Les essais contrôlés randomisés au Royaume-Uni : évaluer les interventions efficaces pour favoriser l'apprentissage des mathématiques

Nathalie ROQUES

Résumé

Depuis sa création en 2011, Education Endowment Foudation (EEF) a réalisé une centaine d'essais contrôlés randomisés (ECR) dans les établissements scolaires britanniques. Il s'agit pour cette fondation d'évaluer de façon scientifique l'impact d'interventions ciblées sur les apprentissages des élèves et de permettre aux acteurs en charge de l'enseignement de définir les actions à mettre en œuvre afin d'élever le niveau scolaire. Les évaluations de l'impact d'une intervention d'une part et de sa mise en œuvre d'autre part, sont encadrées par des procédures clairement définies par EEF et s'appuient sur le cadre plus vaste des études par comparaison de groupes. Leurs études d'efficacité en conditions réelles montrent que les élèves ont progressé en mathématiques au mieux de 3 mois (ce qui correspond à une taille d'effet égale à 0,26). Enfin, d'après les conclusions d'une méta-analyse conduite sur 48 résultats en mathématiques, les élèves de milieux socio-économiques défavorisés n'ont pas plus profité de ces interventions que les autres élèves. L'intérêt porté à ce type d'analyse par le Conseil Scientifique de l'Education Nationale (CSEN) souligne que ce pan de la recherche est probablement amené à prendre de l'importance en France dans le domaine des Sciences de l'Education.

Mots-clés : essai contrôlé randomisé (ECR), interventions, comparaison, taille d'effet, pratiques enseignantes

Abstract

Since its creation in 2011, Education Endowment Foundation (EEF) has carried out around 100 randomized controlled trials (RCTs) in British schools. The objective of this foundation is to scientifically assess the impact of targeted interventions on student learning and to enable those responsible for education to define the actions to be implemented in order to promote learning. The evaluations of the impact of an intervention on the one hand and of its implementation on the other hand, are framed by procedures clearly defined by EEF and are based on the broader framework of comparison group studies. Their effectiveness trials show that pupils have progressed in mathematics at best by 3 months (which corresponds to an effect size equal to 0.26). Finally, according to the conclusions of a meta-analysis conducted on 48 results in mathematics, students from disadvantaged socio-economic backgrounds did not benefit more from these interventions than other students. The interest shown in this type of analysis by the Scientific Council for National Education (CSEN) underlines the importance that this aspect of research is probably brought to play in France in the field of Educational Sciences.

Keywords: randomized controlled trial (RCT), interventions, comparison, effect size, teaching practices

Resumen

Desde su creación en 2011, Education Endowment Foundation (EEF) ha llevado a cabo alrededor de 100 ensayos controlados aleatorios (ECA) en escuelas británicas. El objetivo de esta fundación es evaluar científicamente el impacto de las intervenciones específicas en el aprendizaje de los estudiantes y permitir que los responsables de la educación definan las acciones a implementar para promover el aprendizaje. Las evaluaciones de impacto de una intervención por un lado y de su implementación por otro, se enmarcan en procedimientos claramente definidos por EEF y se basan en el marco más amplio de los estudios por comparación de grupos. Sus estudios de eficacia en condiciones reales muestran que los alumnos han progresado en matemáticas en el mejor de los 3 meses (lo que corresponde a un tamaño del efecto igual a 0,26). Finalmente, según las conclusiones de un metanálisis realizado sobre 48 resultados en matemáticas, los estudiantes de entornos socioeconómicos desfavorecidos no se beneficiaron más de estas intervenciones que otros estudiantes. El interés mostrado por este tipo de análisis por parte del Consejo Científico para la Educación Nacional (CSEN) subraya la importancia que probablemente este aspecto de la investigación está jugando en Francia en el campo de las Ciencias de la Educación.

Palabras clave: ensayo controlado aleatorizado (ECA), intervenciones, comparación, tamaño del efecto, prácticas docentes

C'est en 2011 que le Sutton Trust lance en Grande-Bretagne un nouvel organisme baptisé Education Endowment Foundation (EEF). La mission d'EEF est double : elle consiste à identifier de façon scientifique les interventions permettant d'élever le niveau général des élèves d'une part et de réduire les écarts entre les élèves de groupes socioéconomiques différents d'autre part[1]. Les enseignements tels qu'ils se pratiquent en France et au Royaume-Uni diffèrent en plusieurs points, que ce soit par exemple au niveau des méthodes pédagogiques ou des programmes scolaires. Mais les objectifs et les questionnements de ces communautés éducatives présentent, des deux côtés de la Manche, des points communs, et les éléments de réponse apportés par les travaux d'EEF, peu connus chez nous, pourraient utilement alimenter la réflexion des enseignants français. Les études conduites par EEF ont également été citées à plusieurs reprises par le Conseil Scientifique de l'Education Nationale (CSEN) (CSEN, 2023 ; Dehaene et Pasquinelli, n.d., p.17) dont la vocation est de proposer des recommandations aux enseignants. Dans ce contexte, comprendre les procédures mises en œuvre par les études quantitatives britanniques et analyser leurs résultats permettrait aux enseignants français de déterminer les atouts et les limites de ce type de recherche.

Nous sommes aujourd'hui en 2023, et 12 ans après sa création de nombreux projets ont été menés par EEF. Deux axes complémentaires peuvent être distingués : le premier concerne la rédaction de recommandations pragmatiques à destination des praticiens et fondées sur des méta-analyses réalisées par EEF à partir d'études scientifiques internationales. Le second consiste à concevoir et piloter des études par comparaison de groupes (ce sont des études quantitatives) de grande ampleur sur le territoire britannique et d'en analyser les résultats afin d'identifier les interventions les plus efficaces auprès des élèves. Nous nous intéresserons ici plus particulièrement à ce deuxième volet d'actions.

Les textes dont il sera question ci-dessous ont tous été téléchargés à partir du site internet https://educationendowmentfoundation.org.uk/ en aout 2022, à l'exception de *Statistical analysis guidance for EEF evaluations* (EEF, 2022) mis à jour en octobre 2022 et dont la version antérieure n'est actuellement plus disponible (le site internet d'EEF ne conserve aucune archive).

Présentation générale

L'objectif principal de cet article est de présenter à une large audience les grandes lignes des études par comparaison de groupes conduites par EEF et pour cette raison les aspects techniques calculatoires inhérents aux analyses statistiques ont été limités au maximum (des informations complémentaires peuvent être trouvées sur le site de l'auteure[2]). Les discussions qui interrogent le bien-fondé des analyses quantitatives dans le domaine des Sciences de

[1] https://educationendowmentfoundation.org.uk/about-us/how-we-work
[2] www.mathadoc.fr

l'éducation sont également volontairement laissées de côté ici, mais certains points méthodologiques seront commentés tout au long de ce texte.

Les études par comparaison de groupes menées par EEF sont la plupart du temps des essais contrôlés randomisés (ECR) : les élèves sont affectés aléatoirement soit dans le groupe intervention, soit dans le groupe contrôle, et les moyennes des scores obtenus après l'expérience pour chacun des groupes sont comparées. Elles sont donc conçues comme des expériences scientifiques et leur objectif principal consiste à valider une inférence causale. Les interventions sont en général des programmes constitués d'un ensemble de pratiques clairement identifiées (comme mettre en place des séances de soutien en petits groupes ou utiliser un dispositif pédagogique particulier).

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. 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 national est calculé. Si cette étude aboutit au calcul d'un effet négatif ou très faible (inférieur ou égal à 1 mois de progrès, voir plus loin), 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 franchi les étapes précédentes sont évaluées par une étude à plus grande échelle (*scale-up*) sur un large territoire.

Ce sont les études d'efficacité en conditions réelles qui permettent véritablement d'évaluer l'impact d'une intervention susceptible d'être mise en œuvre dans des conditions ordinaires sur un large territoire (les études à plus grande échelle sont quant à elles très rares), et c'est sur leurs résultats que nous allons porter notre attention.

Pour chaque étude, un rapport d'évaluation (*evaluation report*) est publié. Il inclut deux types d'évaluation :

- une évaluation de l'effet (*impact evaluation*) qui comprend un ensemble de résultats statistiques 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*) qui a comme objectif de montrer 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 derniers documents explicitent les méthodes définies *a priori* pour répondre à un certain nombre de questions et sont rédigés 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 qui 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 énumérés précédemment. Des modèles sont également disponibles : ce sont des fichiers Word qui peuvent être utilisés par les évaluateurs. Une présentation des éléments saillants de ces procédures est proposée ci-dessous. Elle sera suivie d'une description de l'ensemble des études EEF, puis les résultats statistiques publiés pour les études d'efficacité en conditions réelles qui ont comme sujet les mathématiques seront commentés.

Les procédures

Tous les textes qui explicitent les procédures mises en œuvre par EEF pour mener leurs études sur le terrain sont à télécharger sur le site EEF[3].

Trois indicateurs vont être calculés pour chaque intervention étudiée, qui permettent à l'internaute d'évaluer d'un simple coup d'œil l'efficacité d'une intervention : son effet sur les apprentissages des élèves, le niveau de preuve que l'on peut associer à cet effet et le coût de l'intervention. Les informations concernant l'évaluation de la mise en œuvre de l'intervention quant à elles ne sont pas synthétisées sous forme d'indicateurs chiffrés facilement identifiables et devront être recherchées dans le texte résumant les principaux résultats d'une évaluation.

[3] Cliquer sur *Project and evaluation* puis *Evaluation* puis *Evaluation guidance and resources*.

L'effet de l'intervention

Plusieurs résultats sont proposés qui permettent d'évaluer l'effet d'une intervention sur les connaissances et compétences des élèves mesurées par un score (EEF, 2022). L'analyse des scores se fait en conduisant une régression linéaire multiple qui modélise les liaisons entre la variable quantitative à expliquer (le score des élèves après l'intervention ou score posttest) et les variables explicatives (l'exposition des élèves au traitement mais aussi leur niveau de compétence avant l'intervention ou score prétest et parfois d'autres variables comme leur niveau socio-économique). L'affectation est le plus souvent réalisée au niveau des établissements et l'analyse au niveau des élèves ; dans ce cas, les scores des élèves d'un même établissement ne sont pas indépendants et des modèles multiniveaux sont utilisés par les chercheurs. Les statisticiens calculent ensuite une taille d'effet (*Effect Size, ES*) pour évaluer l'impact de l'intervention. Dans tous les essais randomisés d'EEF cette taille d'effet est le g de Hedges : il s'agit du quotient de la différence des moyennes des scores posttest ajustés aux scores prétest des deux groupes d'élèves divisée par une estimation de l'écart-type des scores des élèves. Cette taille d'effet est alors transformée en un nombre de mois de progrès qui est le nombre de mois dont un élève moyen du groupe intervention a progressé par rapport à un élève moyen du groupe contrôle (tableau 1). EEF considère que, dans la plupart des tests nationaux, le score d'un élève augmente en une année scolaire d'un écart-type ; donc 1 mois d'études équivaut à un progrès de 1/12 (soit 0,09) d'écart-type. Cette information a été publiée en 2018 dans un texte qui n'est plus disponible sur le site EEF. Ainsi pour une taille d'effet de 0,16, le nombre de mois de progrès sera égal à 2.

Tableau 1

Conversion d'une taille d'effet en nombre de mois de progrès (EEF, n.d.)

	Nombre de mois de progrès							
	-1 mois	0 mois	1 mois	2 mois	3 mois	4 mois	5 mois	6 mois
Taille d'effet minimale	-0,09	-0,05	0,06	0,10	0,19	0,27	0,36	0,45
Taille d'effet maximale	-0,06	0,05	0,09	0,18	0,26	0,35	0,44	0,52

Les analyses doivent être menées en intention de traiter, c'est-à-dire que tous les élèves inclus au départ dans l'étude doivent être inclus dans l'analyse, quel que soit leur statut en fin d'expérience, ce qui conduit à une estimation conservative de l'effet de l'intervention. Enfin, une attention particulière est portée à l'attrition des données comme à la multiplicité des mesures qui risque d'augmenter le risque de première espèce (et donc de faux positifs).

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

laisse la place à l'intervalle de compatibilité[4]. 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 plus lieu d'être. C'est dans un document publié en février 2020 que ces mises au point sont formulées (EEF, 2020). Elles font suite à un premier texte publié en 2018 (EEF, 2018) qui s'interrogeait sur le nombre non négligeable d'études EEF qui, suivant les procédures *ad hoc* d'un essai contrôlé randomisé, aboutissaient à des résultats positifs mais statistiquement non significatifs. Ces préconisations ont été intégrées à la nouvelle version de *Statistical analysis guidance for EEF evaluations* publiée en octobre 2022, mais elles n'ont pas été appliquées aux études dont nous présentons les résultats ici (et qui ont été publiées avant cette date).

On terminera sur ce sujet en évoquant la comparaison de l'évolution des variances (avant et après l'intervention) entre le groupe intervention et le 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 intervention). Cette question a fait l'objet d'une analyse menée sous l'égide d'EEF (Tymms et Kasim, 2018) mais pour l'instant, face au manque de connaissances théoriques, il est recommandé de conserver la taille d'effet comme résultat principal de l'effet d'une intervention, et de l'accompagner d'une comparaison de l'évolution dans le temps (avant/après l'intervention) des variances pour chacun des groupes intervention et contrôle. Il semblerait que cette dernière recommandation n'ait pas été suivie d'effet à l'heure actuelle.

Le niveau de preuve

Un niveau de preuve est associé au calcul de la taille d'effet et permet de classer l'étude sur une échelle allant de 1 à 5 (5 étant la meilleure évaluation). Quatre points sont considérés (EEF, 2019a) :

1. Le design de l'étude : les essais contrôlés randomisés constituent ici la référence ; en permettant une affectation aléatoire des élèves dans les groupes contrôle et intervention, ce type d'études permet d'éliminer la plupart des facteurs de confusion (ce sont des facteurs qui influent à la fois sur le score des élèves et l'exposition à l'intervention, faussant ainsi l'estimation du lien causal entre l'intervention et son effet sur les scores). Un classement supérieur ou égal à 3, correspondant à des études par comparaison de groupes qui tiennent compte de tous les facteurs de confusion observables, est requis pour les études EEF. Seuls les essais contrôlés randomisés sont classés au niveau 5.
2. La taille d'effet minimale 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). Elle doit être inférieure ou égale à 0,2 pour que l'étude soit classée au niveau 5.

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

3. L'attrition : c'est le niveau global de la perte de données (groupe contrôle et groupe intervention confondus) qui doit être mesuré au niveau des élèves quel que soit le niveau de l'affectation aléatoire. Elle doit être inférieure ou égale à 10 % pour que l'étude soit classée au niveau 5.
4. Les menaces à la validité interne : 7 menaces sont examinées. Il s'agit des facteurs de confusion, de la présence éventuelle d'interventions concurrentes associées à l'intervention évaluée, de l'effet Hawthorne (cet effet est aux études en Sciences de l'éducation ce que l'effet placebo est aux études pharmacologiques, et prend en compte l'influence de la motivation des élèves du groupe intervention sur les résultats de l'expérience), la fidélité de la mise en œuvre, le fait que des données soient manquantes, les questions relatives aux mesures des résultats et la publication sélective des résultats.

Le niveau de preuve d'une étude est attribué en suivant deux étapes. Dans un premier temps les trois premiers critères sont évalués indépendamment les uns des autres sur une échelle de 1 à 5, et la plus faible valeur de ces trois évaluations permet de définir un premier niveau global. Dans un second temps, le 4ème point est alors considéré, et le niveau global précédemment défini est susceptible de perdre un ou deux échelons si des menaces importantes sont identifiées.
Les résultats des études dont le niveau de preuve est inférieur ou égal à 2 devront être considérés avec beaucoup de précaution.

Le coût de l'intervention

Plusieurs principes doivent être suivis qui permettent d'évaluer le coût d'une intervention (EEF, 2019b). 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 partagés 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 faciliter la comparaison entre plusieurs interventions évaluées à des dates différentes. Une grille d'évaluation est proposée dans le modèle à utiliser pour rédiger le rapport d'évaluation et permet de classer le cout d'une intervention sur une échelle de 1 à 5 (correspondant à un coût très élevé) (EEF, n.d.)[5]. Pour être classée au niveau 1, le cout de l'intervention doit être inférieur à 80 £ (environ 90 €) par élève et par année.

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 sur les apprentissages des élèves, mais s'attache également à évaluer sa mise en œuvre. L'objectif est alors de déterminer les contextes ou les circonstances qui ont pu influer sur l'impact d'une intervention, par exemple en précisant son effet sur certains groupes d'élèves ou bien en identifiant les freins à sa mise en œuvre. Pour chaque évaluation de la

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

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 parler. La fidélité (l'intervention a-t-elle été délivrée telle que prévu ?), la conformité (les participants ont-ils reçu l'intervention telle que prévue ?) et les pratiques usuelles (ce sont les pratiques observées dans le groupe contrôle) doivent être analysées et décrites (EEF, 2019c). Cette évaluation doit être établie conjointement à l'évaluation de l'effet, et ce sont ces deux analyses qui constituent le cœur du rapport d'évaluation de l'intervention. Si l'évaluation de l'effet d'une intervention est un thème de recherche classique qui intéresse les scientifiques depuis des dizaines d'années, il en va autrement pour l'évaluation de sa mise en œuvre qui a nécessité une attention particulière de la part de chercheurs commissionnés par EEF pour que des progrès substantiels soient faits dans ce domaine (Humphrey *et al.,* 2016 ; Dawson *et al., 2018*).

Les études EEF

Quand on cherche des informations sur les études EEF[6], on visualise dans un premier temps une page qui recense l'ensemble de ces études sous forme de vignettes comportant les informations suivantes : le type d'études (étude pilote, étude d'efficacité, …), son niveau de complétion (en cours, achevée, …), le nom de l'intervention, le thème de l'étude (qui caractérise 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 preuve ni le coût ne sont indiqués à cette étape.

Sur les 242 études comptabilisées le 6 août 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 pu faire l'objet d'une étude pilote puis d'une étude d'efficacité par exemple).

Sur cette première page, les études peuvent être filtrées selon 7 critères, dont le niveau scolaire, le type d'études, le thème et le sujet abordés. Il s'agit pour ce dernier critère de classer la nature des résultats utilisés pour quantifier l'effet de l'intervention (par exemple des scores en mathématiques). Aucune définition des critères « thème » et « sujet » n'a été trouvée, il s'agit donc ici d'une interprétation personnelle. De la même façon, aucune procédure concernant le codage des études (c'est-à-dire l'association d'une étude à telle ou telle catégorie) n'est publiée. La répartition des études en fonction des sujets est proposée dans le tableau 2. Ni le niveau de preuve, ni l'effet (c'est le nombre de mois de progrès), ni le coût ne sont des critères de sélection.

[6] Cliquer sur *Project and evaluation* puis *Projects*.

Tableau 2

Répartition des études EEF par sujet et par type d'études

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

* : Certaines études n'ayant pas été associées à un sujet (par exemple les études sur l'impact de l'épidémie du covid 19), les totaux sont inférieurs à 242 et 172.

La répartition des études d'efficacité en conditions réelles achevées en fonction de l'effet estimé (en nombre de mois de progrès) est proposée dans le tableau 3. Pour les trois quarts de ces études (26 études sur 33) l'effet calculé est inférieur ou égal à 1 mois et aucune n'a abouti à un effet supérieur à 3 mois de progrès.

Tableau 3

Répartition des études d'efficacité en conditions réelles achevées selon leur effet

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

En cliquant sur une étude donnée on accède à un second niveau d'information, et c'est ici que l'on trouve notamment le niveau de preuve associé à l'effet calculé. Les rapports d'évaluation, protocoles et plans d'analyse statistique sont également téléchargeables sur cette page. La répartition des études d'efficacité en conditions réelles achevées en fonction du niveau de preuve est présentée dans le tableau 4. Presque 90% des études (29 sur 33) ont un niveau de preuve supérieur ou égal à 3.

Tableau 4

Niveau de preuve des études d'efficacité en conditions réelles achevées

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

Si on se limite aux études d'efficacité en conditions réelles achevées avec un niveau de preuve supérieur ou égal à 3, présentant un très faible coût (niveau 1) et un effet positif non nul, on obtient alors les six interventions présentées dans le tableau 5.

Tableau 5

Interventions de coût très faible, efficaces avec un niveau de preuve supérieur ou égal à 3

Interventions	Sujet	Effet	Niveau de preuve
Nuffield early language intervention	Litératie	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	Litératie	1 mois	3

En 2021, EEF publie un rapport (Ashraf *et al.,* 2021) proposant les résultats de plusieurs méta-analyses conduites sur ses propres études par comparaison de groupes publiées entre 2011 et 2019. Il s'agissait pour EEF de répondre aux trois questions de recherche suivantes :

1. Les interventions ont-elles conduit à une amélioration des compétences et connaissances en litéracie et en mathématiques des élèves éligibles au *Free School Meals* (FSM) (ces élèves sont susceptibles de bénéficier d'une aide alimentaire, et sont donc des élèves de milieux socio-économiques défavorisés) ?
2. Quelles sont les catégories d'études ou les types d'interventions associées à une amélioration des compétences et connaissances en litéracie d'une part et en mathématiques d'autre part, 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 ?

Les analyses ont d'emblée été séparées en fonction des compétences explorées (litéracie d'une part, mathématiques d'autre part). Toutes les études proposant des résultats exploitables ont été incluses, y compris des études pilotes et le niveau de preuve associé à l'effet n'a pas été pris en compte. La première étape habituelle d'une synthèse d'études quantitatives qui consiste à sélectionner des études puis à évaluer les études sélectionnées n'a pas été suivie. 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 puis calcul d'une taille d'effet globale). La troisième, qui a été finalement retenue, est une méta-analyse basée sur les données individuelles des participants : il s'agit alors de calculer les tailles d'effets des études et une taille d'effet globale à partir de l'ensemble des scores des élèves de toutes les études sélectionnées.

Les études d'efficacité en conditions réelles portant sur les mathématiques

Dix interventions ont comme sujet les mathématiques et ont été l'objet d'une étude d'efficacité en conditions réelles. L'étude menée pour évaluer l'intervention *Affordable Maths Tuition* devrait 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 6). Les niveaux d'études britanniques n'ont pas été convertis en niveaux d'études français[7] et les thèmes n'ont pas été traduits.
On sait déjà que le niveau de preuve 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.

Tableau 6
Premiers résultats pour les études d'efficacité en conditions réelles

Intervention	Date de publication*	Thème / Niveau	Effet (mois)	Niveau de preuve	Coût
1stClass@Number	Juillet 2018	Feedback Assesment / KS1	2	4	1
Affordable Maths Tuition	Juillet 2016	Maths / KS2	0	3	3
Ark Mathematics Mastery	Février 2015	Learning Behavior/ year 1 et year 7	1	3	2
Catch up numeracy	Février 2019	Maths / KS1	0	2	1
Chess in primary school	Juillet 2016	Learning Behavior/ KS1	0	5	1
Increasing Competence and Confidence in Algebra and Multiplicative Structures (ICCAMS).	Décembre 2021	Feedback Assesment / KS3 (year 7 et 8)	0	3	1
Mathematical Reasoning	Décembre 2018	Learning Behavior/ KS1	1	4	1
Maths Champions	Juillet 2018	Early year / KS1	2	2	1
Shared Maths	2014	Learning Behavior/ KS1	0	4	1
Tutor Trust Affordable Tutoring	Novembre 2018	Maths / KS1 (year 6)	3	4	2

KS1 : maternelle ; KS2 : primaire ; KS3 : collège. * : 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 présentant un effet supérieur ou égal à 1 mois et associées à un niveau de preuve supérieur ou égal à 3.

Des informations plus détaillées issues des rapports d'évaluation de ces 10 interventions sont présentées dans le tableau 7. Il s'agit notamment (pour l'échantillon analytique complet mais

[7] https://frenchradar.com/systeme-educatif-anglais-francais/

aussi pour les élèves bénéficiant de l'aide alimentaire) des effectifs totaux, des tailles d'effets, des intervalles de confiance, et des valeurs-p publiés. Si on s'intéresse plus particulièrement à la précision des estimations des tailles d'effet, seule l'étude de l'intervention *Ark Mathematics Mastery*[8] publie une taille d'effet statistiquement significative au sens classique du terme pour un niveau de confiance égal à 0,95. Enfin, les tailles d'effets calculées pour les élèves éligibles au FSM soit ne sont pas associées à un niveau de preuve, soit ont des niveaux de preuve inférieurs à ceux de l'étude menée sur l'échantillon total ; cela vient du fait que la taille de l'échantillon des élèves éligibles au FSM est inférieure à celle de l'échantillon total. *Tutor Trust Affordable Tutoring* est l'intervention la plus efficace avec une taille d'effet égale à 0,19, mais non significative. Son efficacité semble plus importante pour les élèves de niveaux socio-économiques défavorisés avec pour ce groupe d'élèves une taille d'effet égale à 0,25 (mais toujours non significative). Cette intervention a consisté à faire intervenir des tuteurs de l'organisme de bienfaisance Tutor Trust (des étudiants ou jeunes diplômés formés par cet organisme) auprès d'élèves âgés de 10 à 11 ans (Year 6) et repérés par leurs enseignants comme étant en difficulté. Afin d'améliorer leur niveau en mathématiques, ces élèves ont reçu 12 heures de cours, généralement une heure par semaine pendant 12 semaines. Dans la grande majorité des écoles, des groupes de 3 élèves étaient placés sous la responsabilité d'un tuteur et les séances se sont déroulées durant les heures de classe.

Tableau 7

Résultats détaillés pour les études d'efficacité en conditions réelles

Intervention	**Échantillon total**		**Élèves éligibles au FSM**	
	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]	1610	0,07 [-0,04 ;0,17]
	KS1 + KS2	**0,07]0,00 ; 0,14] p<0,05**	Pas d'informations	
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	2786 (year 3)	0,01 [-0,07 ;0,09] p=0,89	554	-0,05 [-0,17; 0,07] p=0,44
	2683 (year 5)	0,02 [-0,06 ;0,1] p=0,68	535	0,05 [-0,09 ; 0,19] p=0,47
Tutor Trust Affordable Tutoring	1201	0,19 [-0,05 ; 0,44] p=0,10	576	0,25 [-0,02 ; 0,51] p=0,06

N = effectifs des groupes intervention et contrôle ; ES = taille d'effet ; CI = intervalle de confiance ou de compatibilité ; p = valeur-p ; FSM =Free School Meals. En **gras** le résultat statistiquement significatif.

[8] Cette taille d'effet a été calculée en suivant le modèle d'une méta-analyse à partir des deux tailles d'effet calculées pour les niveaux KS1 et KS2.

Impact des interventions sur les élèves éligibles au FSM

On terminera ici en donnant les principaux résultats des méta-analyses conduites par EEF sur ses propres études à partir des scores des élèves éligibles au FSM (Ashraf, B. et al., 2021). Parmi elles, 48 résultats[9] 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 (EEF, 2020), les « intervalles de confiance » sont devenus des « intervalles de compatibilité » et il n'est plus mentionné explicitement qu'un résultat est (ou non) statistiquement significatif. Les auteurs ont classé les tailles d'effets par ordre décroissant ce qui leur a permis de caractériser quatre interventions comme « prometteuses » (tableau 8).

Tableau 8

Résultats des 4 interventions prometteuses pour l'apprentissage des mathématiques (élèves éligibles au FSM)

Interventions	ES	Intervalle de compatibilité
Powerful Learning Conversations	0,31	[-0,25 ; 0,98]
Dialogic Teaching	0,16	[0,03 ;0,29]
Improving Numeracy and Literacy in KS 2	0,13	[-0,10 ; 0,36]
Act, Sing, Play 1	0,12	[-0,10 ; 0,35]

ES = taille d'effet

Les résultats publiés par les rapports originaux d'une part et par la méta-analyse d'EEF d'autre part et concernant une même intervention sont comparés dans le tableau 9. Comme au chapitre précédent, il s'agit d'interventions qui ont fait l'objet d'une étude d'efficacité en conditions réelles. L'intervention *Maths Champion* ne donnait pas d'informations sur les élèves éligibles au FSM et les résultats de l'intervention *ICCAMS* ont été publiés après la réalisation de la méta-analyse. Pour les interventions *Catch up numeracy* et *Tutor Trust Affordable Tutoring*, la méta-analyse s'est basée sur des études d'efficacité (*efficacy trial*) et non sur les études d'efficacités en conditions réelles. Ces quatre interventions sont donc absentes du tableau 9. Enfin pour l'intervention *Ark Mathematics Mastery*, seuls les résultats de l'étude menée au secondaire sont mentionnés (l'étude conduite en primaire et la synthèse de ces deux études n'ayant pas publié de résultats concernant les élèves éligibles au FSM).

[9] Là encore il ne s'agit pas du nombre d'études, plusieurs résultats ayant pu être calculés dans une même étude.

Tableau 9

Résultats comparés publiés par les rapports originaux et par la méta-analyse d'EEF

Intervention		Résultats du rapport original		Résultats de la méta-analyse EEF	
Nom	**Date**	**N**	**ES, [CI]**	**N**	**ES, [CI]**
1stClass@Number	2018	149	-0,03 [-0,30 ; 0,36]	149	-0,02 [-0,37 ; 0,33]
Affordable Maths Tuition	2016	184	-0,08 [-1,23 ; 0,74]	786	-0,04 [-0,22 ; 0,15]
Ark Mathematics Mastery secondary	2015	1610	0,07 [-0,04 ;0,17]	1609	0,06 [-0,08 ; 0,21]
Chess in primary school	2016	1321	0,01 [-0,18 ; 0,19]	1291	0,01 [-0,18 ; 0,19]
Mathematical Reasoning	2018	1323	0,09 [-0,07 ; 0,25]	1342	0,06 [-0,09 ; 0,20]
Shared Maths	2015	554	-0,05 [-0,17 ; 0,07]	554	-0,03 [-0,15 ; 0,10]
	2015	535	0,05 [-0,09 ; 0,19]	535	0,05 [-0,07 ; 0,18]

N = effectifs des groupes intervention et contrôle ; ES = taille d'effet ; CI = intervalle de confiance ou de compatibilité

La taille d'échantillon mentionnée par la méta-analyse concernant l'intervention *Affordable Maths Tuition* est visiblement erronée (erreur probablement due à une faute de frappe). Quand on compare les résultats issus des mêmes études et qui concernent donc 6 interventions, de légères différences peuvent être notées en ce qui concerne la taille d'effet. Ce n'est pas très surprenant, puisque chaque taille d'effet a été recalculée par les auteurs de la méta-analyse en utilisant un modèle différent de celui utilisé par les auteurs de l'étude originale. Plus étonnant, les tailles d'échantillons sont différentes pour 4 résultats sur 7, sans qu'il ait été possible de comprendre pourquoi.

Quand on considère l'ensemble des 48 analyses donnant des informations sur les scores des élèves en mathématiques, la taille d'effet globale publiée 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, par type d'interventions (individuelle, en petits groupes, en classe entière, sur un établissement entier), par design d'études (études multisites où les élèves sont affectés aléatoirement au niveau individuel et cela dans plusieurs établissements, ou bien études avec affectation aléatoire au niveau des établissements) et par type de scores (résultats principaux ou résultats secondaires). Toutes ces tailles d'effet sont proches de zéro et leur intervalle de compatibilité inclut la valeur nulle.

Enfin, la différence entre l'effet des interventions sur les élèves éligibles au FSM d'une part et l'effet sur les élèves non éligibles au FSM d'autre part a été évaluée en calculant là aussi une taille d'effet. En ce qui concerne les compétences mathématiques, elle est égale à -0,01 avec comme intervalle de compatibilité [-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. Cette remarque est écrite sans explicitation des résultats calculés.

D'une façon générale, on retiendra qu'aucune interprétation de l'ampleur des tailles d'effet calculées n'est proposée 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 tailles d'effet globales (calculées pour des sous-groupes d'études) évoquées précédemment. Les seuls commentaires concernent le signe (soit positif, soit négatif) de ces estimations et si certaines interventions ont été classées selon l'amplitude de leurs effets, aucun seuil n'a été explicitement posé.

Conclusions

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'essais contrôlés randomisés en une dizaine d'années, tous réalisés dans des conditions similaires par des équipes qui, bien qu'indépendantes, partagent les mêmes procédures et s'astreignent à suivre un cadre méthodologique rigoureux, est tout simplement exceptionnelle. La rédaction de documents ressources à destination des chercheurs est également à mettre au crédit d'EEF.

Les résultats des premières études EEF se sont avérés assez rapidement ne pas être à la hauteur des attentes : les tailles d'effet calculées étaient modestes, souvent statistiquement non significatives, et contrastaient avec les résultats de la littérature internationale publiés antérieurement (Dawson *et al.,* 2018). C'est aussi à cette époque que des chercheurs ont par ailleurs signalé que les procédures suivies par certaines études quantitatives conduisaient à une surestimation des tailles d'effet calculées (Cheung et Slavin, 2016, Zhang *et al.,* 2013). Les études incriminées étaient souvent conduites par les concepteurs mêmes de l'intervention, qui parfois utilisaient leurs propres tests (on parle de tests « maison ») sur des effectifs très faibles (une ou deux classes d'élèves par exemple). Ces études intégrées dans les méta-analyses (notamment celles réalisées par EEF dans le cadre de son Toolkit[10]) ont largement contribué à la surestimation des tailles d'effets globales publiées, tailles d'effets qui ont servi de référence aux chercheurs d'EEF pour dimensionner leurs premières études. Mais même quand ces études ont été par la suite dimensionnées de façon à ce que de faibles tailles d'effets (de l'ordre de 0,2) puissent être détectées, ces dernières sont la plupart du temps restées statistiquement non significatives. C'est probablement ce qui a conduit EEF à considérer les valeurs-p calculées dans les rapports EEF d'une nouvelle façon. Mais si l'habituelle interprétation dichotomique semble aujourd'hui abandonnée par EEF, elle reste de mise pour d'autres organisations comme le What Works Clearinghouse et le Center for Research and Reform in Education. Et cet abandon laisse aujourd'hui le chercheur démuni quand il se doit d'interpréter les résultats.

[10] Le Toolkit (que l'on peut traduire par boîte à outil) est un ensemble de méta-analyses qui rassemblent les études par thèmes (par exemple des caractéristiques pédagogiques comme l'enseignement collaboratif) dont la vocation est de fournir des informations concises fondées sur des éléments probants.

Dans le même ordre d'idées, on remarquera que EEF ne publie plus sur son site internet de listes d'interventions prometteuses comme c'était le cas par le passé. Si la mise à jour de leur site internet durant l'été 2021 s'est accompagnée de la suppression de nombreux documents publiés antérieurement, les communiqués de presse mentionnant l'existence de ces interventions « prometteuses » aujourd'hui disparues ont échappé à cette purge. Force est de constater que quand la qualité des études augmente, les résultats semblent diminuer. Et que le passage à l'échelle pour une intervention, c'est-à-dire sur un vaste terrain avec des intervenants qui ne sont pas les promoteurs de l'intervention, est une étape cruciale (Lima et Tual, 2021 ; Dawson *et al.*, 2018), étape qui restera encore quelque temps un objet de recherche pour les scientifiques.

On terminera en évoquant à nouveau les acteurs de la communauté éducative française : que peuvent-ils retirer de toutes ces informations ? Les interventions dont nous venons de parler ne les concernent pas directement pour les raisons évoquées au tout début de cet article. C'est donc plutôt au niveau des méthodes de recherche que des pistes sont à explorer, qui pourraient s'avérer pertinentes pour évaluer l'efficacité de recommandations ou d'actions susceptibles d'être posées à l'échelle de notre territoire national. Et dans ce cadre, évaluer l'efficacité de certaines interventions soigneusement décrites en conduisant des essais contrôlés randomisés de bonne qualité semble être un préalable des plus raisonnable. Ce pan de la recherche devrait alors trouver sa place dans le paysage scientifique français et permettre aux chercheurs de répondre à certains types de questions (notamment dans le cadre d'analyses confirmatoires), et aux praticiens d'alimenter leur réflexion en analysant les résultats d'études quantitatives.

Références

Ashraf, B., Singh, A., Uwimpuhwe, G., Coolen-Maturi, T., Einbeck, J., Higgins, S., & Kasim, A. (2021). Individual participant data meta-analysis of the impact of EEF trials on the educational attainment of pupils on Free School Meals: 2011 – 2019. *The Education Endowment Foundation.* 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

Cheung, A., & Slavin, R. (2016). How methodological features affect effect sizes in education. *Educational Researcher, 45* (5), 283-292. https://doi.org/10.3102/0013189X16656615

CSEN. (2023). La boite à idées du CSEN-Quelques pistes pédagogiques fondées sur la recherche.

Dawson, A., Yeomans, E., & Brown, E. R. (2018). Methodological challenges in education RCTs: Reflections from England's Education Endowment Foundation. *Educational Research*, 60(3), 292–310. https://doi.org/10.1080/00131881.2018.1500079

Dehaene, S., & Pasquinelli, E. (n.d.). La recherche translationnelle en éducation. Pourquoi et comment ? CSEN. https://www.reseau-canope.fr/fileadmin/user_upload/Projets/conseil_scientifique_education_nationale/Ressources_pedagogiques/La_recherche_translationnelle_en_education.pdf

EEF. (2018). Statistical uncertainty in Randomised Controlled Trials. https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/evaluation-design

EEF. (2019a). Classification of the security of findings from EEF evaluations (2.0). https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/evaluation-design

EEF. (2019b). Cost evaluation guidance for EEF evaluations https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/evaluation-design

EEF. (2019c). Implementation and process evaluation guidance for EEF evaluations. https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/evaluation-design

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

EEF. (2022). Statistical analysis guidance for EEF evaluations. https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/evaluation-design

La version antérieure, datée de 2018, n'est plus téléchargeable.

EEF (n.d.). Impact evaluation report template. https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/reporting-templates

Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R., & Kerr, K. (2016). Implementation and process evaluation (IPE) for interventions in education settings : An introductory handbook. The Education Endowment Foundation. https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/evaluation-design

Lima, L., & Tual, M. (2021). De l'étude randomisée à la classe : est-il suffisant d'avoir des données probantes sur l'efficacité d'un dispositif éducatif pour qu'il produise des effets positifs en classe ? *Éducation et Didactique*, *16*(1), 153-162. https://doi.org/10.4000/educationdidactique.9899

Tymms, P., & Kasim, A. (2018). Research paper n°2_Standard deviation as an outcome on interventions : a methodological investigation. The Education Endowment Foundation. 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

Zhang, Z., Xu, X., & Ni, H. (2013) Small studies may overestimate the effect sizes in critical care meta-analyses: a meta-epidemiological study, *Critical Care 17*(1):R2. https://doi.org/10.1186/cc11919