Comment enseigner les maths ?
La réponse du Center for Research and Reform in Education

Nathalie ROQUES

Juin 2021

Résumé. Les méthodes et résultats de la méta-analyse conduite par le Center for Research and Reform in Education sur l'enseignement des mathématiques au secondaire publiée en 2009 ont été étudiés. Après analyse des 7 études primaires sélectionnées par la méta-analyse et s'intéressant à un apprentissage de type coopératif, l'efficacité de cette pratique pédagogique est discutée. Certains éléments associés à la qualité des méta-analyse menées dans le domaine des sciences de l'éducation sont alors proposés.

Avant-propos

La méta-analyse dont il sera principalement question ici est le fruit d'un travail mené par Robert SLAVIN, Cynthia LAKE et Cynthia GROFF, sous l'égide du Center for Research and Reform in Education de l'université Johns Hopkins aux Etats-Unis. Robert SLAVIN a dirigé ce centre durant 20 années, et il sera fait référence à cette synthèse en ne citant que son nom. Il est décédé le 24 avril 2021, pendant la rédaction de ce document.

Le terme *méta-analyse* utilisé dans le texte fait référence à la synthèse quantitative d'études primaires sélectionnées selon des critères précisément définis (COOPER, 2019). Pour alléger le texte, le terme *synthèse* sera parfois employé comme synonyme.

Ce document a été écrit après la rédaction de *Comment enseigner les maths ? La réponse du What Works Clearinghouse* qui fournit des informations sur les principes suivis par les méta-analyses et qui rend compte également de quelques questions et discussions relatives à ce type de synthèse. Ces éléments ne seront pas repris ici.

Les études primaires sont écrites en américain et les niveaux d'études sont ceux définis par le système éducatif américain ; ils ont été traduits dans ce texte (par exemple le 7ème grade américain correspond à notre classe de 5ème). Certaines traductions de l'américain sont suivies de l'expression originale.

Des documents annexes (formulaire, fichiers Excel) peuvent être téléchargés sur le site www.mathadoc.fr. A la fin du texte, un glossaire donne les définitions des mots écris en **gras** à leur première apparition dans le texte et les abréviations utilisées sont explicitées.

Introduction

Les **méta-analyses** permettent d'embrasser en quelques minutes des dizaines d'études quantitatives et de se forger ainsi une opinion sur des questions de recherche ayant fait l'objet de plusieurs analyses indépendantes. Dans un précédent livre[a], il a été question des synthèses réalisées par le What Works Clearinghouse qui, depuis 2002, produit des méta-analyses portant sur l'enseignement. Le Center for Research and Reform in Education dirigé par Robert SLAVIN jusqu'à sa mort en 2021, a lui aussi mené un travail de grande ampleur en conduisant plusieurs synthèses que l'on peut répartir en deux groupes. Le premier groupe est constitué de méta-analyses s'intéressant à l'enseignement de la lecture, de la rédaction, des sciences, des mathématiques, aussi bien à l'école primaire qu'au collège ou au lycée. Un second groupe rassemble des méta-analyses qui ont comme objectif de définir des critères méthodologiques afin d'améliorer la qualité de ce type de synthèses conduites dans les sciences de l'éducation. Pour l'essentiel, nous nous intéresserons ici à la méta-analyse qui porte sur l'apprentissage des mathématiques par les élèves du secondaire (SLAVIN, 2009) ; nous ferons parfois référence à la synthèse qui a porté sur l'enseignement des mathématiques à l'école primaire[b] (SLAVIN, 2008) mais aussi à celle qui a concerné l'enseignement assisté par ordinateur en primaire et au secondaire, toujours dans le domaine des mathématiques (CHEUNG, 2013). Enfin, nous reprendrons certaines des conclusions proposées par les synthèses du second groupe dans la troisième partie de ce texte.

La recherche menée par Robert SLAVIN et dont il va être question ici, poursuivait un double objectif : repérer les **interventions** susceptibles de favoriser l'apprentissage des mathématiques par les élèves du secondaire, interventions qui peuvent être mises en place (ou mises en œuvre) par leurs enseignants en milieu scolaire ordinaire pour le premier, mais aussi caractériser ces interventions en identifiant certains de leurs éléments les plus remarquables pour le second. On devra donc distinguer deux groupes de résultats. Les premiers concernent l'évaluation des interventions qui portent sur l'enseignement des mathématiques ; ces interventions sont soit un programme scolaire (qui prennent souvent la forme d'un manuel scolaire), soit un outil numérique (comme un logiciel déployé sur ordinateurs), soit une méthode pédagogique, et sont la plupart du temps identifiées par un nom (par exemple une marque commerciale déposée aux Etats-Unis). Les résultats du second groupe permettront d'évaluer les éléments caractéristiques associés à chacune des interventions, ces éléments permettant de constituer des catégories et des sous-catégories dans lesquelles sont classées les interventions. Nous reviendrons plus loin sur leurs définitions. Ces évaluations partagent bien sûr des points communs avec celles qui ont été

[a] *Comment enseigner les maths ? La réponse du What Works Clearinghouse* (ROQUES, 2021a)

[b] Il s'agira alors de la méta-analyse publiée en 2008 (en 2007 sur le site bestevidence.org) et qui a souvent été associée à celle dont nous allons préciser les résultats. Depuis, les résultats d'une méta-analyse menée par le Center for Research and Reform in Education et portant sur ce même thème ont été publiés en janvier 2021 (PELLEGRINI, 2021) comme nous l'évoquerons plus loin.

menées par le What Works Clearinghouse[a] (WWC) : les *Rapports d'intervention* (pour le premier groupe de résultats) et les *Guides des pratiques* (pour le second groupe).

Dans la première partie de ce document, la méthode mise en œuvre pour la réalisation de cette méta-analyse est décrite et ce sera l'occasion de revenir sur certains des fondements qui structurent ce type de synthèses quantitatives. La présentation des résultats constitue la seconde partie, où une attention plus particulière sera accordée aux éléments concernant l'apprentissage coopératif, clairement mis en avant par Robert SLAVIN. La troisième partie s'appuiera sur cette analyse pour alimenter une discussion sur les points saillants de cette synthèse. Les lecteurs souhaitant comprendre comment les résultats présentés dans les différents tableaux des parties 2 et 3 ont été calculés, sont invités à consulter les fichiers Excel, ainsi qu'une annexe qui donne les formules mathématiques utilisées (téléchargeables sur le site mathadoc.fr). Une présentation des outils statistiques mis en œuvre dans les méta-analyses en éducation est également disponible sur ce site internet. Dans ce qui suit, la technique calculatoire a été volontairement mise en côté ; il n'en reste pas moins vrai qu'elle est intimement liée aux concepts théoriques qui charpentent toute méta-analyse.

[a] Le What Works Clearinghouse est une émanation de l'Institut des sciences de l'éducation du département de l'Éducation des États-Unis

Partie 1 : La méthode

Les données des études par comparaison de groupes qui s'intéressent à l'enseignement des mathématiques constituent la matière première de la méta-analyse menée par le Center for Research and Reform in Education (SLAVIN, 2009) ; ces études seront considérées ici comme des **études primaires**. Leurs objectifs tout comme les procédures qu'elles mettent en œuvre s'appuient sur les concepts qui charpentent les expériences scientifiques dans le domaine des sciences de l'éducation. Deux échantillons sont comparés : un premier échantillon d'élèves qui sont soumis au traitement (l'intervention évoquée ci-dessus) intéressant le chercheur, échantillon dénommé « groupe traitement ». Et un second échantillon d'élèves qui servira de « groupe contrôle ». A la fin de l'expérience, les élèves sont soumis au même test et obtiennent tous un score ; c'est la différence observée entre les deux groupes de scores qui va permettre d'inférer à l'échelle de toute une population (c'est-à-dire de généraliser les résultats). Ces études quantitatives mettent toutes en œuvre des techniques statistiques qui produisent des indicateurs (par exemple des moyennes et des écarts-types) permettant à leurs auteurs de tirer des conclusions dans le cadre défini par ce type de recherche ; ce sont ces indicateurs publiés dans les études primaires qui constituent les données de la méta-analyse qui est de ce fait considérée comme une étude secondaire.

Six étapes sont suivies

La méta-analyse conduite par Robert SLAVIN a suivi six étapes. Ces étapes s'appuient sur le squelette des procédures générales mises en œuvre pour réaliser des méta-analyses dans le domaine des sciences de l'éducation mais également dans d'autres domaines comme celui de la médecine. Elles ont été numérotées ici de 1 à 6 pour une courte présentation (des détails seront donnés par la suite pour certaines d'entre elles). Une étape 0 est ajoutée à cette liste, étape qui n'a pas été suivie par Robert SLAVIN à proprement parlé, mais s'invite de nos jours de façon pressante comme un gage de qualité pour ce type de synthèse ; c'est la raison pour laquelle elle a été mentionnée. Elle servira également d'appui à la discussion en dernière partie de ce texte.

Étape 0 : Définir ou sélectionner un protocole qui fixe les règles des étapes suivantes et le publier *avant* que les résultats soient connus[a]. Ici, c'est la partie Méthode de l'article publié (dont il n'est pas clair qu'elle ait été rédigée a posteriori ou a priori) qui tient lieu de protocole.

Étape 1 : Rassembler les études susceptibles d'avoir publié des résultats intéressants.

Étape 2 : Sélectionner les études selon des critères définis a priori et écarter de l'analyse celles qui ne sont pas conformes à ces critères.

Étape 3 : Regrouper les études. Les études qui analysent la même intervention sont regroupées pour évaluer l'effet de cette intervention. Après avoir défini un ou plusieurs

[a] De nos jours, les auteurs de méta-analyses sont invités à enregistrer leur protocole avant de commencer leur méta-analyse (voir par exemple le registre PROSPERO, www.crd.york.ac.uk/prospero/).

éléments caractéristiques des interventions, l'ensemble des études dont les interventions partagent un même élément caractéristique sont également regroupées. Trois catégories ont été ainsi définies qui ont été elles-mêmes partagées en plusieurs sous-catégories (voir plus loin).

Étape 4 : Calculer des tailles d'effets (pour chaque étude primaire sélectionnée, puis pour chaque intervention, puis pour chaque catégorie et pour chaque sous-catégorie).

Étape 5 : Évaluer les interventions et les éléments caractéristiques définissant les catégories et sous-catégories.

Étape 6 : Publier les résultats.

L'étape 1 consiste à rechercher des études primaires principalement en interrogeant des bases de données, avec une attention particulière portée aux études qui ne sont pas publiées dans des revues à comité de lecture (« littérature grise »), comme les thèses, afin de limiter le biais de sélection (ce biais concerne particulièrement les études de petite taille montrant des effets peu importants voire négatifs dont les résultats ne sont souvent pas publiés). Les bases de données comme JSTOR, ERIC, … ont été interrogés, les études citées entre autres par le What Works Clearinghouse et le National Research Council (aux USA) ont été collectées et les éditeurs ont été sollicités. Ce sont plus de 500 études qui ont été rassemblées à cette étape.

L'étape 2 : sélectionner les études

Toutes les études rassemblées à l'étape 1 sont passées au crible de critères (voir liste ci-dessous) permettant de ne retenir que celles dont la qualité est jugée suffisante pour permettre d'identifier clairement un lien de cause à effet entre l'intervention d'une part et les résultats des élèves d'autre part. Un grand nombre d'études sont exclues de la méta-analyse car ne satisfont pas à ces critères définis a priori. La liste des études non sélectionnées est publiée et le motif de leur rejet succinctement décrit.

Liste des critères de sélection

- Les études doivent évaluer des interventions mises en œuvre durant des cours de mathématiques au collège ou au lycée (de la 6ème à la terminale) [a]. Les études sur des groupes de niveau ou des classes unisexes n'ont pas été incluses.
- Les études comparent des élèves soumis à l'intervention (le groupe traitement) avec des élèves qui suivent un programme « standard » (le groupe contrôle).
- Les études peuvent être menées dans n'importe quel pays, mais les rapports doivent être rédigés en anglais et publiés après 1970.
- Les études doivent attribuer aléatoirement ou par appariement les élèves dans les groupes traitement et contrôle, et prendre en considération le niveau des élèves avant

[a] Aux Etats-Unis, certaines classes de 6ème peuvent être accueillis dans une école primaire.

l'intervention dans leurs calculs (voir plus loin le calcul de la taille d'effet). Les études par régression de la discontinuité peuvent être incluses (mais aucune étude de ce type n'a été repérée dans cette méta-analyse). Les études sans groupe contrôle, comme les comparaisons posttest-prétest ou les comparaisons à des gains attendus ont été exclues.

- Des tests doivent être effectués avant le traitement (on parle de prétests), à moins que les études n'utilisent l'attribution aléatoire d'au moins 30 unités (élèves, classes ou écoles) et qu'elles prouvent qu'il n'y a pas de différence de niveau avant l'intervention. Si les différences entre les scores prétests des groupes traitement et contrôle sont supérieures à 50% d'écart-type, l'étude est exclue.
- Les mesures dépendantes sont des mesures quantitatives de la performance en mathématiques, comme des scores obtenus après passation de tests standardisés. Les mesures issues de tests « maison », c'est-à-dire créés par les auteurs de l'étude, ne sont acceptées que si elles concernent des compétences mathématiques générales et plus précisément des compétences visées également par le groupe contrôle, et non des mesures inhérentes à l'intervention (et qui ne peuvent concerner le groupe contrôle). Selon Robert SLAVIN, ce critère d'exclusion n'était pas activé dans les méta-analyses précédemment publiées sur ce sujet, comme celles du What Works Clearinghouse (WWC).
- Une durée minimum de l'intervention de 12 semaines est requise. Cette exigence a comme objectif de s'attacher à évaluer des interventions qui pourraient être menées sur une année scolaire complète et à ne pas défavoriser le groupe contrôle.
- Les études doivent compter au moins deux enseignants et 15 élèves par groupe.

L'étape 3 : regrouper les études

Deux types de regroupement vont être effectués. De façon assez classique, les études qui portent sur une même intervention sont rassemblées afin d'évaluer l'efficacité de l'intervention en question. Le second type de regroupement sera fait en fonction des éléments caractérisant les interventions (ou les études quand ces dernières ne sont pas associées à une intervention connue) et toutes les études sont réparties dans l'une des trois catégories principales suivantes : programmes mathématiques (A), enseignement assisté par ordinateur (B) ou méthodes pédagogiques (C). Enfin chacune de ces trois catégories a été divisée en plusieurs sous-catégories (voir description ci-dessous). Chaque étude est incluse dans une de ces sous-catégories, et donc dans la catégorie mère. Il peut arriver qu'une étude soit incluse dans plusieurs sous-catégories car présentant des résultats pouvant être associés à plusieurs éléments caractéristiques, mais cela reste exceptionnel : les catégories et sous-catégories ne partagent pas d'études en commun. Pour plus de clarté, catégories et sous-catégories ont été numérotées ici et cette numérotation est reprise dans le texte par la suite. Un tableau récapitulatif est proposé à la fin de cette description (**tableau 1**).

A - Programmes mathématiques (*mathematics curricula, MC*)

Les élèves suivent des programmes mathématiques particuliers qui sont explicités notamment dans les manuels scolaires qu'ils utilisent, manuels souvent édités par une entreprise commerciale. Trois sous-catégories ont été identifiées en les associant aux éléments caractéristiques suivants :

- A1 - Programmes NSF (*National Science Foundation supported programs*) : les interventions suivent un programme présenté comme innovant, qui accorde une grande importance à la résolution de problèmes, l'examen de solutions alternatives et la compréhension des concepts mathématiques tels que préconisés par la National Science Foundation aux USA.
- A2 - Manuels traditionnels (*traditional textbooks*) : la conception des programmes est plus traditionnelle, notamment en ce qui concerne la répartition entre procédures algorithmiques, développement de concepts mathématiques et résolution de problèmes.
- A3 - Retour au bases (*back to basics textbooks*) : les programmes suivent une approche pas à pas des mathématiques.

B - Enseignement assisté par ordinateur (*computer-assisted instruction, CAI*)

Cette catégorie regroupe des interventions qui utilisent des logiciels déployés sur ordinateur. Trois sous-catégories ont été identifiées en les associant aux éléments caractéristiques suivants :

- B1 - Enseignement numérique complet (*core CAI*[a]) : l'enseignant est remplacé par l'ordinateur qui propose aux élèves des activités personnalisées en fonction de leur niveau et de leurs besoins, des évaluations et des prescriptions individualisées.
- B2 - Enseignement numérique complémentaire (*supplemental CAI*) : les élèves utilisent l'ordinateur pour des activités complémentaires (entre 10 et 15 minutes par jour), par exemple pour combler des lacunes.
- B3 - Évaluation de l'élève par ordinateur (*Computer managed learning systems*) : l'élève utilise l'ordinateur pour être évalué, des rapports sont édités à l'attention de l'enseignant et de l'élève.

C - Méthodes pédagogiques (*Instructional Process Programs, IPP*)

Cette catégorie regroupe les interventions qui ont comme objectif de permettre aux enseignants de développer des stratégies pour enseigner efficacement les mathématiques. Dans les études par comparaison de groupes, les manuels, les contenus et les objectifs académiques sont communs aux groupes traitement et contrôle ; les méthodes d'enseignement seules permettent de différencier ces deux groupes. Cinq sous-catégories ont été identifiées en les associant aux éléments caractéristiques suivants :

- C1 - Apprentissage coopératif (*cooperative learning*) : les élèves sont répartis en petits groupes hétérogènes (voir la partie 2 pour plus de détails).

[a] Dénommée *Comprehensive CAI* dans l'autre méta-analyse analysant l'impact de l'enseignement assisté par ordinateur dont nous avons déjà évoqué l'existence (CHEUNG, 2013).

- C2 - Pédagogie métacognitive (*metacognitive strategy instruction*) : l'enseignement est basé sur l'explicitation des processus métacognitifs. Ainsi les élèves sont invités à se poser des questions à voix haute, à faire des liens avec des problèmes/concepts similaires déjà rencontrés et à réfléchir de façon explicite quand ils doivent choisir une stratégie parmi d'autres.
- C3 - Enseignement personnalisé (*individualized instruction*) : l'élève choisi son activité librement, l'enseignant assistant chaque élève individuellement ou en petits groupes.
- C4 - Pédagogie de la maitrise (*mastery learning*) : les élèves sont fréquemment évalués et doivent s'assurer de la maitrise de certaines notions avant de commencer une nouvelle leçon.
- C5 - Refonte complète de l'enseignement (*comprehensive school reform*) : la formation professionnelle des enseignants concernant les méthodes pédagogiques et la définition des programmes, l'organisation particulière de l'école et de la classe, l'investissement attendus des parents sont les principales caractéristiques de cet élément (et correspondent à la description de méthodes pédagogiques déployées dans certains établissements privés français).

A - Programmes mathématiques
A1 - Programmes NSF
A2 - Manuels traditionnels
A3 - Retour aux bases
B - Enseignement assisté par ordinateur
B1 - Enseignement numérique complet
B2 - Enseignement numérique complémentaire
B3 - Évaluation de l'élève par ordinateur
C - Méthodes pédagogiques
C1 - Apprentissage coopératif
C2 - Pédagogie métacognitive
C3 - Enseignement personnalisé
C4 - Pédagogie de la maitrise
C5 - Refonte complète de l'enseignement

Tableau 1 : les 3 catégories et les 11 sous-catégories

Dans la méta-analyse publiée en 2013 sur l'enseignement assisté par ordinateur (CHEUNG, 2013), les études ont été regroupées par intervention mais ces regroupements n'ont pas donné lieu au calcul d'une taille d'effet globale ni à une évaluation (voir plus loin) ; elles ont également été regroupées dans les catégories B1, B2 et B3 décrites ci-dessus, mais aussi en fonction de certaines caractéristiques des élèves (leur niveau d'étude – primaire ou secondaire– et leur statut socio-économique), et en fonction de caractéristiques liées aux

études (années de publication, design,…) ou aux interventions (intensité, qualité de l'implémentation).

Les résultats concernant cette étape sont décrits plus loin dans la seconde partie de ce texte.

L'étape 4 : calculer des tailles d'effet

Dans un premier temps, Robert SLAVIN calcule une **taille d'effet** pour chaque étude sélectionnée. Cette taille d'effet, notée ES dans le texte (pour *Effect Size*), est égale à la différence des moyennes posttests des groupes traitement et contrôle, divisée par l'écart-type du groupe contrôle[a] (cet écart-type est considéré par l'auteur comme reflétant le mieux l'écart-type de la population dont les échantillons sont extraits), voir formule (1). Les moyennes posttests sont dans ce cas ajustées aux moyennes prétests, et si ce n'est pas le cas, c'est la différence entre la taille d'effet posttsest et la taille d'effet prétest qui est calculée, voir formule (2). La méta-analyse menée en 2011 sur l'enseignement assisté par ordinateur (CHEUNG, 2013) utilise une autre formule et remplace l'écart-type du groupe contrôle par un écart-type groupé calculé à partir des écarts-types des groupes traitement et contrôle[b].

$$ES = \frac{m_T - m_C}{s_C} \qquad (1)$$

m_T=moyenne des scores posttest du groupe traitement ajustée aux scores prétest de ce groupe.

m_C= moyenne des scores posttest du groupe contrôle ajustée aux scores prétest de ce groupe.

s_C= écart-type non ajusté des scores posttests du groupe contrôle

$$ES = \frac{m_{T2} - m_{C2}}{s_{C2}} - \frac{m_{T1} - m_{C1}}{s_{C1}} \qquad (2)$$

m_{T1}=moyenne des scores prétests du groupe traitement

m_{C1}= moyenne des scores prétests du groupe contrôle

s_{C1}= écart-type des scores prétests du groupe contrôle

m_{T2}=moyenne des scores posttest du groupe traitement

m_{C2}= moyenne des scores posttest du groupe contrôle

s_{C2}= écart-type des scores posttests du groupe contrôle

Si nécessaire, les écarts-types des classes ou des établissements sont ajustés pour estimer l'écart-type au niveau des élèves (les premiers étant souvent plus faibles que ce dernier). Dans

[a] C'est le Δ de Glass

[b] C'est le g de Hedges

le texte, une estimation de la significativité statistique accompagne le plus souvent les tailles d'effet calculées[a], et dans certains cas une valeur-p est parfois mentionnée (le plus souvent comme étant inférieure à un seuil, la valeur exacte est plus rarement donnée). Aucune information sur les calculs de cette valeur-p n'est fournie, et les cas particuliers (prise en compte dans les calculs des analyses multiples, des clusters, etc.) ne sont pas évoqués. La question se pose en particulier dans le cas des études qui présentent des résultats pour plus de deux groupes ou pour des sous-tests : si des tailles d'effets sont parfois calculées et présentées par Robert SLAVIN comme des résultats partiels, la méthode de calcul de la taille d'effet globale de l'étude n'est pas indiquée (voir plus loin, partie 2, intervention STAD).

Dans un second temps, une taille d'effet globale est calculée pour chaque intervention et également pour chaque catégorie et pour chaque sous-catégorie. Il s'agit d'une moyenne des tailles d'effets des études du regroupement considéré (intervention, catégorie, sous-catégorie) pondérée par la taille de l'échantillon total de chacune des études[b] (formule 3). Le modèle statistique suivi n'a pas permis d'estimer la significativité statistique de ces tailles d'effet globales.

$$ES = \frac{\sum_i N_i \times ES_i}{\sum_i N_i} \qquad (3)$$

N_i= taille de l'échantillon total (groupe traitement et groupe contrôle) de l'étude i.

ES_i= taille d'effet de l'étude i.

Dans la méta-analyse portant sur l'apprentissage des mathématiques par les élèves du primaire (SLAVIN, 2008), la taille d'effet globale d'un regroupement d'études est égale à la médiane des tailles d'effet. Dans la méta-analyse publiée en 2013 sur l'enseignement assisté par ordinateur (CHEUNG, 2013), les tailles d'effet globales (ainsi que leur variance) ont été calculée en appliquant le modèle des effets aléatoires et une analyse de l'hétérogénéité a été menée (voir discussion plus loin).

Les résultats concernant cette étape sont décrits plus loin dans la seconde partie de ce texte.

L'étape 5 : évaluer

Les interventions d'une part et les éléments caractéristiques qui leurs sont associés (et qui définissent les catégories et sous-catégories) d'autre part, sont alors évalués. Pour les premières, des critères ont été définis qui permettent d'attribuer à chaque intervention un échelon allant de « preuve forte d'efficacité » à « pas assez de preuve ». En ce qui concerne

[a] Pour plus de détails sur les calculs, voir le formulaire et le fichier Excel correspondant.

[b] Les détails de ces calculs sont proposés dans *Mesurer l'effet d'un traitement. Méta-analyses et sciences de l'éducation* (ROQUES, 2021), www.mathadoc.fr

les caractéristiques, les règles tout comme les niveaux d'efficacité qu'elles déterminent, sont beaucoup plus flous.

Évaluer les interventions

La qualité du design des études primaires sélectionnées va intervenir ici. Robert SLAVIN distingue 4 designs d'études, classés ci-dessous par ordre décroissant de qualité.

- Les **expériences randomisées** (*randomized experiments*) dans lesquelles les unités (élèves, classes, établissements) assignées aléatoirement au groupe intervention ou au groupe contrôle font l'objet de l'analyse statistique.
- Les **quasi-expériences randomisées** (*randomized quasi-experiments*), dans lesquelles des classes ou des établissements ont été assignés aléatoirement au groupe intervention ou au groupe contrôle, mais où ce sont les élèves qui ont été statistiquement analysés (parce que le nombre de classes ou d'établissements est trop faible et ne permet pas une analyse à leur niveau).
- Les **études par appariement** (*matched studies*) où les élèves des groupes intervention et contrôle sont appariés à l'étape du prétest sur des variables clés, avant que les résultats des posttests ne soient connus (il n'y a donc pas d'assignation aléatoire).
- Les **études par appariement post hoc** (*matched studies post hoc*) où la même méthode est utilisée mais rétrospectivement, une fois les résultats des posttests connus.

Les deux premiers types de recherche correspondent à la catégorie des essais contrôlés randomisés identifiés par le What Works Clearinghouse (*randomized controlled trials*) et les deux derniers à la catégorie des études quasi-expérimentales (*quasi-experimental design*).

La taille de l'échantillon de l'étude (somme des effectifs des groupes traitement et contrôle) sera également prise en compte lors de l'évaluation.

Ce sont la qualité du design des études sélectionnées et la valeur de la taille d'effet globale calculée pour chaque intervention qui permettent de classer l'efficacité de chaque intervention selon les quatre échelons décrits ci-dessous.

- **Preuve forte de l'efficacité**

Au moins deux études ont analysé l'intervention, dont l'une est une expérience randomisée ou une quasi-expérience randomisée ; la taille d'échantillon totale (sur l'ensemble des études) doit être supérieure ou égale à 500 et la taille d'effet globale doit être supérieure ou égale à 0,20[a].

- **Preuve moyenne de l'efficacité**

Au moins deux études ont analysé l'intervention avec une taille d'échantillon totale supérieure ou égale à 500 et une taille d'effet globale supérieure ou égale à 0,20.

- **Preuve limitée de l'efficacité**

Une ou plusieurs études ont analysé l'intervention avec une taille d'effet globale supérieure ou égale à 0,10

[a] Ici, c'est la formulation de l'article publié en 2009 qui est reprise, car la formulation proposée dans le rapport technique prête à confusion (voir aussi la discussion).

- **Pas assez de preuves**

Une ou plusieurs études ont analysé l'intervention avec une taille d'effet globale inférieure à 0,10.

La significativité statistique des tailles d'effet calculée pour chaque étude primaire n'intervient pas dans l'évaluation des interventions.

Évaluer les caractéristiques

Les éléments caractéristiques associés à chaque étude sont également évalués. Comme nous l'avons indiqué précédemment, ces éléments définissent des catégories d'interventions, comme l'enseignement assisté par ordinateur (B), et des sous-catégories, comme l'enseignement numérique complet (B1). Aucune règle n'a été définie a priori (c'est-à-dire avant le calcul des résultats) pour associer les tailles d'effet globales calculées à une évaluation de leur influence sur les apprentissages ; c'est donc leur comparaison les unes aux autres qui va permettre aux auteurs de conclure. La qualité des études est parfois évoquée pour souligner une conclusion.

Les résultats concernant cette étape sont décrits plus loin dans la seconde partie de ce texte.

L'étape 6 : publier

S'il ne peut rivaliser en termes de quantité d'articles ni d'informations fournies avec le site *Find what works*[a] du What Works Clearinghouse, le site internet *Best evidence encyclopedia*[b] du Center for Research and Reform in Education propose tout de même de nombreux éléments à télécharger gratuitement. Pour chacune des méta-analyses conduites, un rapport technique complet accompagne la référence de l'article publié. En ce qui concerne la méta-analyse qui nous intéresse tout particulièrement, le rapport technique est daté d'octobre 2008 (version 1.4) et l'article a été publié en juin 2009 dans la revue à comité de lecture *Review Educational Research.* Le rapport technique donne plus de détails que l'article publié, en décrivant notamment beaucoup plus longuement les interventions étudiées ; c'est pour cette raison que cette méta-analyse est présentée par leurs auteurs comme une *Best-Evidence Synthesis*. Ces deux documents diffèrent également par le nombre des études sélectionnées et donc en partie par les résultats publiés (voir la partie 2).

Un autre rapport technique, daté d'aout 2008 cette fois (version 1.3), était publié sur le site internet encore ces dernières années, mais a disparu depuis. Ont également disparu un résumé de 8 pages et un Guide pour l'enseignant (*Educator's Guide. What Works in Teaching Math ?*) daté de janvier 2010. Ce dernier concernait aussi bien les élèves scolarisés en primaire que les élèves scolarisés au secondaire. La disparition du second texte s'explique probablement par la publication des résultats de la nouvelle méta-analyse sur l'enseignement des mathématiques à l'école primaire en mars 2021 (PELLEGRINI, 2021), cette nouvelle

[a] https://ies.ed.gov/ncee/wwc/
[b] www.bestevidence.org

synthèse rendant les résultats de la précédente synthèse en partie obsolète[a]. Ces textes qui ont disparu sont téléchargeables sur le site www.mathadoc.fr et seront parfois cités dans ce texte.

[a] Contrairement au site internet du What Works Clearinghouse, il n'y a aucun dossier rassemblant les archives et il n'est pas fait mention des évolutions de certains des textes.

Partie 2 : Les résultats

Sélection et regroupement des études primaires (étapes 2 et 3)

Dans le rapport technique d'octobre 2008, 102 études primaires[a] ont été sélectionnées qui concernent 25 interventions et une liste des 271 études rejetées est publiée avec une courte description du motif de leur rejet. Les études sélectionnées ne sont que 100 dans l'article daté de 2009 où ne figurent pas deux études publiées concernant l'enseignement assisté par ordinateur et publiées en 2008 (les études n°23 et 24 de la sous-catégorie B2, voir fichier Excel en annexe). Les résultats présentés dans la suite de ce texte se sont appuyés sur les éléments fournis par le rapport technique. Ces études ont été réparties dans les différentes catégories et sous-catégorie de la façon suivante (**tableau 2**) :

	Nombre d'interventions	Nombre d'études
Total	**25**	**102**
A - Programmes mathématiques	**13**	**40**
A1 - Programmes NSF	8	26
A2 - Manuels traditionnels	3	3
A3 - Retour aux bases	2	11
B - Enseignement assisté par ordinateur	**9**	**40**
B1 - Enseignement numérique complet	4	17
B2 - Enseignement numérique complémentaire*	4	20
B3 - Évaluation de l'élève par ordinateur	1	3
C - Méthodes pédagogiques	**6**	**22**
C1 - Apprentissage coopératif	3	8
C2 - Pédagogie métacognitive	Pas d'intervention	2
C3 - Enseignement personnalisé	Pas d'intervention	2
C4 - Pédagogie de la maitrise	Pas d'intervention	6
C5 - Refonte complète de l'enseignement	3	4

Tableau 2 : nombre d'études et d'interventions dans les différentes catégories et sous-catégories

* : dans cette sous-catégorie, 15 études ne sont pas associées à une intervention.

Il peut être intéressant de comparer les études sélectionnées dans cette méta-analyse avec celles qui ont été sélectionnées dans la méta-analyse sur l'enseignement assisté par ordinateur (CHEUNG, 2013). Les 3 sous-catégories B1, B2 et B3 sont utilisées de la même manière dans les deux études. On ne va considérer dans notre comparaison que les études qui concernent les élèves du secondaire (dans sa synthèse de 2013, CHEUNG s'est aussi intéressé aux élèves du primaires). Bien entendu toutes les publications postérieures à 2007

[a] Plus précisément, 102 résultats associés à une étude. Certaines études proposent plusieurs résultats pour des interventions différentes par exemple, et ont pu être citées plusieurs fois.

et sélectionnées par la seconde synthèse n'ont pas été comptabilisées ici. Les critères de sélection étaient très proches et on pourrait s'attendre à ce que, à 3 ans d'intervalle seulement, on retrouve pratiquement les mêmes études[a]. On constate pourtant que ce n'est pas le cas (**tableau 3**). Ainsi 20 études sélectionnées dans la méta-analyse publiée en 2009 ne sont pas sélectionnées par la suivante[b].

	Uniquement dans la synthèse de 2009	Uniquement dans la synthèse de 2013	Dans les deux synthèses
B1-Enseignement numérique complet	9	0	8
B2-Enseignement numérique complémentaire	11	3	9
B3-Évaluation sur ordinateur	0	1	3
TOTAL	**20**	**4**	**20**

Tableau 3 : nombre d'études sélectionnées dans les deux méta-analyses de 2008 et de 2011.

Efficacité des interventions (étape 4 et 5)

Revenons à la méta-analyse qui nous intéresse plus particulièrement. Les interventions ont été classées selon les échelons définis auparavant. Deux interventions ont montré avoir une efficacité forte sur les apprentissages : IMPROVE et STAD[c]. Aucune étude montrant un effet modéré n'est citée dans le rapport technique daté de 2008 ; ce n'est pas le cas de l'article publié en 2009 qui mentionne l'intervention Prentice Hall Course 2 comme ayant un effet modéré, mais il s'agit d'une erreur (puisqu'il faut deux études au moins pour obtenir cet échelon, et qu'une seule étude sélectionnée analyse cette intervention). Dans le rapport technique l'intervention New Century a été oubliée dans le tableau récapitulatif (mais elle est bien présentée comme ayant un effet limité dans le corps du texte p.44). Elle disparaissait totalement du résumé et du Guide de l'enseignant, tous deux aujourd'hui disparus. Enfin aucune taille d'effet globale n'a été publiée pour évaluer l'impact du logiciel Plato web learning Network sur les apprentissages (analysé par les deux études n°20 et 21, sous-catégorie B2, voir annexe). Cette taille d'effet a été calculée ici comme égale à 0,23. La taille totale de l'échantillon étant de 589, cette intervention aurait pu être signalée comme présentant un effet modéré selon les critères de Robert SLAVIN ; elle a été placée à ce niveau d'efficacité dans ce texte (voir **tableau 4**). Les sous-catégories auxquelles chaque intervention

[a] Dans les deux méta-analyses, la date d'arrêt des recherches menées sur les bases de données n'est pas indiquée, mais la première publication des résultats de la méta-analyse qui nous intéresse datant d'aout 2008, il parait raisonnable d'exclure les études sélectionnées dans la seconde méta-analyse qui ont été publiées après janvier 2008 ; de plus les deux études (n°23 et 24, sous-catégorie B2) qui ont été rajoutés in extrémis dans le dernier rapport technique ont été publiées en mars 2007. Dans la méta-analyse de 2013, aucune date minimale n'est clairement indiquée. Les études les plus anciennes datent des années 70, ce qui laisse sous-entendre que les publications doivent être postérieures à 1970, ce qui était le cas pour la méta-analyse publiée en 2009.

[b] La liste des publications rejetées n'est pas publiée par la méta-analyse de 2013.

[c] Voir les descriptions des ces interventions et les commentaires au sujet du classement plus loin.

Efficacité forte	N	ES	N total	ES globale
IMPROVE (C1)	**182**	**0,79**		
	247	0,61		
	265	0,25	694	0,52
STADS (C1)	**588**	**0,21**		
	80	**0,2**		
	208	**1,09**		
	50	0,38	926	0,42

Efficacité modérée	N	ES	N total	ES globale
Plato Web Learning Network* (B2)	467	0,21		
	122	0,29	589	0,23

Efficacité limitée	N	ES	N total	ES globale
Cognitive Tutor (B1)	541	0,03		
	444	0,32		
	777	0,12		
	445	-0,07	2207	0,12
Core-Plus Mathematics (A1)	113	0,05		
	98	0,28		
	86	0,05		
	1050	0,12		
	14463	0,11	15810	0,11
The Expert Mathematician (B1)	70	0,38		
Jostens/compass Learning (B2)	90	0,22		
MATH Thematics (A1)	1792	0,25		
New Century (B2) **	306	0,28		
Path Mathematics (C5)	100	0,47		
Prentice Hall Course 2 (A2) ***	453	0,55		
Saxon Math (A3)	454	0,19		
	212	-0,25		
	211	0,39		
	?	0,07		
	?	0,25		
	185	-0,13		
	36	0,15		
	174	0,12		
	278	-0,44		
	276	-0,02		
	165	0,17	?	0,14****
Talent Development Middle School Mathematics Program (C5)	2130	0,18		

Tableau 4 : efficacités des interventions

* : cette intervention a été rajoutée ici d'après la remarque énoncée dans ce texte.

** : n'est pas mentionnée dans le tableau 4 du rapport technique (p.111), mais figure bien dans la liste (p.44) du rapport technique et est bien mentionnée dans le tableau de l'article (p.884).

*** : efficacité modérée annoncée (par erreur) dans *Review Educational Research* (2009)

**** : n'a pas pu être recalculée car il manquait les tailles d'échantillons pour deux études.

appartient sont rappelées en utilisant les notations du **tableau 1**. Les résultats des études expérimentales ou quasi-expérimentales sont écrits en **gras** (il n'y en a que 4), la première colonne donne le nom de l'intervention, la seconde la taille d'échantillon des études notée N, la troisième colonne la taille d'effet des études notée ES, la quatrième colonne l'effectif cumulé et la quatrième colonne donne la taille d'effet globale de l'intervention. Toutes les tailles d'effet globale ont été recalculées, à l'exception de l'intervention Saxon Math pour laquelle il manquait les tailles d'échantillon de deux études, en utilisant la formule (3) et les valeurs publiées par Robert SLAVIN ont été retrouvées.

Efficacité des différents éléments caractéristiques

Les différents éléments caractéristiques associés aux catégories et sous-catégories d'études ont été définis précédemment. Aucune information n'est donnée sur la procédure suivie pour associer une étude à l'un de ces éléments. Leur effet sur les résultats des élèves a été évalué en calculant là aussi une taille d'effet globale, mais sans faire référence aux échelons utilisés pour le classement des interventions, comme cela a déjà été signalé auparavant. L'évaluation se base uniquement sur l'observation des tailles d'effet accompagnée parfois d'un commentaire sur la quantité et la qualité des études appartenant à la catégorie ou sous-catégorie concernée, et sur la comparaison que les auteurs sont amenés à faire entre les tailles d'effet globales de différentes catégories ou sous-catégories. Dans l'article (SLAVIN, 2009) comme dans le rapport technique, ces tailles d'effet globales n'ont pas été rassemblées dans un tableau mais sont disséminées dans le texte principal.

Pour deux sous-catégories, les tailles d'effet globales ne sont pas publiées dans le rapport technique (mais elles le sont dans l'article de 2009). Les tailles d'effet globales des catégories A2 et C5 n'ont-elles pas été publiées. A l'aide des éléments fournis par Robert SLAVIN, toutes les tailles d'effet globales ont été recalculées. Pour 6 sous-catégories, les tailles d'échantillon de toutes les études regroupées étaient publiées et les tailles d'effet recalculées sont égales à celles publiées. Pour les autres, le manque d'information ne permet bien sûr pas de comparer les tailles d'effet recalculées avec celles qui ont été publiées[a]. Dans le **tableau 5** ci-dessous, la première colonne donne le nom de la catégorie ou de la sous-catégorie, la deuxième colonne indique le nombre d'études dont on ne connait pas la taille d'échantillon sur le nombre total d'études incluses dans la catégorie ou sous catégorie ; la troisième colonne donne la taille d'effet recalculée à partir des éléments disponibles ; la dernière colonne donne la taille d'effet publiée par Robert SLAVIN.

[a] Seuls les tailles d'effet des études primaires dont la taille d'échantillon est connue ont pu être utilisées ici comme cela est signalé dans le tableau (deuxième colonne).

	Effectifs manquants	ES globale recalculée*	ES globale publiée
A - Programmes mathématiques	**4/40**	***0,06***	**0,03**
A1 - Programmes NSF	2/26	*0,05*	0,00
A2 - Manuels traditionnels	0/3	0,13	n.calc
A3 - Retour aux bases	2/11	*0,01*	0,14
B - Enseignement assisté par ordinateur	**4/40**	***0,09***	**0,08**
B1 - Enseignement numérique complet	3/17	*0,09*	0,09
B2 - Enseignement numérique complémentaire	1/20	*0,12*	0,07
B3 - Évaluation de l'élève par ordinateur	0/3	-0,02	-0,02
C - Méthodes pédagogiques	**1/22**	***0,21***	**0,18**
C1 - Apprentissage coopératif	0/8	0,42	0,42
C2 - Pédagogie métacognitive	0/2	0,31	0,31**
C3 - Enseignement personnalisé	0/2	0,36	0,36**
C4 - Pédagogie de la maitrise	0/6	-0,05	-0,05
C5 - Refonte complète de l'enseignement	1/4	*0,19*	n.calc

Tableau 5 : tailles d'effet globales des catégories et sous-catégories

* : en italique les tailles d'effets recalculées à partir de données incomplètes.

** : ces tailles d'effet sont publiées dans l'article de 2009 mais ne sont pas mentionnées dans le rapport technique de 2008

L'impact fort d'un enseignement mettant en œuvre un apprentissage coopératif (C1) sur les compétences académiques des élèves est le principal résultat souligné par Robert SLAVIN. La taille d'effet globale de cette sous-catégorie est égale à 0,42 et a été calculée à partir de 8 études dont 5 avec attribution aléatoire d'élèves ou de classes dans les groupes traitement et contrôle. Les deux interventions STAD et IMPROVE qui participent à elles seules à ce résultat[a], sont d'ailleurs les deux seules interventions identifiées comme montrant une efficacité élevée (voir ci-dessus). Dans le Guide pour l'enseignant (aujourd'hui disparu), la recommandation « mettre en place un apprentissage coopératif » était placée en première position. A l'opposé, le manque d'efficacité des interventions basées sur l'exploitation d'un manuel particulier (et donc du suivi par les élèves d'un programme particulier) est souligné, la taille d'effet globale calculée pour la catégorie A étant égale à 0,03. Ici, c'est le faible impact des programmes soutenus par le NFS (A1) qui est mis en avant par Robert SLAVIN. L'auteur indique que les compétences visées par ces programmes innovants sont probablement moins évaluées par les tests standardisés que des compétences plus « classiques ». Aucune remarque qualitative n'est formulée concernant l'enseignement assisté par ordinateur, mais les chiffres parlent sans doute d'eux-mêmes (la taille d'effet moyenne de la catégorie B étant égale à 0,08. Dans le résumé et le Guide pour l'enseignant (aujourd'hui disparus), la rapide évolution des outils informatiques était repérée comme compliquant l'analyse de l'effet d'un enseignement assisté par ordinateur. En effet, de nombreuses études primaires évaluaient des logiciels qui n'étaient plus en usage ou qui avaient subi des modifications importantes au moment de la

[a] Une troisième intervention a été plus ou moins écartée de ces résultats, voir plus loin

réalisation de Robert SLAVIN (SLAVIN, 2009). On remarquera par exemple que deux études de cette catégorie datent de 1971, et évaluent donc des outils utilisés plus de 30 ans avant la conduite de cette synthèse[a]. Les résultats publiés par l'article de 2009 pour cette catégorie sont différents, puisque deux études publiées en 2008 et concernant l'enseignement assisté par ordinateur (les études n°23 et 24 de la sous-catégorie B2, voir annexe) n'ont pas été incluses. Les tailles d'effet globales calculées dans l'article de 2009 sont égales à 0,10 pour la catégorie B et 0,19 pour la sous-catégorie B2. Enfin, la méta-analyse de 2013 (CHEUNG, 2013) n'apporte ici pas d'éléments nouveaux : les tailles d'effet globales ont été calculées pour les trois catégories B1, B2 et B3, mais les études concernent les élèves du primaire et les élèves du secondaire dans leur ensemble[b].

En conclusion, Robert SLAVIN suggère aux enseignants (dans son article de 2009 comme dans son rapport technique) de s'intéresser davantage à l'organisation de la classe et à la mise en œuvre de certaines méthodes pédagogiques pour augmenter l'implication et la motivation des élèves, qu'au choix des programmes.

L'apprentissage coopératif

Nous allons revenir plus en détail sur la sous-catégorie définie par l'apprentissage coopératif (C1), car c'est cette caractéristique qui semble montrer un impact fort sur les résultats des élèves en mathématiques. Les études primaires rassemblées dans cette sous-catégorie ont été analysées afin de comprendre dans le détail les conclusions de Robert SLAVIN. Elles sont au nombre de 8[c], mais l'étude publiée par CALHOON[d] en 2003 sur une intervention combinant la méthode PALS (Peer-Assisted Learning Strategies) et la méthode CBM (Curriculum-Based Measurement) a été écartée de mon analyse. En effet, elle a concerné 92 élèves de la classe de la 3ème à la terminale, présentant des troubles de l'apprentissage et répartis dans 10 classes, et ne permet donc pas d'inférer sur une population plus large, inférence qui reste l'objectif de cette méta-analyse[e]. Si l'on exclue l'étude de CALHOON, deux interventions ont été incluses dans cette catégorie : l'intervention STAD pour laquelle 4 études ont été sélectionnées, et l'intervention IMPROVE, avec 3 études. Dans les études par comparaison de groupes mettant

[a] La récente méta-analyse sur l'enseignement des mathématiques en primaire du Center for Research and Reform in Education (PELLEGRINI , 2021) n'a inclus que des études postérieures à 2000 dans la catégorie Enseignement assisté par ordinateur.

[b] Ces tailles d'effet sont égales à 0,06 ; 0,19 et 0,09 respectivement (en utilisant le modèle des effets aléatoires).

[c] Et non pas 9 comme indiqué dans le résumé du rapport technique.

[d] Étude n°5 (C1). Cette étude a été sélectionnée par la méta-analyse réalisée par Campbell sur l'enseignement des mathématiques (et de la lecture) aux élèves scolarisés au secondaire et considérés comme en difficulté (DIETRICHSON, 2020)

[e] Robert SLAVIN publie une taille d'effet ES = - 0,30 (non significativement différent de zéro) ; cette taille d'effet n'a pas été recalculée ici. Cette étude a d'ailleurs été d'une certaine façon mise de côté par Robert SLAVIN lui-même puisqu'il calcule une taille d'effet globale pour les deux interventions STAD et IMPROVE égale à 0,46.

en œuvre ces deux interventions, les leçons, les objectifs et le matériel pédagogique (manuels, logiciels utilisés, etc.) sont identiques pour tous les élèves des groupes traitement et contrôle.

Toutes les tailles d'effet des études primaires ont pu être recalculées[a] ici grâce aux données publiées par les études primaires (c'est-à-dire les moyennes et écart-types des scores) et les résultats publiés par Robert SLAVIN ont a chaque fois été confirmés au centième près, à l'exception d'un calcul pour l'intervention STAD (voir ci-dessous). Des valeurs-p ont également pu être recalculées dans certains cas et comparées à celles publiées le cas échéant ; des résultats discordants ont cette fois été parfois trouvés. Un fichier Excel (à télécharger sur www.matahdoc.fr) permet de comprendre comment ces résultats ont été recalculés à partir des données de chaque étude. Dans ce qui suit, les notations utilisées restent les mêmes que ci-dessus, avec notamment N désignant la taille totale des échantillons de l'étude. Les études sont identifiées par le premier nom de leurs auteurs.

L'intervention STAD

La constitution de groupes hétérogènes est la clé de l'intervention STAD (*Student Teams-Achievement Divisions*) mise au point par Robert SLAVIN en 1980. Les élèves sont rassemblés en petits groupes en tenant compte de leur niveau scolaire : un élève de niveau élevé, deux élèves de niveau moyen et un élève de niveau faible travaillent ensemble en s'entraidant pendant 15 jours, période au bout de laquelle les groupes sont modifiés. Les élèves sont évalués individuellement, mais toutes les semaines un bilan est également dressé pour chacun des groupes (à partir des évaluations individuelles) et le groupe ayant la meilleure moyenne est volontairement mis en avant ; cet élément est présenté par Robert SLAVIN comme essentiel à l'intervention.

SLAVIN (1984)

Étude n°1, expérience randomisée de grande taille (N=588)

Plusieurs enseignants ont été enrôlés dans cette étude et la durée de la mise en œuvre du traitement est d'une année scolaire. La pédagogie de la maitrise a également été étudiée et 4 groupes d'élèves ont été constitués : dans le premier (groupe 1, 125 élèves), le traitement a associé pédagogie de la maitrise et travail en groupe, dans le second (groupe 2, 138 élèves) seul le travail en groupe a été mis en œuvre, dans le troisième (groupe 3, 165 élèves) seule la pédagogie de la maitrise a été étudiée ; le quatrième groupe constituait le groupe contrôle (groupe 4, 160 élèves). Deux tailles d'effet pour deux sous-groupes sont calculées par SLAVIN, ainsi qu'une taille d'effet globale égale à 0,21. Pour cette dernière, la méthode de calcul n'a pas été explicitée.

[a] C'est la formule (2) qui a été utilisée, sauf pour l'étude n°6 où la formule (1) a été employée.

Cette publication a également fait l'objet d'un examen par le What Works Clearinghouse[a], et leur procédure utilisée pour calculer la taille de l'effet permettant de mesurer l'influence de l'apprentissage coopératif sur les résultats des élèves a été reprise ici[b] : les groupes 1 et 2 constituent pour eux le groupe traitement (263 élèves), les groupes 3 et 4 constituent le groupe contrôle (325 élèves)[c]. La taille d'effet calculée et publiée par le WWC est égale à 0,21, c'est donc la même valeur que celle calculée par Robert SLAVIN. Une valeur-p inférieure à 0,03 est proposée par l'auteur et reprise par le WWC (recalculée ici comme comprise entre 0,012 et 0,12 en fonction des groupes considérés, voir annexe[d]).

Tailles d'effet recalculées

Groupe 1 versus groupe 4	Groupe 2 versus groupe 4	**Moyenne**	Groupes 1+2 versus groupes 3+4
$ES = 0{,}23$* (p=0,056)	$ES = 0{,}18$ (p=0,120)	$ES =$ **0,21**	$ES = 0{,}21$ (p=0,012)

* : c'est le seul résultat discordant parmi toutes les tailles d'effet recalculées

Tailles d'effet publiées par SLAVIN

STAD et pédagogie de la maitrise	STAD sans pédagogie de la maitrise	Taille d'effet globale
$ES = 0{,}24$	$ES = 0{,}18$	$ES = 0{,}21$ ($p < 0{,}03$)

Taille d'effet publiées par le WWC

Enseignement en groupe versus enseignement sans groupe
$ES = 0{,}21$ ($p < 0{,}03$)

NICHOLS (1996)

Expérience randomisée de petite taille (N=80)

Un seul enseignant a été enrôlé dans cette étude et trois groupes ont été constitués à partir d'un ensemble de 80 élèves au total (68 élèves de seconde, 10 élèves de première, et 2 élèves de terminale) : un groupe (groupe 1) où l'intervention STAD a été mise en œuvre 9 semaines suivies de 9 semaines de cours habituel, un autre (groupe 2) où ces deux périodes ont été inversées (d'abord 9 semaines de cours habituel puis 9 semaines de traitement) et le troisième groupe constituait le groupe contrôle (groupe 3). Les tailles des groupes ne sont pas publiées. Les critères de sélection fixent un nombre minimum d'enseignants égal à deux, et une durée d'intervention d'au moins 12 semaines ce qui pose alors la question de la sélection de cette

[a] Cette étude est classée comme conforme avec réserves aux normes WWC, cette réserve provenant de la forte attrition de l'échantillon (sur 1 092 élèves, 588 seulement ont complétés les deux tests et constituent l'échantillon analytique).

[b] Pour satisfaire aux méthodes de calculs utilisées par le WWC, les auteurs de l'étude primaire ont communiqué les moyennes posttests ajustées aux moyennes prétests et les écarts-types groupés (le WWC calcule le g de Hedges comme taille d'effet).

[c] Dans leur publication, le WWC publie les moyennes des scores traitement et scores contrôles, ce qui a permis de comprendre comment les groupes traitement et contrôles ont été constitués.

[d] La valeur-p dépend de l'effectif du groupe traitement et de l'effectif du groupe contrôle.

étude (voir discussion plus loin). Deux tailles d'effet ont été recalculées avec une moyenne égale à 0,20 ce qui correspond au résultat publié par Robert SLAVIN qui donne une estimation de la significativité statistique avec p<0,05 ; la valeur recalculée est égale à 0,40, ce qui indiquerait une absence de significativité statistique. Cette étude n'a pas été examinée par le WWC.

Tailles d'effet recalculées

Groupe 1 versus groupe 3	Groupe 2 versus groupe 3	Moyenne
ES = 0,16	ES = 0,24	ES = 0,20 (p = 0,403)

Taille d'effet publiée par SLAVIN

Traitement versus contrôle
ES = 0,20 (p < 0,05)

BARBATO (2000)

Quasi-expérience randomisée de petite taille (N=208)

Un seul enseignant a été enrôlé dans cette étude (ce qui pose à nouveau la question de sa sélection) où l'intervention STAD a été mise en œuvre sur le groupe traitement durant une année scolaire. La taille d'effet calculée est égale à 1,09 pour une taille d'échantillon totale de 208 élèves de seconde (107 élèves dans le groupe traitement, 101 élèves dans le groupe contrôle) et Robert SLAVIN propose une valeur-p telle que p<0,001 (le même résultat est recalculé ici). L'auteur de l'étude primaire souligne en conclusion qu'il convient de ne pas inférer sur une population différente de celle de l'étude (des élèves de seconde). Cette étude a été évaluée par le WWC comme non conforme à leurs normes, l'équivalence entre le groupe traitement et le groupe contrôle avant l'expérience n'ayant pas été démontrée.

Taille d'effet recalculée

ES = 1,09 (p = 0,000)

Taille d'effet publiée par SLAVIN

ES = 1,09 (p < 0,001)

REID (1992)

Étude par appariement de petite taille (N=50)

Le nombre d'enseignants est ici inconnu. L'échantillon au départ était constitué de 70 élèves de 5ème, mais l'échantillon analytique est constitué de 25 élèves pour le groupe traitement et 25 élèves pour le groupe contrôle. Aucune information n'est donnée sur l'enseignement concernant le groupe contrôle. Le traitement a été mis en œuvre durant une année scolaire et la taille d'effet calculée est égale à 0,38 ; Robert SLAVIN propose une valeur-p inférieure à 0,05 (la valeur-p recalculée est égale à 0,19 ce qui indiquerait une absence de significativité statistique). Cette étude a été évaluée par le WWC comme non conforme à leurs normes pour la même raison que l'étude précédente.

Taille d'effet recalculée

ES = 0,38 (p = 0,19)

Taille d'effet publiée par SLAVIN

ES = 0,38 (p < 0,05)

L'intervention IMPROVE

Cette méthode pédagogique a été développée par Zemira MEVARECH et Bracha KRAMARSKI en 1997 en Israël. IMPROVE est l'acronyme des sept points constitutifs de cet enseignement que l'on peut traduire par : Introduction de nouveaux concepts, Métacognition, Pratique, Retravailler les difficultés, Obtenir la maitrise, Vérifier et Enrichir. Les auteurs désignent les trois composantes principales de cette méthode comme étant la mise en place de processus métacognitifs et l'acquisition de méthodes stratégiques (notamment pour résoudre des problèmes) ; le travail en groupes constitués de 4 élèves de niveaux différents ; les retours fréquents faits aux élèves pour corriger leurs erreurs et/ou enrichir leurs connaissances. Cette intervention a fait l'objet de trois études par leurs créateurs (deux études sont publiées dans un même article). Aucune de ces trois études n'a été examinée par le WWC ni aucune estimation de la significativité statistique proposée par Robert SLAVIN. Quand cette dernière a été calculée ici, elle a toujours été inférieure à 0,05 (ce qui correspond donc à une taille d'effet statistiquement significative au niveau 0,05, voir annexes).

KRAMARSKI (2001)

Quasi-expérience randomisée de petite taille (N=182)

L'intervention a été mise en œuvre par 6 enseignants pour une durée d'une année scolaire sur un échantillon de 182 élèves de 6ème. L'auteur signale que le travail en groupe est pratiqué de façon habituelle en Israël (et qu'il a donc aussi concerné le groupe contrôle). Dans cet article, l'accent est clairement mis sur la pédagogie métacognitive. Les élèves sont répartis en trois groupes : dans le premier cette méthode pédagogique est mise en place dans les cours de mathématique et aussi dans les cours d'anglais (groupe 1, 60 élèves), dans le second (groupe 2, 60 élèves), la méthode pédagogique est mise en place uniquement dans le cours de mathématiques et le troisième groupe (groupe 3, 62 élèves) constitue le groupe contrôle. Deux tailles d'effet ont été recalculées ici et leur moyenne est égale à 0,79, ce qui correspond au résultat publié par Robert SLAVIN.

Tailles d'effet recalculées

Groupe 1 versus contrôle	Groupe 2 versus contrôle	Groupe 1+ 2 versus contrôle
ES = 0,93	ES = 0,65	ES = 0,79

Taille d'effet publiée par SLAVIN

ES = 0,79

MEVARECH (1997)

Étude par appariement de petite taille (N=247) - Première étude

L'intervention a été mise en œuvre par plusieurs enseignants, pour une durée d'un semestre. 247 élèves de 5ème ont été répartis dans le groupe traitement (99 élèves) et le groupe contrôle (148 élèves). Les résultats des élèves sont regroupés selon 3 niveaux de compétence, et 6 groupes analytiques ont été constitués : un groupe traitement de niveau faible (groupe 1), un groupe traitement de niveau moyen (groupe 2), un groupe traitement de niveau élevé (groupe 3), un groupe contrôle de niveau faible (groupe 4), un groupe contrôle de niveau moyen (groupe 5), un groupe contrôle de niveau élevé (groupe 6). Les tailles d'effet ont été recalculées en comparant les groupes traitement et les groupes contrôles de même niveau, puis en calculant une moyenne. Les élèves de niveau faibles sont 67, ceux de niveau moyen sont 71 et ceux de niveau élevé sont 109 (les effectifs de chacun des 6 groupes ne sont pas communiqués). Les élèves des groupes contrôles sont répartis dans des classes en fonction de leur niveau (donc les classes sont homogènes). Les élèves ont passé un test général (baptisé introduction à l'algèbre) ; une sous-partie des questions de ce test a été analysée dans un second temps et baptisée raisonnement mathématique. La taille d'effet moyenne recalculée est égale à 0,54 pour l'introduction à l'algèbre (premier tableau) et est égale à 0,68 pour le raisonnement mathématique (second tableau) ce qui correspond aux résultats publiés par Robert SLAVIN qui propose aussi une taille d'effet globale égale à 0,61.

Les auteurs de l'étude signalent que les résultats sont statistiquement significatifs pour les 6 comparaisons, à l'exception de celle concernant les élèves faibles dans le premier tableau (MEVARECH, 1997, p.380) ; Robert SLAVIN quant à lui affirme que les effets sont similaires pour tous les niveaux (SLAVIN, 2009, p.36).

Tailles d'effet recalculées

Introduction à l'algèbre			
Niveau faible	Niveau moyen	Niveau élevé	**Moyenne**
ES = 0,36	ES = 0,70	ES = 0,55	ES = **0,54**

Raisonnement mathématique			
Niveau faible	Niveau moyen	Niveau élevé	**Moyenne**
ES = 0,67	ES = 0,79	ES = 0,57	ES = **0,68**

Moyenne pour les deux tests : ES = 0,61

Tailles d'effet publiées par SLAVIN

Introduction à l'algèbre	**Raisonnement mathématique**	**Taille d'effet globale**
ES = 0,54	ES = 0,68	ES = 0,61

MEVARECH (1997)

Étude par appariement de grande taille (N=265) – seconde étude

L'intervention a été mise en œuvre par plusieurs enseignants, pour une durée d'une année scolaire. 265 élèves de 5ème ont été répartis dans le groupe traitement (164 élèves) et le groupe contrôle (101 élèves). Là aussi les résultats des élèves sont regroupés selon 3 niveaux de compétences (76 élèves de niveau faible, 92 élèves de niveau moyen et 97 élèves de niveau élevé). Là encore, les effectifs de chacun des 6 groupes ne sont pas communiqués. Les résultats portent cette fois sur le calcul algébrique, les exercices étant partagés en 4 rubriques (calculs numériques, substitution dans une expression algébrique, calcul littéral et problèmes). La taille d'effet moyenne est égale à 0,25 ce qui correspond au résultat publié par Robert SLAVIN.

Les auteurs de l'étude primaire soulignent que pour toutes les rubriques les résultats sont statistiquement significatifs à l'exception du calcul littéral ; et que les résultats des élèves les plus faibles ne sont pas statistiquement significatif (MEVARECH, 1997, p.385). Mais là encore Robert SLAVIN insiste sur la similitude des effets quel que soit le niveau des élèves (SLAVIN, 2009, p.36).

Tailles d'effet recalculées

Niveau faible	Niveau moyen	Niveau élevé	**Moyenne**
$ES = 0{,}26$	$ES = 0{,}24$	$ES = 0{,}25$	$ES = \mathbf{0{,}25}$

Taille d'effet publiée par SLAVIN

$$ES = 0{,}25$$

Partie 3 : discussion

Les calculs statistiques

D'une façon générale, les résultats publiés dans les méta-analyses sont séduisants car semblent relativement simples à comprendre : peu de nombres sont proposés et les conclusions sont formulées de façon claire. Cette simplicité est relative, et cache des choix méthodologiques obscures au plus grand nombre. Il n'est besoin que de se plonger dans la lecture de quelques articles publiant les résultats d'études primaires pour constater que la diversité des indicateurs et des méthodes statistiques utilisées requièrent de la part du méta-analyste une expertise aussi bien dans le domaine statistique que dans celui des sciences de l'éducation.

C'est le Δ de Glass qui a été utilisé ici pour calculer la taille d'effet pour chaque étude. Ce choix pose question, car il ne permet pas d'aller vraiment au-delà de la simple observation d'une série de résultats bruts. C'est encore la même chose si on considère les estimations de la significativité statistique des tailles d'effet des études qui parsèment l'article et qui n'ont pas été prises en compte de façon explicite dans cette méta-analyse : leur seule mention laisse finalement le lecteur décider de l'impact à leur donner. C'est la procédure statistique choisie qui est en cause ici. De nos jours, les méta-analyses suivent d'autres modèles statistiques (le modèle de l'effet fixe, mais plus souvent le modèle des effets aléatoires) qui permettent d'aller au-delà d'une simple description : estimation de la significativité statistique des tailles d'effet globales, analyse de l'hétérogénéité des résultats et analyse des sous-groupes d'études sont quelques uns de leurs atouts[a] (voir plus loin).

D'autres questions se posent quand plusieurs groupes de résultats sont publiés dans une étude primaire (ce qui est souvent le cas). Comment faire quand des sous-groupes d'élèves sont constitués ? Quels sont les résultats qui doivent être retenus par le méta-analyste ? Quels sont les résultats qui devront être considérés comme annexes et qui ne participeront pas au calcul d'une taille d'effet globale ? Quelles informations utiliser quand des tests ont été subdivisés pour étudier un aspect plus particulier des compétences, ou bien passés à des moments différents une fois l'expérience terminée ? Ces éléments techniques devraient être connus des lecteurs, car ils permettent d'abord de comprendre comment les indicateurs ont été calculés, puis également de procéder à d'autres calculs pour poursuivre le travail. C'est ce que des organisations internationales comme la Cochrane Collaboration et le What Works Clearinghouse ont su faire en publiant des documents complets sur leurs méthodes de calculs[b]. Il a été difficile de comprendre comment Robert SLAVIN a calculé certaines tailles d'effet pour certaines études, et les valeurs-p recalculées souvent différentes (voire même

[a] Mais pour utiliser ces modèles, c'est le g de Hedges qui doit être calculé, et non le Δ de Glass.

[b] Voir aussi les recommandations de PRISMA http://www.prisma-statement.org/

très différentes) de celles publiées sèment la confusion, particulièrement quand ces valeurs-p publiées sont favorables aux conclusions tirées.

Des règles parfois confuses

La sélection des études est une étape fondamentale pour toute méta-analyse et le respect des règles édictées par leurs auteurs ne peut souffrir aucune défaillance. D'après notre analyse, il semble ici que sur les 7 études primaires sélectionnées dans la catégorie apprentissage coopératif (C1), deux au moins auraient dû être exclue de cette méta-analyse (NICHOLS, 1996 et BARBATO, 2000). Pour une troisième, une absence d'information ne permet pas d'affirmer ici que l'étude est bien conforme aux critères définis par Robert SLAVIN (REID, 1992). Ces études concernent toutes l'intervention STAD et s'il était avéré que les critères de sélection n'étaient pas remplis, alors cette intervention ne saurait afficher l'efficacité forte telle qu'elle est annoncée par l'auteur. Quand nous avons comparé les études sélectionnées dans la méta-analyse concernant l'enseignement des mathématiques au secondaire (SLAVIN, 2009) avec celles sélectionnées dans la méta-analyse conduite sur l'enseignement assisté par ordinateur au primaire et au secondaire (CHEUNG, 2013), il n'a pas été possible de comprendre pourquoi des études sélectionnées en 2008 ne l'étaient plus 3 années plus tard. Aucune liste des publications exclues n'est fournie pour la seconde méta-analyse et il est donc impossible de connaitre la raison de ces absences.

Une certaine confusion règne également entre les différentes publications concernant la méta-analyse qui nous a particulièrement intéressé ici (SLAVIN, 2009). Deux études ont été rajoutées dans le rapport technique et ce rajout n'est annoncé nul part. Des distorsions concernent également l'évaluation des interventions (avec une intervention présentée par erreur comme moyennement efficace dans l'article publié 2009 et l'oubli d'une autre dans un tableau du rapport technique d'octobre 2008). Certaines tailles d'effet publiées dans l'article ne le sont pas dans le rapport technique. Il est également étonnant que certaines tailles d'effet globales n'aient pas été calculées (comme pour l'intervention Plato Web Learning Network mais aussi pour les deux sous-catégories A2 et C5), car on s'attend à ce qu'une méta-analyse applique des procédures de façon systématique. Les différentes formulations précisant les critères définissant les interventions dont l'efficacité est jugée forte sont également gênantes et rendent les évaluations confuses et peu fiables (dans le rapport technique, une taille d'échantillon minimale de 250 est mentionnée, elle est de 500 dans l'article de 2009). On regrettera également que le terme *median* soit utilisé en lieu et place de *weighted mean* à plusieurs reprises dans le rapport technique mais aussi dans l'article de 2009, cette coquille s'expliquant très probablement par le fait que la taille d'effet globale utilisée dans la méta-analyse concernant les élèves du primaire précédemment conduite était la médiane des tailles d'effet

Les éléments caractéristiques

Le point qui offre à mon avis le flanc à la discussion la plus épineuse ici concerne le regroupement des études en catégories. En effet, regrouper les études dans des catégories et/ou sous catégories et calculer pour chacun de ces regroupements une taille d'effet globale, donne aux éléments caractéristiques qui définissent ces catégories et ces sous-catégories une importance considérable. Cela revient en quelques sortes à associer, voire à réduire l'intervention à ce seul (ou à ces deux seuls) élément(s), et ce de manière implicite. Cette association peut poser problème pour plusieurs raisons.

Rassembler des études à l'intérieur de très larges catégories escamote des différences entre les interventions qui peuvent être très dissemblables et dont le seul point commun est finalement très ténu. Cela est particulièrement vrai pour la catégorie Programmes mathématiques (A). En effet, les auteurs des études primaires rassemblées dans cette catégorie n'ont pas comme objectif de comparer l'utilisation dans les cours de mathématiques de deux programmes différents, mais bien d'identifier l'impact de l'utilisation d'un programme « particulier » quand on le compare à une situation qualifiée de « traditionnelle », et ces programmes « particuliers » sont très différents les uns des autres. Calculer une taille d'effet globale ici n'a donc pas beaucoup de sens. Cette remarque peut évidemment être étendue aux deux autres catégories Enseignement assisté par ordinateur (B) et Méthodes pédagogiques (C). C'est plutôt l'hétérogénéité des tailles d'effet à l'intérieur d'une catégorie aussi vaste qu'il conviendrait d'analyser ; en fait, l'observation remarquable qui devrait être faite quand on considère l'ensemble des tailles d'effet des interventions classées dans la catégorie Programmes mathématiques, c'est effectivement que les tailles d'effet sont faibles mais surtout qu'elles sont homogènes. Imaginons une situation où la taille d'effet globale est toujours égale à 0,03, mais avec des tailles d'effets très différentes les unes des autres : la conclusion aurait bien évidemment été toute autre. Malheureusement les méthodes de calculs statistiques utilisées par Robert SLAVIN ne permettent pas ici de mener une telle analyse (le modèle des effets aléatoires aurait sans doute permis de répondre avec plus de pertinence à cette question comme cela a été remarqué auparavant).

Une autre raison concerne plus particulièrement les études de la troisième catégorie, Méthodes pédagogiques. L'apprentissage coopératif (sous-catégorie C1) est très certainement l'élément caractéristique majeur de l'intervention STAD, mais cela semble nettement moins clair en ce qui concerne l'intervention IMPROVE. Les auteurs des trois études primaires qui ont été sélectionnées pour analyser cette dernière intervention reconnaissent eux-mêmes ne pas pouvoir déterminer les rôles respectifs des trois éléments majeurs de leur méthode pédagogique (à savoir la pédagogie métacognitive, le travail en groupes et le retour sur les erreurs) ni leurs éventuelles interactions. On pourra souligner également que le travail en groupe est habituel en Israël et que cet élément de la méthode IMPROVE n'est peut-être pas aussi discriminant que la pédagogie métacognitive qui elle n'était clairement pas mise en œuvre dans les groupes contrôle. En pour finir, si on devait éliminer de la catégorie Apprentissage coopératif deux (voire trois) études pour non respect des critères de sélection, puis les trois études qui font état des résultats de l'intervention IMPROVE, il resterait bien peu

de grain à moudre pour faire de l'apprentissage coopératif la méthode pédagogique efficace telle que présentée par Robert SLAVIN[a]. Deux méta-analyses récentes[b] (l'une sous l'égide de la collaboration Campbell, l'autre sous l'égide du What Works Clearinghouse) associent également les études primaires à des éléments caractéristiques, mais sans imposer d'étanchéité entre les catégories. Ainsi, une étude peut être associée à plusieurs éléments, ce qui correspond sans doute mieux à la réalité du terrain. Dans ce cas, c'est une association entre des caractéristiques et des tailles d'effet qui sont estimées, et cela peut ne pas correspondre à un lien de cause à effet (voir à ce sujet la note publiée sur le site www.mathadoc.fr).

Enfin un dernier point concerne l'absence de règles qui permettraient l'évaluation de l'impact des éléments caractéristiques sur les apprentissages. Elles devraient pourtant être définies de façon transparente et peut-être même avant tout calcul[c]. Ce manque de clarté donne à l'évaluation de ces éléments un caractère subjectif contraire aux principes même d'une synthèse de ce type.

Le What Works Clearinghouse

On a plusieurs fois évoqué dans ce texte le travail du What Works Clearinghouse. Établi sur le même territoire, partageant des objectifs similaires, soutenu par le gouvernement et les autorités locales, ce « grand frère » a parfois été un peu bousculé par Robert SLAVIN. On pourra prendre comme exemple l'évaluation de l'intervention Saxon Maths. SLAVIN calcule une taille d'effet globale égale à 0,14 et souligne avoir évincé de sa synthèse l'article de WHITE (1986), ce dernier ayant utilisé un test élaboré par le chercheur (p.22 du rapport technique de 2008). Cette même étude aurait été considérée par le WWC comme conforme à leurs normes, ce qui l'aurait conduit à donner son crédit à cette intervention[d]. Aussi bien dans l'article que dans le rapport technique, aucune étude n'a été référencée sous le nom d'auteur WHITE avec comme année de publication 1986. Par contre, c'est l'article de WILLIAMS (1986) que l'on retrouve aussi bien dans la liste des publications rejetées par Robert SLAVIN (p.116 du rapport technique de 2008) que dans les rapports d'intervention du WWC[e].

De façon plus large, les différences observées entre les méthodes et donc les résultats des méta-analyses conduites par des organisations différentes, mais parfois à l'intérieur d'une même organisation par des chercheurs différents, devraient susciter un débat sur les règles à fixer pour répondre à une même question en utilisant les synthèses quantitatives.

[a] On ne peut également pas ne pas voir qu'il est lui-même à la source de cette préconisation, ce qui n'ajoute qu'un peu plus de confusion à toute cette discussion.
[b] DIETRICHSON (2021 ; 2020), FUCHS (2021).
[c] Et si des règles sont définies a posteriori, cela doit être clairement mentionné.
[d] En 2007, le WWC présente l'intervention Saxon Math comme ayant des effets positifs.
[e] Avec une taille d'effet égale à 0,65. Notons qu'en 2017, le WWC a réévalué cette intervention qu'il considère aujourd'hui comme n'ayant pas montré d'efficacité sur les apprentissages des élèves ; l'étude de WILLIAMS (1986) a été considérée cette fois comme non conforme aux normes WWC.

Pour conclure

Nous avions évoqué au début de ce texte une étape 0 que toute méta-analyse est de nos jour encouragée à suivre et qui consiste en la formulation explicite, détaillée et transparente de l'ensemble des règles rédigées sous la forme d'un protocole qui doit être publié avant la réalisation de la synthèse. La rédaction d'un tel protocole aurait sans doute permis d'éviter certaines des faiblesses repérées dans la méta-analyse de Robert SLAVIN.

La complexité des méta-analyses, tant au niveau des outils statistiques employés que des protocoles élaborés dans le but notamment de sélectionner rigoureusement des études, de les décrire en déterminant leurs caractéristiques, puis d'évaluer l'influence de ces dernières sur les variables dépendantes, requiert un travail de fond qui exige des moyens humains et financiers importants. Pour développer cette assise conceptuelle et technique, des institutions comme le What Works Clearinghouse, la collaboration Cochrane ou la collaboration Campbell ont développé un arsenal d'outils disponibles gratuitement sur leurs sites internet qui donne à voir l'ampleur du travail fourni. Les techniques statistiques ont encore progressé depuis la publication en 2009 de la méta-analyse de Robert SLAVIN, et aujourd'hui les regards se tournent plus volontiers sur l'analyse de l'hétérogénéité des tailles d'effets ou de l'influence de facteurs pertinents sur ces tailles d'effet (DIETRICHSON, 2020), que sur le simple calcul d'une taille d'effet globale. C'est d'ailleurs cet objectif que s'est fixé la récente méta-analyse publiée par le Center for Research and Reform in Education (PELLEGRINI, 2021) et qui concerne l'enseignement des mathématiques à l'école primaire.

On pourra mettre au crédit de Robert SLAVIN d'avoir tout particulièrement cherché à définir les caractéristiques qui permettent de définir les qualités d'une bonne méta-analyse et la synthèse menée sur l'enseignement assisté par ordinateur (CHEUNG, 2013) s'inscrit tout à fait dans cette démarche. Toujours sur le site Best evidence encyclopedia, une étude publiée en 2015 montre que plusieurs éléments caractéristiques des études, comme la qualité de leur design, la taille de leur échantillon ou la qualité des mesures utilisées sont associés à une diminution des tailles d'effet[a] (CHEUNG, 2016).

La preuve statistique est fondée sur un principe fondamental simple : le chercheur doit prouver qu'il n'a pas tort. Lutter contre les erreurs de jugement souvent alimentées par nos convictions intimes exige une discipline parfaite. Les méta-analyses se présentent comme un outil efficace pour analyser l'effet que certains traitements peuvent avoir sur les apprentissages et elles ne peuvent pas se permettre d'être de qualité discutable. De nombreuses voix se lèvent pour que cette qualité s'améliore, et des recommandations concernant la réalisation, mais aussi la rédaction et la publication des rapports (PRISMA) voient le jour. Il est important de suivre leurs conseils car nous avons encore besoin d'enrichir nos connaissances sur les liens qui unissent pratiques d'enseignement et apprentissages.

[a] Plus la qualité du design de l'étude, la taille de son échantillon ou la qualité des mesures est élevée et moins la taille de l'effet est élevée.

Glossaire

Méta-analyse	Synthèse quantitative d'études primaires sélectionnées utilisant des procédures statistiques.
Étude primaire	Étude expérimentale prospective dont proviennent les données utilisées par les méta-analyses.
Catégorie (sous-)	A la fois un élément caractéristique partagé par un ensemble d'études et le regroupement que ces études constituent.
Taille d'effet	Différence standardisée des moyennes : écart entre les moyennes divisé par l'écart-type des scores.
Intervention	Cela peut être un programme scolaire ou un ensemble de pratiques qualifié de méthode pédagogique ou un produit commercial (comme une application informatique ou un manuel), dont l'objectif est d'améliorer les compétences des élèves. On parle aussi de traitement.

Abréviations

ES	Effect size traduit par Taille d'effet
IMPROVE	Introduction de nouveaux concepts, Métacognition, Pratique, Retravailler les difficultés, Obtenir la maitrise, Vérifier et Enrichir
STAD	Student Teams-Achievement Divisions
WWC	Waht Works Clearinghouse

Références

Barbato, R. (2000). Policy implications of cooperative learning on the achievement and attitudes of secondary school mathematics students. Unpublished doctoral dissertation, Fordham University, New York.

Calhoon, M. B., & Fuchs, L. S. (2003). The effects of peer-assisted learning strategies and curriculum-based measurement on the mathematics performance of secondary students with disabilities. *Remedial and Special Education*, 24(4), 235–245.

Cheung, A., & Slavin, R. E. (2013). The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: A meta-analysis. *Educational Research Review, 9*, 88-113.

Cheung, A., & Slavin, R. (2016). How methodological features affect effect sizes in education. *Educational Researcher, 45* (5), 283-292.

Cooper, A., Hedges, L. & Valentine, J. (2019). *The Handbook of Research Synthesis and Meta-Analysis***.** NEW YORK: Russell Sage Foundation.

Dietrichson J, Filges T, Klokker RH, Viinholt BCA, Bøg M, Jensen UH. Targeted school - based interventions for improving reading and mathematics for students with, or at risk of, academic difficulties in Grades 7 – 12: A systematic review. *Campbell Systematic Reviews.* 2020;16:e1081. https://doi.org/10.1002/cl2.1081

Fuchs, L.S., Newman-Gonchar, R., Schumacher, R., Dougherty, B., Bucka, N., Karp, K.S., Woodward, J., Clarke, B., Jordan, N. C., Gersten, R., Jayanthi, M., Keating, B., and Morgan, S. (2021). Assisting Students Struggling with Mathematics: Intervention in the Elementary Grades (WWC 2021006). Washington, DC: National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U.S. Department of Education. Retrieved from http://whatworks.ed.gov/.

Kramarski, B., Mevarech, Z. R., & Lieberman, A. (2001). Effects of multilevel versus unilevel metacognitive training on mathematical reasoning. *Journal of Educational Research*, 54(5), 292–300.

Mevarech, Z., & Kramarski, B. (1997). IMPROVE: A multidimensional method for teaching mathematics in heterogeneous classrooms. *American Educational Research Journal*, 34, 365-394.

Nichols, J. D. (1996). The effects of cooperative learning on student achievement and motivation in a high school geometry class. Brief research report. *Contemporary Educational Psychology,* 21, 467–476.

Pellegrini, M., Neitzel, A., Lake, C., & Slavin, R. (2021). Effective programs in elementary mathematics: A best-evidence synthesis. *AERA Open*, 7 (1), 1-29.

Reid, J. (1992). The effects of cooperative learning with intergroup competition on the math achievement of seventh grade students. (ERIC Document Reproduction Service No. ED 355106)

Roques N. (2021a), Comment enseigner les maths ? La réponse du What Works Clearinghouse, www.mathadoc.fr

Roques N. (2021b), Mesurer l'effet d'un traitement. Les méta-analyses en sciences de l'éducation www.mathadoc.fr

Slavin, R. E., & Karweit, N. L. (1984). Mastery learning and student teams: A factorial experiment in urban general mathematics classes. *American Educational Research Journal*, 21(4), 725-736.

Slavin R. E. and Lake C., (2008) Effective Programs in Elementary Mathematics : A Best-Evidence Synthesis, *Review of Educational Research*, 78 (3) , 427-515.

Slavin R. E., Lake C. et Groff C. (October, 2008), Effective Programs in Middle and High School Mathematics: A Best-Evidence Synthesis, Johns Hopkins University, Rapport technique, Version 1.4

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.

Les deux textes supprimés du site internet

Effective Programs in Middle and High School Mathematics : A Best Evidence Synthesis, Last Updated March 11, 2009 (résumé de 4 pages, supprimé du site au 01/05/2021)

Slavin Robert E., Lake Cynthia, Groff Cynthia, January 2010, Educator's Guide What Works in Teaching Math? (supprimé du site au 01/05/2021)