critères définis *a priori*. Une synthèse peut être qualitative ou quantitative.

Taille d'effet : Différence standardisée des moyennes : écart entre les moyennes des scores des groupes traitement et contrôle divisé par l'estimation de l'écart-type des scores de la population.